1 The AI of Ethics

Robert C. Williamson

Abstract. The spectacular rise of AI has lead to anxiety regarding the *ethics of AI* — how might one conceive and control the ethical impacts of AI technology. I will first sketch the traditional framing of "the ethics of AI." I will then provide an inverse view, that sees AI as an extension of human reasoning, with concomitantly different conclusions regarding decisional autonomy. I will further argue that the solution to many ethical problems relies upon the better use of cognitive technologies such as AI. That is, we have largely had it backwards — it is the *AI of ethics* that warrants our attention, and the root of the harm is the *use* of AI, not the *technology itself*.

Invert, always invert!
— Carl Gustav Jacob Jacobi

1.1 Introduction

Anxiety regarding the ethical implications of the new(ish) technology of Artificial Intelligence (AI) is widespread, and arguably is the contemporary version of the recurring civilisational anguish described by Arthur Herman (1997). Like the earlier worries regarding "civilising the machine" arising from the apparent conflict between the new technologies of the day (railroads, steam, and factories) and higher community ideals in the 19th century (Mumford, 1944; Kasson, 1999), there is a common confounding of the technology itself, and the means by which it is used. This chapter is primarily concerned with how the technology can improve our ethical reasoning and behaviour. But this analysis presumes a subject — who is it that the machine serves? In the conclusion, I will return to the point, and locate, as eventually the 19th century critics of the then-new technologies did, the root of harm in the choices made by humans as to how the new technologies are used, rather than intrinsic failings of the technologies themselves. Thus my message, while simple, is two-fold: rather than an autonomous *machine*, AI can be better thought of as a *tool*, that can help humans make better ethical decisions, but as with any tool, much depends upon *whose hands control it*, and *to what end it is put*.

Many technical problems open themselves to progress by the application of Carl Jacobi's dictum quoted above. In this chapter I invert the problem of *the ethics of AI* by reframing

it as *the AI of ethics*. I introduce the reframing, arguing why it is warranted and useful. I then summarize, in caricature form, seven common theses from the literature on the ethics of AI. I subsequently offer some anti-theses from the inverse view of the AI of ethics.

The ethics of AI is typically framed as the problem of ensuring the ethical behaviour of AI systems, especially with regard to the "making of decisions." It is cast as a problem for us humans to civilise and manage the dangerous technology of AI (which is considered to be fundamentally different to all prior technologies), to control it, and to ensure that it implements what we, in our unaided human decision making, have decided is the right thing to do. A popular response is argumentation over which abstract principles should apply (Jobin et al., 2019), with little to no recognition that these principles alone are far from sufficient for the task, nor as universal as naively thought — compare the list of universals elaborated by Brown (1991) with the non-universality of fairness both linguistically (Wierzbicka, 2006, pages 141–167) and politically (Fischer, 2012).

Inversely, the premise (and promise) of the AI of ethics is that the technology of AI, being a part of human culture and inextricable from the modern human mind, is a powerful tool which humans can use in order to make better ethical decisions. In the same way that a craftsman can make better furniture by use of tools (which they may well have designed and built), humans can raise their ethical standards and quality of ethical decision making by virtue of delegating some tasks to AI technologies. Like any tool, since the first use of a stick or rock, AI can be ethically misused, but the ethical harm comes from the *use*, and is not something intrinsic to the technology.

My goal in this chapter (a longer version of which will be published elsewhere), is to suggest that by alternatively framing the problem as "the AI of ethics" we are likely to make more progress on what ultimately matters – improving the standard of ethical behavior of the only moral agents that we care about – humans, including ourselves. I do not offer any solutions; I am simply suggesting that a reframing of the problem could be very helpful; a reframing which allows us to face squarely our own ethical deficiencies, allows a fair comparison with alternative methods of making ethical decisions, and illuminates the location of the real source of the ethical problem with AI, namely with the behavior of individuals and groups who sometimes use the tool of AI for ethically indefensible activities.

1.2 Seven theses on the ethics of AI

I now tersely sketch seven theses implicit in much recent consideration of the ethics of AI.

1.2.1 The problem is clear — AI machines have "bias" and this is one reason we need to prevent them from making ethical decisions about people

There are several claims here. First, there is a well defined notion of "bias" that one can sensibly talk about (this might be bias of a machine, or of data). Second, the distinction between a human making a decision and a machine making one is clear. Third, indeed the

distinction between human and machine (and machine and tool) is also clear. Fourth, the problem is intrinsic to the technology — some AI technologies have properties that make them ethically suspect; the job of technologists is to fix this problem with the technology.

1.2.2 AI is fundamentally different to other technologies; the artefact is the problem

This viewpoint is common regarding many new technologies: from the railroad to electricity, they are often considered as *unprecedented*, *autonomous* (Winner, 1977), and *threatening* because of the agency it has which threatens to "re-engineer humanity" (Frischmann and Selinger, 2018). The changes threatened are absolute, and the introduction of the new technology will lead to a total revolution (this last point is the "fallacy of total revolution" as expressed by Corn (1986)). The essence of technology is the *artefact*, and the essence of AI is the *algorithm*, which is a well defined notion. It is these AI algorithms that need to be regulated, certified or otherwise controlled. The reason AI is different is because it, unlike all prior technologies, is a *cognitive* technology that "makes decisions" and thus encroaches upon what was previously the exclusive preserve of humans. Because of this difference, none of the lessons learned from previously new technologies is of relevance.

1.2.3 Goodness is manifest to humans, as is their ethical reasoning, and human ethical reasoning is the gold standard, but this is not quantifiable

Humans just "know" what is good — we have an intuitive sense of right and wrong, and this is a qualitative, transcendental notion. Any attempt to reduce it calculation is a denial of what is greatest about humanity. Likewise, any attempt to use mathematical reasoning to analyse different notions of fairness just proves that mathematicians have missed the transcendental and ineffable nature of ethical choice (Powles and Nissenbaum, 2018).

1.2.4 Ethics is reasoning which is all done inside a human's head

Ethics is the use of reason to solve moral problems, which needs to be done (wholly) inside people's heads (this is the internalist, representational or "neurocentric" view of mind (Malafouris, 2013)). Reasoning is conscious thinking, which like our knowledge, lives in, and is done, purely inside our head. Furthermore, *we know our own minds*:

> Knowledge of our own mental states seems to be authoritative, in the sense that if we think we are in a particular mental state that cannot be challenged. Knowledge of our own mental states seems also to be privileged, meaning that we know the contents of our own minds always better than we know the contents of the minds of other people. Another important feature that is related to introspection is immediacy. This notion implies that introspective beliefs, as opposed to perceptual beliefs, are non-inferential and non-evidence based (Aydin, 2015, page 77).

1.2.5 Ethical decisional autonomy is binary and unambiguous

If one delegates some decisional autonomy to a machine, then one is morally negligent: a common example is the use of GPS systems to navigate which (supposedly) implies "to

navigate by GPS requires no sense of where you are" (Frischmann and Selinger, 2018). Humans have autonomy; machines should not. And autonomy is a thing that single individuals have, not groups, contra Mackenzie and Stoljar (2000). Autonomy is essentially related to the notion of *undue influence* (Niker et al., 2018). If we do not watch out, we will lose our human self-rule to that of machines.

1.2.6 Decisions are conceptually, spatially, and temporally atomic

The notion of "making a decision" is unambiguous. Decisions are conceptually indivisible (atomic). Furthermore, they are "made" at a unique and precise point in time and space by a single agent. This implies that a decision has to be "made" *either* by a human, or by a machine.

1.2.7 Ethical decision making should be reserved to unaided humans, not machines

There is a deep seated, justifiable belief, that machines simply should not make ethical decisions (Bigman and Gray, 2018). Implicit in this is that anything mechanical is anti-human (Button et al., 1995, Chapter 3) and values are fundamentally irreducible to mechanism, in the sense that no mechanism can possibly represent them (Midgley, 1994, Chapter 9).

Misunderstandings of the nature of machines exacerbate the belief in this thesis:

> An evolved being is not one made like a machine. Unlike machines, which typically have a single, fixed function, evolved organisms have a plurality of aims, held together flexibly in a complex but versatile system. It is only this second, complex arrangement that could make our kind of freedom possible at all (Midgley, 1994, pages 163–164).

1.3 Seven (anti)theses on the AI of ethics

I now build on the ground of the previous section by developing an antithetical perspective on the seven theses above.

1.3.1 "Bias" is a misleading way to frame the problem

There is no doubt that significant harm has been caused to many people from the use of AI decision technologies. And there is also no doubt that many varieties of "bias" can be partially held to blame in part for this. Without diminishing the value of analysing such "biases", I want to suggest an alternate framing of the problem. One difficulty of framing the ethics of AI in terms of the search for removal of "bias" is that the word can mean so many different things (Olteanu et al., 2019) and furthermore, these meanings are often not made explicit (Blodgett et al., 2020). Other examples include (Cowgill and Tucker, 2019) which leans on the definition of "biased action" due to (Becker, 1971), and the well known, but still not well understood, phenomenon of selection bias (Meng, 2018).

However, my concern with "bias" is that it tends to try and isolate the cause of a problem that can not be so isolated. What causes harm is the complete decision making system, which is affected by the formulation of the problem being solved, the decision as to what

data to collect and how, various corruptions of that data, the goals specified for the decision (which may encode all manner of strategic choices), and the potentially flawed implementation of goals via inadequate algorithms. There are indeed a multitude of ways of getting things wrong. But there are also a multitude of ways of getting things "right" in the sense that there are many potential ethical desiderata one may wish to impose. Suppose one concludes that matching the false positive rate of a classifier to the population proportions across sensitive features is your desired goal. Then generically you can not expect, necessarily, that the false negative rate will also match. This is *not* a consequence of "bias" in the algorithm or the data: it is simply the fact that you made one ethical choice out of the infinite number you may have.

Contrast the common way AI technologies are used to make decisions with a properly run scientific experiment. In the experiment, the goal is declared first, and data is gather subsequently with a view to that goal. All sources of potential error are assessed and mitigated as much as possible *before gathering the data*. The starting point of using AI to make decisions is often some pre-existing data, perhaps accumulated for other purposes. The particular data available is a consequence of decisions made long ago for reasons perhaps unknown. One can give the name "bias" to the errors arising in this setting, but I claim this is unhelpful because it is not the removal of this "bias" that is needed, but rather that the end-to-end system needs to be considered holistically and analysed *before the data is gathered*.

1.3.2 AI is a technology, which shares many attributes with other technologies, and is part of human culture

In a very real sense, tools created Homo Sapiens.
— Sherwood L. Washburn (1959)

AI is a fantasy, an idea, an artefact, a tool, a mechanism, a machine, a technology, a business model, or a marketing tool. Perhaps it is all of these. Clarifying how we should conceive of AI seems necessary in order to reason sensibly about its ethical ramifications. Artefacts are the most tangible manifestations, but this causes problems since artifacts have multiple meanings (Engeström, 1990). Technological artefacts are often ascribed politics, most famously by Langdon Winner (1980) regarding the bridges spanning Robert Moses' Long Island parkway: "Moses wanted to keep poor African Americans, who would have had to use buses that could not negotiate his over-passes, from certain suburban areas" (Mitcham, 2014, page 17). Alas, as shown by Woolgar and Cooper (1999), Winner's ascription is false, as evidenced by their documentation of long operating bus routes on the parkway. If one can get the "politics of an artefact" as mundane as a bridge so spectacularly wrong, what hope is there with something more complex like AI?

The history of technology sees all such categorisations (of what technology is) as fuzzy at best (McNeil, 1990), with some authors concluding there is an irreducible plurality of viewpoints necessary to understand cybernetics and AI (Cordeschi, 2002). For the pur-

poses of this chapter, noting the ambiguity, I argue it is better to think of AI as a decisional or prediction *technology* (Agrawal et al., 2018), which requires answering the question: *what is a technology?* Technology is not merely its artefacts; a compelling definition is that it is simply *knowledge* (Layton, 1974), particularly *useful* knowledge (Mokyr, 1990) or (reflecting technology's evolutionary trajectory) *successful* knowledge (Levinson, 1988), which extends our human capabilities, and this is the critical difference between technological artefacts and other objects (Lawson, 2008, 2010).

Another important distinction is that between *tool* and *machine*; AI is almost universally construed as the latter. The distinction is partly one of familarity (the modern hand-held electric power drill would be considered a fantastical *machine* a century ago; now it is a mundane *tool* that one takes for granted). This reflects the observation of Gunkel (2012, page 31): "as Karl Marx pointed out, [experts in mechanics] often confuse these two concepts, calling 'tools simple machines and machines complex tools,' there is an important and crucial difference between the two, and that difference ultimately has to do with the location and assignment of agency." Tools have evolved over time, from the earliest primitive rocks and stones (Childe, 1944; Oakley, 1965), but their development is inextricably tied to the growth of human knowledge and human intelligence (Sternberg and Preiss, 2005). The additional agency that is sometimes ascribed to machines (Hodges, 2008) does indeed seem relevant for the consideration of AI, but most of the literature seems to adopt the thesis that autonomy is dichotomous (see the discussion of autonomic hierarchies in subsection 1.3.7) which forces an unjustifiable hard choice.

All technologies offer benefit and cause harm: "Technology is neither good nor bad; nor is it neutral" (Kranzberg, 1986, page 545). Fair comparisons with non-technological processes for making ethical decisions are rare. Cowgill and Tucker (2019, page 39) observed that ethical failings can be more manifest in digital systems because one can more easily measure the effects. Indeed, some jurisdictions *prohibit* the aggregation of statistics regarding human decisions made by certain privileged groups, such as judges, which would be necessary to detect such ethical shortcomings (Taylor, 2019), thus guaranteeing a technological tool (the performance of which *can* be measured) will *always* look worse! The notorious Amazon recruiting tool reportedly would have worked better than humans, but because its perceived ethical shortcomings were so manifest, it was deemed (far) worse than the poorer performing but opaque human alternative (Cowgill and Tucker, 2019, page 44). It seems unreasonable to worry about ethical shortcomings of technological decision systems when shortcomings of current human systems are not merely invisible, but, in some cases, are guaranteed so by legislation, for example by prohibiting the gathering relevant data as in the case of the French judges (Taylor, 2019).

1.3.3 Ethical reasoning is not manifest, but can be quantified

Moral decision making is built upon reason: Sapolsky (2017, page 479) writes of the "primacy of reasoning in moral decision making." Anguish over the knowability and trans-

parency of algorithmic moral reasoning suggests we should, to be fair, ask to what extent is our *own* unaided moral reasoning knowable and transparent to us. Whether our *moral* reasoning is manifest is comprehensively answered by (Haidt, 2007; Graham et al., 2011): experiments with challenging ethical scenarios repeatedly show subjects holding very firm views as to the ethically right decision, but being utterly unable to articulate *why*. Greene (2013) calls this "the tragedy of commonsense morality". This is arguably a consequence of a deeper problem, namely that we do not know our own minds, neither its immediate percepts, nor its reasons (Nisbett and Wilson, 1977; Wilson, 2002): "Few things are more completely hidden from my observation than those hypothetical elements of thought which the psychologist finds reason to pronounce 'immediate'" (Pierce, 1958, CP8.144); "There is a substantial and growing body of evidence suggesting that much of what we do, we do unconsciously, and for reasons that are inaccessible to us" (Greene, 2008).

Greene (2008), argues that the difference between deontological and consequentlist moral reasoning is a difference of cognition and information processing: the difference between "stereotyped and flexible behavior." Crockett (2013) describes this as "multiple decision systems." Greene (2008) distinguishes "intuitive emotional responses [which] drive prepotent moral intuitions while 'cognitive' control processes sometimes rein them in. Education is to a large extent the development of one's 'cognitive' capacities, learning to think in ways that are abstract, effortful, and often either nonintuitive or counterintuitive." But it is well recognised that "only in hypothetical examples in which 'all else is equal' does consequentialism give clear answers. For real-life consequentialism, everything is a complex guessing game, and all judgments are revisable in light of additional details. There is no moral clarity in consequentialist moral thought, with its approximations and simplifying assumptions. It is *fundamentally actuarial.*" (page 64, italics added). It is precisely the difficulty of performing this consequentialist reasoning in the presence of uncertainty that motivates the use of external "actuarial" technologies designed for that very purpose — namely Machine Learning, a type of AI.

Human moral reasoning is *not* manifest, and it seems that further understanding requires mathematization, and mathematical reasoning relies heavily on external technologies, the most recent of which is formal statistical decision theory (French and Insua, 2000), which underpins modern machine learning and artificial intelligence.

1.3.4 Unaided human ethical decision making is not the gold standard, and proper behaviour has to take account of context

Ethical decision making is almost always reasoning under uncertainty, and unaided humans are terrible at this (Tversky and Kahneman, 1974). We can not even easily accept alternative beliefs (Golman et al., 2016), and there is a strong sense of "value homophily." Many hold "a belief that the rules governing proper behavior should be universal vs. a preference for particularistic approaches that take into account the context and the nature of the relationships involved" (Nisbett, 2003, pages 61–62) and (literally) fight to the death

over the principles, rather than the decision itself; confer the story related in (Jonsen and Toulmin, 1988) regarding unanimous agreement about a challenging ethical decision, but vehement disagreement about the reasons supporting the decision.

The technology of AI is not a gold standard either, but it is a tool that has the potential to allow us to *raise* our standards in a manner that takes account of context better than we can because of its better information processing abilities.

1.3.5 Human ethical decision making has always relied upon technology

Just as you cannot do very much carpentry with your bare hands, there is not much thinking you can do with your bare brain.
— Daniel Dennett (2000)

What is to be made of the fact that philosophical thinking cannot be carried on by the unaided human mind but only by the human mind that has familiarized itself with and deeply interiorized the technology of writing? — Walter Ong (2012, pages 169–170)

A traditional view of moral reasoning is that it is a unique human privilege, and arguably defines our very humanity. The idea that not only could future technologies assist in this endeavour, but that current and past technologies have been doing so for ages, would likely be met with disbelief by many. Nevertheless, if one thinks of technology as a *mediator* it is clear that they do indeed play such a role:

> Technologies play an important role in virtually every moral decision we make. The decision how fast to drive and therefore how much risk to run of harming other people is always mediated by such things the layout of the road, the power of the cars engine, the presence or absence of speed bumps and speed cameras. The decision to have surgery or not is most often mediated by all kinds of imaging technologies and blood tests, which help to constitute the body in specific ways and organize specific situations of choice (Verbeek, 2011, page 59).

The aversion to the use of technology in moral decision making perhaps stems from the rule-like nature of technology, which (appears to) remove the act of choice from the human:

> My claim is that the resistance against the idea that technologies are morally significant is in fact a resistance against the need to give up the modernist idea that actions and decisions can only be moral when they are the sole product of individual human choice without external influences (Verbeek, 2014, page 76).

The concern is perhaps the apparent blackness of the boxes AI lives in:

> Something about the phrase 'black box' — a common description of machine learning techniques — may make machine learning sound incompatible with notions of accountable government. But it would be more accurate to view machine-learning algorithms, or any other statistical procedures, not as complete black boxes, but rather as extensions of existing human decision making. . . Democratic government itself, in a collective sense, is decision making according to the algorithm of majority rule (Coglianese and Lehr, 2016).

1.3.5.1 Cognitive technologies and AI We have delegated some cognition to "cognitive technologies" for centuries: "civilization advances by extending the number of important operations which we can perform without thinking about them" (Whitehead, 1911, Chapter 5); "the more of the details of our daily life we can hand over to the effortless custody of automatism, the more our higher power of mind will be set free for their own proper work" (James, 1918, page 122). Young children autonomously adopt the viewpoint: "'When you program a computer, there is a little piece of your mind and now its a little piece of the computers mind,' said Deborah, a sixth- grade student in an elementary school that had recently introduced computer programming into its curriculum" (Turkle, 2005, page 1). In this subsection I explore the notion of a *cognitive technology*, and argue that this is a useful way to think about AI, especially with regard to ethical concerns. As will be obvious to anyone who consults his paper, I am greatly indebted to Peter Skagestad (1993), who carefully developed the cognitive technology perspective that the digital computer is fruitfully regarded as a means for *augmenting* human thought by automating and accelerating the production and manipulation of symbols, rather than as an *independent source* of thought.

One can only get so far with hand tools; for harder problems we need power tools. Doug Engelbart (1962) pursued the amplificatory idea of Ross Ashby (1957): "it seems to follow that intellectual power, like physical power, can be amplified." Engelbart sought to achieve such intelligence amplification via *augmentation* with computers, and he argued that such a goal warranted the most serious consideration because "man's problem solving capability represents possibly the most important resource possessed by a society" (Engelbart, 1962, page 131). Many of the computing technologies we now take for granted are the consequence of the line of work that Englebart initiated. To date, most of the problems considered for such augmentation by computer have been non-ethical. But there is no reason why our ethical reasoning abilities can not be similarly amplified and extended. We should consider the IA of ethics — that is *Intelligence Augmentation* — (Skagestad, 1996) to aid our ethical reasoning.

Modern cognitive technologies do raise a worry (for some) that we humans will be made redundant. Howard Rheingold (2000), in his lively history of "tools for thought," observed "Few people object to the notion of understanding things that nobody understands – until it is suggested that the agent for achieving that understanding might be an intelligence that is made of silicon rather than protoplasm." He suggested a remarkably simple way to alleviate the concern about redundancy: *one should not automate the work, only the materials*: "if you like to play music, do not build a 'player piano'; instead program yourself a new kind of instrument." This perspective can be seen to be one where the tool is a *mediator*, not a replacement (Verbeek, 2008a). What might such ethical instruments look like?

"Non-cognitive cognizers" (the phrase is due to Hayles (2017)) such as computers have long been recognised as language technologies, not mere number crunchers (Nofre et al.,

2014). But language itself, especially alphabetic written language, is a technology *par excellence*. Understanding written language as a technology, and its role in ethical reasoning is arguably helpful to understand the potential role of AI in solving ethical problems. Finally, while viewing AI as an intelligence amplifier is indeed helpful, it does not mean that the human mind (in one's head) is not changed by it; on the contrary, the continued use of such cognitive technologies changes our thinking even when we are not using them (Malafouris, 2013, page 81).

1.3.5.2 From orality to morality — technologising philosophy

The Greek alphabet . . . [is] a piece of explosive technology, revolutionary in its effects on human culture, in a way not precisely shared by any other invention. — Eric Havelock (1982, page 6)

Language does for intelligence what the wheel does for the feet and the body. — Marshall McLuhan (2011)

The relationship between philosophy and technology is almost universally considered to one of "the philosophy of technology," and the idea that technology has something to offer philosophy would be seen by many as absurd (Levinson, 1982). The philosopher sits in his armchair and ponders the philosophical implications and assumptions of technology. But as Levinson (1988, page 72) observes, he would be well served to pay more attention to the technology of the armchair itself; one can not do armchair philosophy without it! More fundamentally, *one can not do philosophy without the information technology of writing*. In this subsection I go beyond this observation to argue that the introduction of the technology of alphabetical writing commenced the exteriorisation of our intellect, and underpinned the very development of moral philosophy, and that this insight offers us guideance for the study of the ethical ramifications of AI.

It is common to consider language first as a *communication* technology, and only second as a *cognitive* technology, although this view is now being challenged (Reboul, 2015) via the notion that the communicative function is a consequential exaptation of its original cognitive function. Walter Ong described the effects of "technologising of the word":

> Philosophy and all the sciences and 'arts' . . . depend for their existence on writing,. . . they are produced not by the unaided human mind but by the mind making use of a technology that has been deeply interiorized, incorporated into mental processes themselves. . . . Philosophy . . . should be reflectively aware of itself as a technological product – which is to say a special kind of very human product. Logic itself emerges from the technology of writing (Ong, 2012, page 169).

Jack Goody (2000, pages 140-141) considered writing as a "technology of the intellect." In doing so, he "was not thinking primarily of the immediate implications of different types of script, the level of technique, but rather of what kinds of cognitive or intellectual operations could be carried out in writing that were impossible, difficult, or less efficient in speech." Marshall McLuhan (2011, page 389) suggested some direct answers to such a

question: "The uniformity and repeatability of print created the *political arithmetic* of the seventeenth century and the *hedonistic calculus* of the eighteenth."

Crediting language and writing with much in human development is hardly new (Sproat, 2010), and the alphabet in particular is singled out for especial credit (McLuhan and Logan, 1977; Logan, 2004). Writing forms an *external memory system*, greatly increasing our information processing power (Logan, 2007). There is no going back:

> The growth of the external memory system has now so far outpaced biological memory that it is no exaggeration to say that we are permanently wedded to our great invention, in a cognitive symbiosis unique in nature. External memory is the well of knowledge at which we draw sustenance, the driving force behind our ceaseless invention and change, the fount of inspiration in which succeeding generations find purpose and direction and into which we place our own hard-won cognitive treasures (Donald, 1991, page 356).

Harold Innis (1950), noting the danger of such broad generalizations, observed

> [T]he art of writing provided man with a transpersonal memory. Men were given an artificially extended and verifiable memory of objects and events not present to sight or recollection. Individuals applied their minds to symbols rather than things and went beyond the world of concrete experience into the world of conceptual relations created within an enlarged time and space universe. ... Writing enormously enhanced a capacity for abstract thinking .... Man's activities and powers were roughly extended in proportion to the increased use and perfection of written records.

Nowadays,"external memory is a critical feature of modern human cognition" (Donald, 1991, page 312). Clearly the development of writing, and then printing, vastly influenced the modes and means of communication. But, for our purposes, what is more interesting and suggestive is the way it changed how people thought. Richard Nisbett (2003, page 156) suggested

> Greek and other Indo-European languages encourage making properties of objects into real objects in their own right – simply adding the suffix 'ness' or its equivalent. ... [T]his practice may foster thinking about properties as abstract entities that can then function as theoretical explanations.

But *how* can features of a language affect the way we think about moral questions, and what has this to do with the ethics of AI? The first question has been comprehensively answered by Eric Havelock (1963, 1978, 1986), who argued that the critical event was Plato's discovery of the method of abstraction:

> Here is a new frame of discourse and a new kind of vocabulary offered to the European mind. We take it for granted today as the discourse of educated men. It does not occur to us that once upon a time it was necessary for it to have been discovered and defined and insisted on, so that we could easily and complacently inherit it. This discovery is essentially Plato's, even though he is building on a great pioneering effort in this same direction which had preceded him (Havelock, 1963, page 26).

Without abstraction, there is limited communication (Levinson, 1988, page 120). Without abstraction, there is no conceptual thinking, only "image-based" thinking — like Funes the

Memorious, one would be "not very capable of thought" because "[t]o think is to forget differences, generalize, make abstractions" (Borges, 1970, page 94). And without abstraction, there is no concept of "justice" only "situational thinking" (Havelock, 1978). Psychologists have shown the importance of levels of abstraction in the way humans (Rosch et al., 1976; Rosch, 1978) and other species (Vonk and MacDonald, 2004) categorise the world, and for determining our intellectual and managerial ability (Jaques, 1978).

A "philosopher" is a person who has mastered this device of abstraction:

> The clues to the history of the word 'philosopher', and therefore to a history of the idea of philosophy, are first fully supplied in the *Republic* itself, where the type of person symbolised by this word is identified simply as the man who is prepared to challenge the hold of the concrete over our consciousness, and to substitute the abstract (Havelock, 1963, page 281).

Later, Havelock (1986, page 4) distilled his argument as follows:

> The argument as completed offered the twin proposal that the notion of a moral value system which was autonomous, while at the same time capable of internalization in the individual consciousness, was a literate invention and a Platonic one, for which the Greek enlightenment had laid the groundwork, replacing an oralist sense of 'the right thing to do,' as a matter of propriety and correct procedure.

If the idea of material objects playing such a significant role in the development of our thinking abilities is hard to swallow, it is perhaps worthwhile instead considering the idea of externality and social relations: "Social relations make artefacts out of persons" (Strathern, 1998, page 135). This helps make the notion of external storage (and indeed external cognition) perhaps less confronting — it simply becomes the way your society "thinks". This idea is hardly new:

> Cognition is the most socially-conditioned activity of man, and knowledge is the paramount social creation. The very structure of language presents a compelling philosophy characteristic of that community, and even a single word can represent a complex theory. . . . every epistemological theory is trivial that does not take the sociological dependence of all cognition into account in a fundamental and detailed manner (Fleck, 1935/1979, page 42).

Thus the very development of conceptual and theoretical thinking, which is essential to ethical reasoning, relies upon technology! Bentley and O'Brien (2012, page 2) describe "the appearance of technology capable of accumulating and manipulating vast amounts of information outside humans" as a "cultural tipping point" to which "all other features [are] derivative." AI is arguably the next step in this journey.

1.3.5.3 Implications for AI The significance (for questions of AI) is that the cognitive structures underpinning these theoretic cultures "exist mostly outside the individual mind" and rely upon technological hardware in the form of external memory devices: "Theoretic culture was from its inception externally encoded; and its construction involved an entirely

new superstructure of cognitive mechanisms external to the individual biological memory" (Donald, 1991, page 274).

Any legitimate statistical reasoning requires abstract thinking; perhaps this is why the ecological fallacy remains so prevalent (since it the very notion of a statisitical population is an abstraction) (Robinson, 1950; Freedman, 1999), and why there are such political anxieties about the use of statistics at all (Desrosières, 1998). Since much moral reasoning requires actuarial reasoning about aggregates, our unaided human intuition needs all the help it can get from a range of cognitive technologies in order to perform sound moral reasoning. Foremost amongst these are mathematical and statistical reasoning, which is best performed by machine (because machines can be built to perform it more accurately and quickly than unaided humans). AI thus becomes essential for ethical reasoning.

1.3.6 Decisions are nonatomic

What does it mean to "make a decision"? As usual, dictionaries are of little help: the Oxford English Dictionary defines "to make a decision" simply as "to decide". And "decide" is defined as "To determine upon as a course of action; to resolve to do (something) or bring (something) about" or "the action, fact, or process of arriving at a conclusion regarding a matter under consideration; the action or fact of making up one's mind as to an opinion, course of action, etc" which surfaces the core problem — how does this "determining" or "resolving" occur?

There are several distinct ideas hidden inside the phrase "making a decision":

> To say that a person has made a decision may mean (1) that he has started a series of behavioral reactions in favor of something, or it may mean (2) that he has made up his mind to do a certain action, which he has no doubts that he ought to do. But perhaps the most common use of the term is this: 'to make a decision' means (3) to make a judgment regarding what one ought to do in a certain situation after having deliberated on some alternative courses of action (Ofstad, 1961, page 15).

The presence of technological augmentation is usually considered as interfering with the autonomy of this judgement: Eilon (1969) analyses a situation where a "data processing facilty" may "encroach" on a decision makers domain by "by taking over parts or the whole function of analysis". He singles out the control that the person has: "In the extreme case, when control is completely impersonalistic, the decision-maker ceases to have a meaningful role; he ceases to be a decision-maker." But this presumes that the person was not the commissioner of the data processing facility (which may or may not be true).

The standard presentation of statistical decision theory is that there is a set of theories or outcomes X and a set of actions (or "decisions") A and a loss function $L\colon X \times A \to \bar{\mathbb{R}}$. The set X can be thought of as states of the world, one's features, a representation of the world, or as "outcomes" and for each x, $a \mapsto L(x, a)$ is typically viewed as a (negative) outcome-contingent utility (French and Insua, 2000) that codifies the cost of making decision/action a in the situation that x occurs. The goal, in such theories, is to construct a function d from

some the set X to the action set A that minimizes $L(x, a)$ in some sense (one might minimize certain averages with respect to given probability measures, or one might minimize over a the maximum over x). In the sequential decision theory set-up, one plays an iterated game in which each move is effectively a problem as described above (Wald, 1947).

Thus, from the standpoint of (statistical) decision theory, "making a decision" reduces to determining the entire decision *function* $d\colon X \to A$. Whereas in common parlance, "making a decision" is more akin to *evaluating* d *at some particular* $x \in X$. This distinction highlights that one needs to think about the problem at two levels: the determination of the overall decision function $d\colon X \to A$, and the evaluation of it in a particular instance $d(x)$.

This illustrates the conceptual non-atomicity of "making a decision": when is the decision really "made"? When one constructs the decision function, or when one evaluates it? This distinction can be further refined: one can extend the notion of a decision function to a higher order function $d\colon X_1 \to X_2 \to A$. Thus given some $x_1 \in X_1$, one has a more specific function $d(x_1)\colon X_2 \to A$. When $x_2 \in X_2$ becomes available, one can evaluate $d(x_1)(x_2) \in A$ to determine the final action. Things get interesting when you think that today, using information $x_1 \in X_1$, you will compute $d(x_1)$ and tomorrow, when $x_2 \in X_2$ becomes available, you will compute $d(x_1)(x_2)$. Consider the simple example due to French and Insua (2000, Page 9), which they used to distinguish between "decision" (made using full powers of reason) and "choice" (which is instinctive and intuitive): "When offered a cigarette, an individual makes an unthinking choice and refuses it. But long ago, he or she may have deliberated and decided upon a policy of no smoking." His "policy" is simply the function $d(x_1)(\cdot)$.) This is (modulo notation, and ignoring the overall evaluation criteria) no different to the classical sequential decision theory set-up. But your future self of tomorrow is not you now, so there seems little difference to suppose that *you* compute $d(x_1)$, but *another person* determines $d(x_1)(x_2)$. Obviously this can be extended to arbitrary hierarchies with $d\colon X_1 \to X_2 \to \cdots \to X_n \to A$. And there is no need for the passage of any time between the two steps, as long as they are done in order.

But if one is conceptually comfortable delegating the computation of $x_2 \mapsto d(x_1)(x_2)$ to another *person*, one can equally well delegate it to a *machine* ...a "non-conscious cognizer" in Katherine Hayles terminology. All the machine needs to do is to evaluate $x_2 \mapsto d(x_1)(x_2)$. The person (agent) might do all the cognitive heavy-lifting in evaluating $d(x_1)$, leaving only very limited "discretion" to the machine contingent upon x_2. In such a case, demanding that there is a single entity that "makes the decision" is clearly silly: the overall "making of the decision", the evaluation of the second order function $d(x_1)(x_2)$, is a cooperative effort between the person and machine.

A useful concept for thinking about decisional autonomy is notion of *second order responsibility* due to Illies and Meijers (2009). They assume two moral agents (I changed their notation to a boss B and subordinate S). The level of moral responsibility is couched in terms of the different possible courses of action available to the two agents. If the second agent has no choice whatsover, they are not held to any account. When one thinks of two

persons in the sequential set-up described above, it does indeed make sense to hold the second accountable *only within the set of choices available*. They say

> The second-order responsibility of B does not diminish the normal (or first-order) responsibility of S; S remains fully responsible for her choice on the basis of her Action Scheme at a certain time. But we do hold B responsible for having influenced S's Action Scheme as that is indeed what he can be blamed or praised for (Illies and Meijers, 2009, page 433).

But this does *not* imply that one should hold a machine to which one delegates the second order choice similarly responsible — for if B designed or commissioned the machine, all that is happening is that B is taking full responsibility for evaluating $d(x_1)(x_2)$, but has broken the execution into two parts. The machine does not have the moral agency, or conscious thought that would allow it to take on *any* moral responsibility. We do need to consider the *behaviour* of the entire system into account. But the *responsibility* rests with B: the designer and commissioner.

When the roles are reversed it is easy to mistakenly assign moral responsibility to the machine as is done for example by Verbeek (2008b) and Illies and Meijers (2014). That is, machine M chooses $d(x_1)$ and the human S gets to only choose $d(x_1)(x_2)$ and responsibility is assigned to M. But this is misleading since the machine M (a technological artefact) was the result of an earlier decision by a (potentially different) person D, choosing some $x_0 \in X_0$, and thus S is actually evaluating (deciding) upon $x_2 \mapsto d(x_0)(x_1)(x_2)$; The designer D is responsible for $d(x_0)$, $d(x_0)(x_1)$ and the subordinate S is responsible for $d(x_0)(x_1)(x_2)$. This is true whether M is a speedbump (Verbeek, 2008b; Latour, 1999), another "mundane artefact" (Latour, 1992) or an advanced piece of AI technology.

I do agree with Verbeek (2008b, page 98) that technological artefacts help *constitute* and *mediate* human moral choices. Ultimately one needs to assess the moral implications of the entire system. But it is the human designers, owners, commissioners, operators, and beneficiaries of the system, that need to be held accountable, not the mere technological artefacts (Illies and Meijers, 2014). It is the persons who have the intentionality, freedom, ability to understand the implications, general purpose reason, and capacity for empathy and moral sentiment.

Decisions are generically non-atomic; there is *always* a prior context, which was (at least in part) a consequence of earlier decisions by some (potentially other) person. Thus the notion of "second order responsibility" is generically necessary. Such second-order responsbility does not absolve any person of the choices that *they* make, but "no one is to be blamed or praised for the choices of *others*" (Illies and Meijers, 2014, pages 171ff).

Peterson and Spahn (2011) argue that that technological artefacts themselves can "never be (part of) moral agents" viewing them only as "neutral tools" While recognizing that strong neutrality can hardly be tenable (given the influence technologies have), they compellingly argue for a "weak neutrality thesis". Consider the three subtheses that technological artefacts (1) never figure as moral agents, and are never (2) morally responsible for their effects, and (3) never affect the moral evaluation of actions. The "strong neutrality

thesis" demands (1–3); the weak version only 1 and 2 (Peterson and Spahn, 2011, pages 422ff).

1.3.7 Ethical decisional autonomy is graded

> *For the twenty-first century, then, autonomy is human action removed in time. This, in a sense, is the essence of the term 'programming' – telling the computer what to do at some point in the future, when the program is run*
> — David Mindell (2015)

"Autonomy," literally "self-rule" or "self-legislation," is often held to mean acting according to rules that one imposes upon oneself.

> At the core of the moral philosophy of Kant is the claim that morality centers on a law that human beings impose upon themselves, necessarily providing themselves, in doing so, with a motive to obey. Kant speaks of agents who are morally self-governed in this way as autonomous (Schneewind, 1998, page 483).

Other conceptions of autonomy are possible: Dworkin (1988, page 20) conceives of it as

> A second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher-order preferences and values.

My purpose here is not to try and *define* autonomy precisely; on the contrary, I argue that it cannot be precisely defined because of its graduated nature. This gradation is centrally important to considerations of AI in ethical decision making. Mindell (2015) talks of "the myth of full autonomy" of technological devices, which if true would be a legitimate concern; but neither any technological artefact, nor any human, is entirely autonomous.

Autonomy is considered vital for human flourishing: "it is widely held that being autonomous is the default presumption under which we interact with adults in the modern world" (Schneewind, 2013, page 147). Although much modern philosophy makes little or no use of the concept (Schneewind, 2013, pages 150–154), in the last 50 years there has been an explosion of interest in it. But recent neuroscientific evidence shows that our perception of our own autonomy is quite exaggerated (Felsen and Reiner, 2011), which is not to say that we lose our moral responsibility (Dubljević, 2013). Such misperception affects our economic (Bossaerts and Murawski, 2015) and moral and social decision making (Yoder and Decety, 2018). Chemicals (hormones) directly affect our economic decisions (Nadler and Zak, 2016) and our ethical behaviours such as pro-sociality (Crockett, 2009). These influences merely mean our autonomy is not absolute but graded (Walker, 2011; Nagel, 2013).

Partial autonomy means we retain some discretion, but not complete control. But we remain responsible relative to the discretion we have (Walker, 2011). Many technological artefacts can partially reduce our autonomy, without reducing our accountability. A speed-bump limits our otherwise morally free choice on how fast we may drive on a road: "Moral agency, therefore, does not require complete autonomy. Some degree of freedom can be

enough for one to be held morally accountable for an action. And not all freedom is taken away by technological mediations, as the [example] of . . . driving speed make[s] clear" (Verbeek, 2011, page 59). And this perception of technological influence on autonomy is widely felt with regard to such behavioural "nudges" such as speedbumps (Felsen et al., 2013).

Our tools have no autonomy, nor indeed agency. But we have a "second-order responsibility" for what our tools do on our behalf — namely enacting our own moral choices (Illies and Meijers, 2009, 2014). Their moral significance is only as mediators (Verbeek, 2011) or as amplifiers of our own ethical intelligence, which we remain entirely responsible for — if a human uses the "wrong tool" for a job, or uses the tool badly, no responsibility rests with the tool; the tool does not decide — we merely use the tool to decide for us.

1.4 AI Tools — Technologies of the Intellect

> *A regrettably common misperception of particularly some newer, less familiar, seemingly more menacing forms of technology is that they are an invasive, extra-human force.* — Robert McC Adams (1996, page 265)

> *In no sense is technology nonhuman, since it is developed and used by human minds and hands. . . . The example to which I turn again (and there are many possibilities) is the mathematical table. That is essentially the product of writing, but one that can be taught to and learned by those who can neither read nor write. Yet it provides those who use it with a special cognitive tool, a technology of the intellect.* — Jack Goody (2000)

> *All of human history, upon close scrutiny, ultimately resolves into the history of the invention of better tools.* – Edmund Reitlinger

Artificial Intelligence technology, like any tool, expands our capabilities:

> [T]hanks to the hammer, I become literally another man, a man who has become 'other', since from that point in time I pass through alterity, the alteration of that folding (Gibson, 2015). This is why the theme of the tool as an 'extension of the organ' makes such little sense. Those who believe that tools are simple utensils have never held a hammer in their hand, have never allowed themselves to recognize the flux of possibilities that they are suddenly able to envisage (Latour, 2002, page 250).

Like the hammer, AI can be used for ethical harm. Compared to decision making by humans, at least *in principle* AI can ensure auditability and accountability at a level hard to achieve otherwise (Kleinberg et al., 2017; Parker and Danks, 2019): one is unlikely to see laws passed prohibiting the gathering of statistics on AI decisions like that recently passed for human judges in France (Taylor, 2019). Evaluating AI systems that are cognitive extenders, is a different task to that of evaluating them as pure autonomous agents (Hernández-Orallo and Vold, 2019; Ienca, 2017). The distinction is akin to Richard Harper et al.'s distinction (2008) between two different notions of a "smart home": one that is smart in itself, and one which helps the occupants be smarter.

Technologizing ethical decisions is akin to demanding one's argument and position is written down. As Peter Skagestad (1993, page 167) observed, following Karl Popper's lead, "only once something has been written down does it become criticizable and hence epistemologically interesting." I claim that the same sentence is true if one replaces "epistemological" by "ethically." Being forced to "write-down" one's ethics with AI will demand quantification with its concomitant increase in precision and auditability: "[A]gencies should recognize that the use of algorithms will often compel agency decision makers to engage in quantitative coding of value judgments that have typically been made qualitatively" (Coglianese and Lehr, 2016, page 1218).

1.4.1 The real problem — The ethically harmful *use* of AI

Comparing AI to other technologies, and how their harms have been controlled and mitigated (Williamson et al., 2015), I suggest that we need to focus on the *uses* of the technology, rather than the technology itself. I illustrate this idea with the two notions of privacy and autonomy. As Dworkin (1988, page 104) puts it "One way of interfering with your autonomy is to deceive you. This interference with information is, however, just the opposite kind from that involved in interference in privacy. What is controlled is the information coming to you, not the information coming from you."

This is hardly a new problem, as illustrated by the following example regarding an earlier control technology (Beninger, 1986). Dworkin (1988, page 38) quotes from *Anna Karenina* regarding how Stefan Arkadyevitch's moral principles are not their own because they regularly read a newspaper. Stefan's "beliefs are not his because they are borrowed; and they are borrowed *without even being aware of their source; and, it is implied, Stefan is not capable of giving some account of their validity . . .* " (italics added).

Consider the distinction made by Susser et al. (2018); Susser (2019); Vold and Whittlestone (2019) between *persuasion* (a third party attempting to convince you in a manner that is manifest to you and with reasons) versus *manipulation* (Coons and Weber, 2014) (changing your mind in a manner that you are not even aware of, and thus have no power to resist). By virtue of their power to build accurate predictive models of people on the basis of large amounts of seemingly innocuous information, AI technologies especially enable manipulation, which is the business model of advertising (Bartholomew, 2017). AI did not create advertising, but by making it more precise and efficient, it has weaponised it to an extent that violates human moral autonomy (Villlarán, 2017; Gunn et al., 2018) to a degree where great harms are being caused (Bernal, 2018).

But such harms are largely ignored in the debate about the ethics of AI. Indeed they are quite invisible. Jørgensen (2017) interviewed employees at large AI based platform companies and found

> None of the people I spoke to associate the company's dedication to privacy with limits on the information that is collected about its users. Data collection and targeted advertising is the taken-for-granted context in the sense that it is a premise for using the service.

There is little point in making the advertising algorithms fair to different ethnic groups if the greater harm of invisible maniuplation of voting preferences (for example) remains unchecked. The very business model is one of treating people as means, not ends in themselves.

Rather than using AI in an unethical manner, we should look to more convivial ways (McQuillan, 2016), returning to the ideal behind the creation of the personal computer. Lewis Mumford's definition of democracy included the requirement for "protection against arbitrary external controls, and a sense of individual responsibility for behavior that affects the whole community" (Mumford, 1964, page 1). I claim such protection is necessary against the wielding of the external controls enabled by AI technology in the hands of a few. We have seen this before. Randolph Probstfield, a senator from the state of Minnesota, wrote in 1896 of the appropriation and control of the the powerful technologies of his day:

> [O]ur would be masters have a corner on the whole outfit of the inventions, and they are now just as much employed to the destruction of human rights as formerly in the absence of those inventions the peoples ignorance was used as a means. (Quoted in (Pollack, 1962, page 23))

Focussing upon the harmful *use* of AI is simply an updated version of a common theme from 19th centry critics of the impact of the then-new technologies of railroads, steam and factories. As Kasson (1999, page 192) observes of Edward Bellamy's *Looking Backwards*, "the root of this new social barbarism ... lay in the rise of a new industrial aristocracy rather than in the technology itself" — that is, it was the "grossly inequitable distribution of the fruits of the technology" that was the root cause of the immense harms that were perpetrated. Bellamy's contemporaries were in consensus that focussing on the benefits to individuals, rather than the one dimensional pursuit of money, was necessary to reverse the harm.

We could do well to learn from them, for like the gradual transformation of the new factory town of Lowell Masachussetts described by Kasson (1999, Chapter 2), which was explicitly founded on high republican ideals, only to degenerate, after a generation, into an exploitative and miserable existence for those who worked there, we see a similar pattern today, with companies founded on wide-eyed promises of the new technology of AI to "universally" serve humanity and "do no evil" only to become, after less than a generation, beholden only to the interests of power, and corrupting and quashing any notion of higher purpose with a torrent of informational manipulation and exploitation in the service of the highest bidder. We do, however, have a choice:

> Nothing that man has created is outside his capacity to change, to remold, to supplant, or to destroy: his machines are no more sacred or substantial that the dreams in which they originated. — Lewis Mumford (1944, page 415)

Acknowledgement Thanks to insightful criticism from Julia Haas and Jennifer Wortmann Vaughan.

Bibliography

Adams, Robert McC. 1996. *Paths of fire: An anthropologist's inquiry into western technology*. Princeton University Press.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. 2018. *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Ashby, W. Ross. 1957. *An introduction to cybernetics*. Chapman and Hall.

Aydin, Ciano. 2015. The artifactual mind: overcoming the 'inside–outside' dualism in the extended mind thesis and recognizing the technological dimension of cognition. *Phenomenology and the cognitive sciences* 14 (1): 73–94.

Bartholomew, Mark. 2017. *Adcreep: The case against modern marketing*. Stanford Law Books.

Becker, Gary S. 1971. *The economics of discrimination*. The University of Chicago Press.

Beninger, James R. 1986. *The control revolution*. Harvard University Press.

Bentley, R. Alexander, and Michael J. O'Brien. 2012. Cultural evolutionary tipping points in the storage and transmission of information. *Frontiers in Psychology* 3 (569): 1–14.

Bernal, Paul. 2018. Fakebook: why Facebook makes the fake news problem inevitable. *NILQ* 69 (4): 513–530.

Bigman, Yochanan E., and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181: 21–34.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp, Technical report, arXiv:2005.14050v2.

Borges, Jorge Luis. 1970. Funes the memorious. In *Labyrinths*, 87–95. Penguin.

Bossaerts, Peter, and Carsten Murawski. 2015. From behavioural economics to neuroeconomics to decision neuroscience: the ascent of biology in research on human decision making. *Current Opinion in Behavioural Science* 5: 37–42.

Brown, Donald E. 1991. *Human universals*. McGraw-Hill.

Button, Graham, Jeff Coulter, John R. E. Lee, and Wes Sharrock. 1995. *Computers, minds and conduct*. Polity press.

Childe, V. Gordon. 1944. *The story of tools*. Cobbett Publishing Co. Ltd..

Coglianese, Cary, and David Lehr. 2016. Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal* 105: 1147–1223.

Coons, Christian, and Michael Weber, eds. 2014. *Manipulation: Theory and practice*. Oxford University Press.

Cordeschi, Roberto. 2002. *The discovery of the artificial: Behaviour, mind and machines before and beyond cybernetics*. Springer.

Corn, Joseph J. 1986. *Imagining tomorrow: History, technology and the American future*. MIT Press.

Cowgill, Bo, and Catherine Tucker. 2019. Economics, fairness and algorithmic bias. Preprint, Columbia and MIT.

Crockett, Molly J. 2009. The neurochemistry of fairness: Clarifying the link between serotonin and prosocial behavior. *Annals of the New York Academy of Sciences* 1167: 76–86.

Crockett, Molly J. 2013. Models of morality. *Trends in cognitive sciences* 17 (8): 363–366.

Dennett, Daniel C. 2000. Making tools for thinking. In *Metarepresentations: A multidisciplinary perspective*, ed. Dan Sperber, 17–30. Oxford University Press.

Desrosières, Alain. 1998. *The politics of large numbers: A history of statistical reasoning*. Harvard University Press.

Donald, Merlin. 1991. *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.

Dubljević, Veljko. 2013. Autonomy in neuroethics: Political and not metaphysical. *AJOB Neuroscience* 4 (4): 44–51.

Dworkin, Gerald. 1988. *The theory and practice of autonomy*. Cambridge University Press.

Eilon, Samuel. 1969. What is a decision? *Management Science* 16 (4): 172–189.

Engelbart, Douglas C. 1962. Augmenting human intellect: A conceptual framework, Technical Report SRI Project 3578 / AFOSR-3223, Stanford Research Institute.

Engeström, Yrjö. 1990. When is a tool? Multiple meanings of artifacts in human activity. In *Learning, working and imagining: Twelve studies in activity theory*, 171–195. Orienta-Konsultit Oy.

Felsen, Gidon, and Peter B. Reiner. 2011. How the neuroscience of decision making informs our conception of autonomy. *AJOB Neuroscience* 2 (3): 3–14.

Felsen, Gidon, Noah Castelo, and Peter B. Reiner. 2013. Decisional enhancement and autonomy: public attitudes to overt and covert nudges. *Judgement and Decision Making* 8 (3): 202–213.

Fischer, David Hackett. 2012. *Fairness and freedom: A history of two open societies: New Zealand and the United States*. Oxford University Press.

Fleck, Ludwick. 1935/1979. *Genesis and development of a scientific fact*. The University of Chicago Press.

Freedman, David. 1999. Ecological inference and the ecological fallacy, Technical Report 549, Department of Statistics, University of California, Berkeley.

French, Simon, and David Ríos Insua. 2000. *Statistical decision theory*. Arnold.

Frischmann, Brett, and Evan Selinger. 2018. *Re-engineering humanity*. Cambridge University Press.

Gibson, James J. 2015. *The ecological approach to visual perception (classic edition)*. Psychology Press.

Golman, Russell, George Loewenstein, Karl Ove Moene, and Luca Zarri. 2016. The preference for belief consonance. *Journal of Economic Perspectives* 30 (3): 165–88.

Goody, Jack. 2000. Technologies of the intellect: Writing and the written word. In *The power of the written tradition*, 132–151. Smithsonian Institution Press. Chap. 8.

Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. 2011. Mapping the moral domain. *Journal of Personal and Social Psychology* 101 (2): 366–385.

Greene, Joshua. 2013. *Moral tribes: Emotion, reason, and the gap between us and them*. The Penguin Press.

Greene, Joshua D. 2008. The secret joke of Kant's soul. In *Moral psychology (3) — The neuroscience of morality: Emotion, brain disorders, and development*, ed. Walter Sinnott-Armstrong, 35–80. MIT Press.

Gunkel, David J. 2012. *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Gunn, Sylvia, Ebba Brunnstrom, Grace Engelman, and Matthew Flathers. 2018. Moral manipulation: A Kantian take on advertising and campaigning. *Brown Unversity Journal of Philosophy, Politics and Economics* Fall.

Haidt, Jonathan. 2007. The new synthesis in moral psychology. *Science* 316: 998–1002.

Harper, Richard, Alex Taylor, and Michael Molloy. 2008. Intelligent artefacts at home in the 21st century. In *Material agency: Towards a non-anthropocentric approach*, eds. Carl Knappett and Lambros Malafouris, 97–120. Springer.

Havelock, Eric A. 1963. *Preface to Plato*. Harvard University Press.

Havelock, Eric A. 1978. *The Greek concept of justice: From its shadow in Homer to its substance in Plato*. Harvard University Press.

Havelock, Eric A. 1982. *The literate revolution in Greece and its cultural consequences*. Princeton University Press.

Havelock, Eric A. 1986. *The muse learns to write: Reflections on orality and literacy from antiquity to present*. Yale University Press.

Hayles, N. Katherine. 2017. *Unthought: The power of the cognitive nonconscious*. University of Chicago Press.

Herman, Arthur. 1997. *The idea of decline in western history*. The Free Press.

Hernández-Orallo, José, and Karina Vold. 2019. AI extenders: The ethical and social implications of humans cognitively extended by AI. In *AAAI/ACM conference on artificial intelligence, ethics and society*.

Hodges, Andrew. 2008. What did Alan Turing mean by "machine"? In *The mechanical mind in history*, eds. Philip Husbands, Owen Holland, and Michael Wheeler, 75–90. MIT Press.

Ienca, Marcello. 2017. Cognitive technology and human-machine interaction: The contribution of externalism to the theoretical foundations of machine and cyborg ethics. *Annals of the University of Bucharest (Philosophy Series)* 66 (2): 91–115.

Illies, Christian, and Anthonie Meijers. 2009. Artefacts without agency. *The Monist* 92 (3): 420–440.

Illies, Christian F. R., and Anthonie Meijers. 2014. Artefacts, agency, and action schemes. In *The moral status of technical artefacts*, eds. Peter Kroes and Peter-Paul Verbeek, 159–184. Springer.

Innis, Harold A. 1950. *Empire and communications*. Oxford University Press.

James, William. 1918. *The principles of psychology*. Henry Holt and Company.

Jaques, Elliot. 1978. *Levels of abstraction in logic and human action: A theory of discontinuity in the structure of mathematical logic, psychological behaviour, and social organization*. Heinemann.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. Artificial intelligence: the global landscape of ethics guidelines, Technical report, arXiv preprint arXiv:1906.11668.

Jonsen, Albert R., and Stephen Edelston Toulmin. 1988. *The abuse of casuistry: A history of moral reasoning*. Univ of California Press.

Jørgensen, Rikke Frank. 2017. What platforms mean when they talk about human rights. *Policy & Internet* 9 (3): 280–296.

Kasson, John F. 1999. *Civilising the machine: Technology and republican values in America, 1776–1900*. Hill and Wang.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133 (1): 237–293.

Kranzberg, Melvin. 1986. Technology and history: "Kranzberg's laws". *Technology and Culture* 27 (3): 544–560.

Latour, Bruno. 1992. Where are the missing masses? The sociology of a few mundane artifacts. In *Shaping technology/building society: Studies in sociotechnical change*, eds. Wiebe E. Bijker and John Law, 225–258. MIT Press.

Latour, Bruno. 1999. *Pandora's hope: essays on the reality of science studies*. Harvard university press.

Latour, Bruno. 2002. Morality and technology: The end of means. *Theory, Culture and Society* 19 (5/6): 247–260.

Lawson, Clive. 2008. An ontology of technology: Artefacts, relations and functions. *Techné* 12 (1): 48–64.

Lawson, Clive. 2010. Technology and the extension of human capabilities. *Journal for the Theory of Social Behaviour* 40 (2): 207–223.

Layton, Edwin T. 1974. Technology as knowledge. *Technology and Culture* 15 (1): 31–41.

Levinson, Paul. 1982. What technology can teach philosophy: Ruminations along Kantian/Popperian lines. In *In pursuit of truth: Essays on the philosophy of Karl Popper on the occasion of his 80th birthday*, ed. Paul Levinson, 157–175. Humanities Press.

Levinson, Paul. 1988. *Mind at large: Knowing in the technological age*. JAI Press.

Logan, Robert K. 2004. *The alphabet effect: A media ecology understanding of the making of western civilization*. Hampton Press.

Logan, Robert K. 2007. *The extended mind: The emergence of language, the human mind, and culture*. University of Toronto Press.

Mackenzie, Catriona, and Natalie Stoljar, eds. 2000. *Relational autonomy: Feminist perspectives on autonomy, agency, and the social self*. Oxford University Press.

Malafouris, Lambros. 2013. *How things shape the mind: A theory of material engagement*. MIT Press.

McLuhan, Marshall. 2011. *The Gutenberg galaxy: The making of typographic man*. University of Toronto Press.

McLuhan, Marshall, and Robert K. Logan. 1977. Alphabet, mother of invention. *Et Cetera: A Review of General Semantics* 34 (4): 373–383.

McNeil, Ian. 1990. Basic tools, devices, and mechanisms. In *An encyclopedia of the history of technology*, 1–44. Routledge.

McQuillan, Dan. 2016. Algorithmic paranoia and the convivial alternative. *Big Data and Society*.

Meng, Xiao-Li. 2018. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics* 12 (2): 685–726.

Midgley, Mary. 1994. *The ethical primate: Humans, freedom and morality*. Routeledge.

Mindell, David A. 2015. *Our robots, ourselves: Robotics and the myths of autonomy*. Viking.

Mitcham, Carl. 2014. Agency in humans and in artifacts: A contested discourse. In *The moral status of technical artefacts*. Springer. Chap. 2.

Mokyr, Joel. 1990. *The lever of riches: Technology, creativity and economic progress*. Oxford University Press.

Mumford, Lewis. 1944. *The condition of man*. Martin Secker and Warburg.

Mumford, Lewis. 1964. Authoritarian and democratic technics. *Technology and Culture* 5 (1): 1–8.

Nadler, Amos, and Paul J. Zak. 2016. Hormones and economic decisions. In *Neuroeconomics*, eds. Martin Reuter and Christian Montag, 41–66. Springer.

Nagel, Saskia K. 2013. Autonomy — a genuinely gradual phenomenon. *AJOB Neuroscience* 4 (4): 60–61.

Niker, Fay, Peter B. Reiner, and Gidon Felsen. 2018. Perceptions of undue influence shed light on the folk conception of autonomy. *Frontiers in Psychology* 9: 1400.

Nisbett, Richard E. 2003. *The geography of thought: How asians and westerners think differently . . . and why*. The Free Press.

Nisbett, Richard E., and Timothy D Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review* 84 (3): 231–259.

Nofre, David, Mark Priestley, and Gerard Alberts. 2014. When technology became language: The origins of the linguistic conception of computer programming, 1950–1960. *Technology and Culture* 55 (1): 40–75.

Oakley, Kenneth P. 1965. *Man the tool-maker*, 5th edn. Trustees of the British Museum (Natural History).

Ofstad, Harald. 1961. *An inquiry into the freedom of decision*. Norwegian University Press.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (13): 1–33.

Ong, Walter J. 2012. *Orality and literacy: The technologizing of the word (30th anniversary edition)*. Routledge.

Parker, Jack, and David Danks. 2019. How technological advances can reveal rights. In *AAAI/ACM conference on artificial intelligence, ethics and society*.

Peterson, Martin, and Andreas Spahn. 2011. Can technological artefacts be moral agents? *Science and Engineering Ethics* 17 (3): 411–424.

Pierce, Charles Saunders. 1958. *Collected papers*. Harvard University press.

Pollack, Norman. 1962. *The populist response to industrial America: Midwestern populist thought*. Harvard University Press.

Powles, Julia, and Helen Nissenbaum. 2018. The seductive diversion of 'solving' bias in artificial intelligence. *Medium* 7th December.

Reboul, Anne. 2015. Why language really is not a communication system: a cognitive view of language evolution. *Frontiers in Psychology* 6 (1434): 1–12.

Rheingold, Howard. 2000. *Tools for thought: The history and future of mind-expanding technology*. MIT Press.

Robinson, William S. 1950. Ecological correlations and the behaviour of individuals. *American Sociological Review* 15 (3): 351–357.

Rosch, Eleanor. 1978. Principles of categorization. In *Cognition and categorization*, eds. Eleanor Rosch and Barbara B. Lloyd, 27–48. Lawrence Erlbaum Associates.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology* 8: 382–439.

Sapolsky, Robert M. 2017. *Behave: The biology of humans at our best and worst*. Penguin.

Schneewind, Jerome B. 1998. *The invention of autonomy*. Cambridge University Press.

Schneewind, Jerome B. 2013. Autonomy after Kant. In *Kant on moral autonomy*, ed. Oliver Sensen, 146–168. Cambridge University Press.

Skagestad, Peter. 1993. Thinking with machines: Intelligence augmentation, evolutionary epistemology and semiotic. *Journal of Social and Evolutionary Systems* 16 (2): 157–180.

Skagestad, Peter. 1996. The mind's machines: The Turing machine, the Memex, and the personal computer. *Semiotica* 111 (3/4): 217–243.

Sproat, Richard. 2010. *Language, technology, and society*. Oxford University Press.

Sternberg, Robert J., and David D. Preiss, eds. 2005. *Intelligence and technology: The impact of tools on the nature and development of human abilities*. Lawrence Erlbaum Associates.

Strathern, Marilyn. 1998. Social relations and the idea of externality. In *Cognition and material culture: the archeology of symbolic storage*, eds. Colin Renfrew and Chris Scarre, 135–147. McDonald Institute for Archeological Research.

Susser, Daniel. 2019. Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures. In *AAAI/ACM conference on artificial intelligence, ethics and society*.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2018. Online manipulation: Hidden influences in a digital world. Preprint, available at SSRN 3306006.

Taylor, Simon. 2019. France bans data analytics related to judges' rulings. *Legal Week (4 June)*.

Turkle, Sherry. 2005. *The second self: Computers and the human spirit (twentieth anniversary edition)*. MIT Press.

Tversky, Amos, and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185 (4157): 1124–1131.

Verbeek, Peter-Paul. 2008a. Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and the Cognitive Sciences* 7: 387–395.

Verbeek, Peter-Paul. 2008b. Morality in design: Design ethics and the morality of technological artifacts. In *Philosophy and design*, 91–103. Springer.

Verbeek, Peter-Paul. 2011. *Moralizing technology: Understanding and designing the morality of things*. The University of Chicago Press.

Verbeek, Peter-Paul. 2014. Some misunderstandings about the moral significance of technology. In *The moral status of technical artefacts*, eds. Peter Kroes and Peter-Paul Verbeek, 75–88. Springer.

Villlarán, Alonso. 2017. Irrational advertising and moral autonomy. *Journal of Business Ethics* 144 (3): 479–490.

Vold, Karina, and Jess Whittlestone. 2019. Privacy, autonomy, and personalised targeting: Rethinking how personal data is used. Preprint, Cambridge University.

Vonk, Jennifer, and Suzanne E. MacDonald. 2004. Levels of abstraction in orangutan (*Pongo abelii*) categorization. *Journal of Comparative Psychology* 118 (1): 3–13.

Wald, Abraham. 1947. Foundations of a general theory of sequential decision functions. *Econometrica* 15 (4): 279.

Walker, Tom. 2011. Full autonomy, substantial autonomy, and neuroscience. *AJOB Neuroscience* 2 (3): 56–57.

Washburn, Sherwood L. 1959. Speculations on the interrelations of the history of tools and biological evolution. *Human Biology* 31 (1): 21–31.

Whitehead, Alfred North. 1911. *An introduction to mathematics*. Williams and Norgate.

Wierzbicka, Anna. 2006. *English: Meaning and culture*. Oxford University Press.

Williamson, Robert C., Michelle Nic Ragnhaill, Kirsty Douglas, and Dana Sanchez. 2015. *Technology and Australia's future: New technologies and their role in Australia's security, cultural, democratic, social and economic systems*. Australian Council of Learned Academies.

Wilson, Timothy D. 2002. *Strangers to ourselves: Discovering the adaptive unconscious*. Harvard University Press.

Winner, Langdon. 1977. *Autonomous technology: Technics-out-of-control as a theme in political thought*. MIT Press.

Winner, Langdon. 1980. Do artifacts have politics? *Daedalus* 109 (1): 121–136.

Woolgar, Steve, and Geoff Cooper. 1999. Do artefacts have ambivalence? Moses' bridges, Winner's bridges and other urban legends in S&TS. *Social Studies of Science* 29 (3): 433–449.

Yoder, Keith J., and Jean Decety. 2018. The neuroscience of morality and social decision-making. *Psychology, Crime and Law* 24 (3): 279–295.