Final draft of chapter to appear in: Dubber, M., Pasquale, F. and Das, S. (eds.), Oxford Handbook of Ethics of AI, Oxford University Press, New York, reproduced by permission of Oxford University Press: https://global.oup.com/academic/product/the-oxford-handbook-of-ethics-of-ai-9780190067397?cc=gb&lang=en&

Autonomy

Michael Wheeler

Abstract

Unease regarding autonomous (self-governing) AI is most vividly expressed in the vision of an artificial super intelligence whose self-generated goals and interests diverge radically from those of humankind, and which thus places our well-being, and maybe even our survival, at risk. The first question addressed by this chapter, then, is this: what are the conditions that would need to be met by an intelligent machine, in order for that machine to exhibit the kind of autonomy that is operative in this dystopian scenario? However, there is arguably a more pressing concern regarding a different class of AI systems, those that are autonomous in only the milder sense that, in their domains of operation, we are ceding, or will cede, some significant degree of control to them. Systems of this kind include self-driving cars and autonomous weapons systems. The second question addressed by this chapter, then, is this: are these already-in-the-world autonomous AI systems a genuine cause for concern? A key issue here concerns the properties of so-called deep learning networks. The chapter ends by suggesting briefly that the two kinds of autonomy discussed are connected in an interesting way.

Keywords: autonomous AI, autonomous weapons systems, control, deep learning, self-driving cars.

Introduction

There are many ethical challenges in the vicinity of AI, but perhaps our greatest anxieties concern *autonomous AI* – AI that is, in some relevant sense, *self-governing*. In their most extreme form, these anxieties are most vividly expressed in the prediction that humankind will soon share the planet with an autonomous artificial super intelligence whose self-generated goals and interests diverge radically from our own. As a result of this divergence, so the prediction goes, there is a palpable risk that this machine will exercise its autonomy in ways that are detrimental to our well-being or survival. Such visions of a not-too-distant future populated by at least one super-intelligent machine with malicious intentions (or maybe just intentions in which our well-being simply doesn't figure) will no doubt strike some readers as a disturbing specification of a clear and credible danger in need of urgent consideration by a robustly funded international task force, while it will strike others as pure science fiction in need of nothing more expensive than a healthy dose of technical reality. The truth is almost certainly somewhere in between, which is surely enough to make the issue worthy of consideration.

In light of the foregoing, it seems that one important question we might ask is this: what are the conditions that would need to be met by an intelligent machine, in order for that machine to exhibit the kind or degree of autonomy that is operative in our dystopian scenario? The guiding intuition here is that it is only when a machine is a fully autonomous agent that the threats in question arise, so it makes sense to have ways of determining if and when that point

has been reached. After all, understanding what the bar is for artificial autonomy may help us to decide how worried we should be. In what follows, then, an attempt will be made to bring the notion of autonomy at issue so far into better view.

That said, there is arguably a more pressing concern regarding a different notion of autonomous AI. Recent years have witnessed enormous advances in areas such as machine learning, sensor technology, and robotics. Indeed, it seems that we are already building, or are on the verge of building, AI systems that, although they may fail to exhibit autonomy in any metaphysically demanding sense, are self-governing in the milder sense that, in their domains of operation, we are ceding, or will cede, some significant degree of control to them. Existing and imminent examples of systems of this kind (some of which are discussed below) include weapons, vehicles, financial management applications, and medical assistants that have been AI-enhanced so as to take control in some sphere of intelligent, often life-critical, action. So, one might reasonably be moved by the thought that debates about what are (at present anyway) mere thought experiments should take a back seat to debates about the nature and implications of real AI systems, embedded in the actual world, that are, or soon will be, taking important decisions, sometimes with profound consequences, on our behalf.[1]

Given all this, the following treatment of autonomous AI will focus not only on autonomy as it figures in relation to some future, post-singularity dystopia, but also on autonomy as it figures in contemporary, concrete AI systems taking sensitive decisions for us in the wild, a state of affairs that may itself be a legitimate cause for concern. There will, however, be a twist in our tale, since, as we shall see, the two kinds of autonomy are actually connected in an interesting way.

Autonomy and Control

An autonomous entity is an entity that has the capacity for self-governance, in some relevant sense of that term. Understood as such, the notion of autonomy looms large in many debates of ethical and political importance, debates over, for example, the aspirations of particular counties or regions to be constitutionally independent from existing external power structures, the rights of patients to make informed and uncoerced decisions about medical treatments, and the ideal of living a maximally authentic life free from manipulating or self-distorting influences. Examples could be multiplied indefinitely, and, in different contexts, different aspects of what matters for or about autonomy will come to the fore. Given this kaleidoscope of issues and problems, it is worth homing in on one's target domain to highlight the concepts or principles that have local currency. Thus, we can begin by noting that when the topic is the autonomy of machines, or, more generally, autonomy in a mechanistic universe, the notion that, it might reasonably be said, defines the territory is that of *control*. Thus, in this machine-related context, control is what we mean by governance (consider the Watt governor, a device for controlling the speed of a steam engine), and self-governance is control over oneself, or some relevant aspect of one's activity.

The idea that the concept of control is central to the appropriate understanding of autonomy has what we might think of as a negative justification and a positive one. Let's take the

[1] See, e.g.: David A. Mindell, *Our Robots, Ourselves: Robotics and the Myths of Autonomy* (New York: Penguin, 2015); Filippo Santoni de Sio and Jeroen van den Hoven, 'Meaningful Human Control over Autonomous Systems: A Philosophical Account,' *Frontiers in Robotics and AI* 5 (2018): 15.

negative one first. What is it to *lack* autonomy? It is, it seems, to lack control over one's own behaviour or, on a larger scale, over one's destiny. To a first approximation, then, a non-autonomous entity is one whose behaviour or destiny is controlled by external causal forces. Thus an autonomous entity is one which is in control of its own behaviour or destiny. This is only a first approximation, because there remain intricate matters of detail. For example, as Dennett points out during his classic discussion of control in relation to free will (a notion that is, of course, conceptually intertwined with that of autonomy), when one is in control of something, including oneself, one doesn't achieve that feat by controlling all the causal forces that act on that thing.[2] In other words, I may rightly be said to be in control of my physical actions, even though those actions are constrained and shaped by factors such as the force of gravity, the ambient temperature, and the strength of the wind. Indeed, a skilled soccer player with enough weather-related information may anticipate, accommodate, and maybe even exploit the wind – an external, active factor that is beyond his control – in order to score from a majestic, and thus beautifully under control, free kick. There are other subtleties: one can sometimes control a self-controlling entity, without thereby undermining that entity's basic claim to autonomy, by controlling the external factors that, via its own self-controlling mechanisms, cause it to act in certain predictable ways[3]; there are circumstances under which any sensible autonomous agent should, in *a* sense of control, want to be controlled by external factors, such as when imminent danger results in an agent adopting avoidance behaviour in a purely reactive, stimulus-response (but thereby appropriately speedy) manner (cf. Dennett's discussion of Skinnerian control[4]); and sometimes, in an act of what we might call meta-autonomy, it is rational (e.g. to meet time-constraints or to avoid being overly predictable to a competing self-controlling agent) for an agent deliberately to give up control, often to practical randomness, in order to achieve a desired outcome, such as when a coin or racquet is flipped to determine who will serve or receive first in tennis (Dennett identifies similar and more complex cases[5]). All of these niceties – and many others besides – would need to be sorted out, but let's write a philosophical blank cheque to those who would complete the hard thinking here (Dennett does more than make a start) and agree that compromised autonomy is, among other things, a matter of compromised control.

The positive justification for the intimate connection between autonomy and control comes from the thought that we can exploit the notion of different aspects of control not only to make sense of the idea that autonomy is a graded quality, rather than a binary, 'all or nothing' property, but also to carve out a notion of autonomy that applies to machines and mechanisms. In the context of the present treatment, it is the latter result that most obviously concerns us, since it is of direct significance to our understanding of autonomous AI. In other contexts, however, the same idea might be developed to ground the claim that human beings are biological machines whose autonomy is founded on the operation of biological/psychological mechanisms, a view whose most prominent manifestation in philosophy and psychology conceives of the human mind as an integrated set of neurally realized computational processes.

To illustrate the way in which a framework involving different aspects of control might be used to build an account of autonomy in the realm of the artificial, we can build on an

---

[2] Daniel C. Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, Mass.: MIT Press, 1984, especially chapter 3).
[3] Dennett, *Elbow Room*, 56.
[4] Dennett, *Elbow Room*, 57-8.
[5] Dennett, *Elbow Room*, 67.

analysis due to Boden.[6] Inspired by work in both AI and artificial life (ALife – the construction and study of artificial systems that exhibit various features characteristic of biological systems), Boden draws a distinction between three different aspects of control that (she suggests) are crucial to the possession of autonomy. The first is the extent to which the behaviour of an agent is governed not by inner mechanisms that respond to environmental triggers in ways that were programmed into the agent at 'birth', but by mechanisms that have been shaped by that agent's own past experience of the world. Boden's thought here is something like this: intra-lifetime learning matters for autonomy, at least because, given an agential capacity to learn, different historical paths of learning will produce agents that possess 'individuality', in the sense that the behavioural response of any two such agents to the same environmental variable may differ. Under such circumstances, it is not merely the present state of the environment plus some 'innate' (unlearned, preprogrammed) mechanical set-up shared by an entire group of agents that determines the behaviour of some particular agent, but the present state of the environment *plus individual experiential history*, a history during which a suite of shared, 'innately specified' learning mechanisms will have modified that agent's inner mechanical set-up so as to produce a behavioural profile that may well differ from that of an 'innately' identical agent with a different experiential history. Of course, the area of AI known as machine learning, from classical induction systems such as ID3 and AQ11, to traditional connectionist approaches in unsupervised and supervised learning, to recent successes in Bayesian inference and so-called deep learning, provides a rich suite of ways in which such adaptive inner modifications to individual experiential histories may be realized.

The second autonomy-critical aspect of control that Boden identifies is the extent to which the behaviour-directing mechanisms at work are self-generated by the agent in question, rather than imposed by external design. As Boden herself notes, this may initially look like a repeat of the point about learning. However, the appeal to self-generation is designed to invite a different observation, namely that the behaviour of some systems is the product of *emergent self-organization*. To explain: A self-organizing system is one in which certain intra-systemic components, on the basis of purely local rules (i.e. without the direction of some global executive control process), interact with each other in nonlinear ways so as to produce the emergence and maintenance of structured global order. Self-organization is now recognized as being a widespread phenomenon in nature. Regularly cited examples in the literature include the Beloussov-Zhabotinsky chemical reaction, slime moulds, foraging by ants, and flocking behaviour in creatures such as birds. The final example is instructive, because, as it happens, our scientific understanding of flocking was arguably enhanced by a computer simulation due to Reynolds[7], a simulation that has been enormously influential in the ALife community. In this system, adaptive flocking behaviour (e.g. flocks that maintained their integrity while navigating obstacles) emerged from an arrangement in which individual virtual birds each followed just three simple, purely local rules. These rules are imperfectly but intuitively captured by the following ordinary language paraphrases: don't get too close to other the birds around you, don't get too far away from them, and move at roughly the same speed as them. Of course, since, as we have just seen, self-organization is exhibited by all kinds of systems, its presence is certainly not sufficient for autonomy in the agent-centric sense we require. Nevertheless, applying the concept in this context – and more

[6] Margaret A. Boden, 'Autonomy and Artificiality,' in Margaret A. Boden, ed., *The Philosophy of Artificial Life* (Oxford: OUP, 1996), 95-107.
[7] Craig W. Reynolds, 'Flocks, Herds, and Schools: A Distributed Behavioral Model,' *Computer Graphics* 21:4 (1987): 25-34.

specifically within hierarchies of emergent behaviour-directing mechanisms, in which higher layers of self-organization are generated on the basis of primitives which are in fact emergent structures from the lower levels[8] – gives us another way to make sense of the idea that a purely mechanistic system might exhibit behaviour that is not environmentally determined (which here includes the idea of being essentially prefigured in an externally designed executive program), but rather generated by the agent itself.[9]

Boden's third autonomy-critical aspect of control is the extent to which an agent's behaviour-directing mechanisms may be reflected upon and selectively modified by that agent, so as to explore and transform, in a self-governed fashion, the conceptual spaces of thought and action. The paradigm cases of such deliberate inner modification by an agent of its own mechanisms are episodes of conscious thought in human beings in which 'higher' levels of processing access and amend states and processes occurring at 'lower' levels. It is at least arguable that, in AI, the best models we have for such reflective processing still hail from classical AI. These are models marked out by their deployment of explicit, language-like rules and representations that are algorithmically manipulated in ways that are often inspired by human introspection.[10]

For Boden, then, when we ask whether an entity is autonomous, we should ask whether its behaviour-directing mechanisms (i) may be shaped by the entity's experiential history, (ii) are emergent in nature, and (iii) are reflectively modifiable by that entity. All of these control-related properties are realizable in the realm of the artificial. Indeed, their status as autonomy-relevant is inspired precisely by a consideration of achievements in that domain. Moreover, they are to be conceived as defining something like a three-dimensional coordinate system that gives an entity a position in what we might call 'autonomy space'. The higher the values on the different axes, the more autonomous an entity is. And that's what delivers the idea that autonomy is a graded, rather than a binary (on or off), phenomenon. As Boden puts it, '[a]n individual's autonomy is the greater, the more its behaviour is directed by self-generated (and idiosyncratic) inner mechanisms, nicely responsive to the specific problem-situation, yet reflexively modifiable by wider concerns'.[11]

---

[8] Boden, 'Autonomy and Artificiality,' 103.

[9] Although Boden doesn't pursue this thought, a more formal relationship between self-organization and autonomy may be found in the theoretical framework provided by autopoiesis, a framework that has been influential in the field of ALife. According to this framework, a self-organizing system counts as autonomous if it is a network of interdependent processes whose recurrent activity (a) produces and maintains the very boundary that determines the identity of that network as a unitary system), and (b) defines the ways in which that system may encounter perturbations from what is outside it while maintaining its organization and thus its viability (see e.g. Francisco J. Varela, *Principles of Biological Autonomy* (New York: Elsevier North Holland, 1979); for useful discussion, see Xabier E. Barandiaran, 'Autonomy and Enactivism: Towards a Theory of Sensorimotor Autonomous Agency,' *Topoi* 36 (2017): 409-430). Of course, the connection between this technical notion of autonomy and the more common usage in ethics would need to be worked out. For a related development, see Gunther Teubner, *Law as an Autopoietic System.* (Oxford/Cambridge: Blackwell, 1993).

[10] Boden, 'Autonomy and Artificiality,' 105.

[11] Boden, 'Autonomy and Artificiality,' 102.

Boden's analysis of autonomy, as useful as it is, will not take us all the way to what we need. Recall that our first aim in this chapter is to bring into better view the conditions that would need to be met by a machine, in order for that machine to exhibit the kind or degree of autonomy that might make us take seriously a vision in which an autonomous artificial super intelligence whose self-generated goals and interests diverge radically from our own exercises that autonomy in ways that are detrimental to our well-being or survival. In light of this goal, Boden's account is productive in that it succeeds in characterizing a robust sense of agential autonomy in such a way that we can see that phenomenon as being built from, or emerging out of, purely mechanistic processes. However, even though, by emphasizing distinctive learning histories, it hints at the presence of a self-spawned life-plan structured by idiosyncratic goals and desires, and even though, by stressing the reflective modification of behaviour-directing mechanisms, it almost points us in the direction of a self-modifiable individual world-view, it fails adequately to foreground, or to account for, the demand that a fully autonomous agent must be able to arrive at its own life-plan and then adaptively modify that plan in light of experiences and evidence.[12] And those capacities, one might reasonably think, will need to be found in our artificial super intelligence, if the apocalyptic scenario is to look plausible. So, can such capacities be delivered by additional, purely mechanistic, control-related features, thus making available new dimensions and higher points in our autonomy space?

Some of the questions waiting in the wings here present formidable philosophical challenges. For example, what establishes that a life-plan is the *agent's own*? The answer to this question presumably requires an account of cognitive ownership (for one such account, see Rowlands[13]) and thus of the self. And is consciousness, or self-consciousness, required for adaptive life-planning? In the present context, this raises the issue of whether artificial consciousness is possible[14] and so might be an invitation to the recalcitrant *hard problem of consciousness* (the problem of explaining why any purely physical system is conscious rather than non-conscious).[15] Some commentators might take comfort in the fact that these are long-standing, deeply perplexing puzzles, which might make it seem as if fully autonomous AI remains a long way off. However, one should not underestimate the power of science to chip away at such recalcitrant problems. For example, a common thought in philosophical discussions of autonomy is that each autonomous agent possesses a set of so-called 'pro-attitudes' (roughly, higher-order desires, values and beliefs that record approval, admiration, or preference towards things) that governs its approach to, and its engagement with, the world. This set of pro-attitudes is often taken to define in part what is meant by 'the self'.[16] Moreover, a fully autonomous agent will be able to incorporate new pro-attitudes (beliefs, desires, values) into its governing set, on the basis of its unfolding experience and evidence. And this capacity for pro-attitude maintenance and revision is also a pivotal aspect of

---

[12] Steven Weimer, 'Evidence-Responsiveness and Autonomy,' *Ethical Theory and Moral Practice* 16 (2013): 621–642.
[13] Mark Rowlands, *The New Science of the Mind: from Extended Mind to Embodied Phenomenology* (Cambridge, MA: MIT Press, 2010).
[14] Ronald L. Chrisley, 'Philosophical, Foundations of Artificial Consciousness,' *Artificial Intelligence in Medicine* 44 (2008): 119-137
[15] David J. Chalmers, 'Facing up to the Problem of Consciousness,' *Journal of Consciousness Studies* 2 (1995): 200-19.
[16] Fay Niker, Peter B. Reiner, and Gidon Felsen, 'Updating our Selves: Synthesizing Philosophical and Neurobiological Perspectives on Incorporating New Information into our Worldview,' *Neuroethics* 11 (2018): 273–282.

autonomy, since the agent's goal in that activity will be to plan its life in accordance with its pro-attitudes. So, rather than ask directly whether an AI system could adaptively modify a life-plan in light of experience and evidence, we can ask the related, perhaps less daunting, question of whether an AI system could incorporate new pro-attitudes (beliefs, desires, values) into its behaviour-governing set in light of experience and evidence. Drawing on recent work in neuroscience, Niker et al.[17] argue that the latter feat may be achieved by a specific kind of computational mechanism in the brain, one that works according to principles of Bayesian inference that tell us how to update the probabilities of prior beliefs (or other attitudes, thought of as hypotheses) given evidence. Of course, Bayesian inference techniques are an established and a long-standing part of the AI toolkit (e.g. in pattern recognition and machine learning). Indeed, at least some of their popularity in neuroscience can be traced to their success in AI.

If autonomy is a graded phenomenon, characterizable in terms of different varieties or levels of mechanizable control that eventually top-out in full autonomy of the kind required by our (thankfully still fictional) super-intelligent AI, then, in principle, we have both a road map to such autonomy in the realm of the artificial and a way of recognizing how far down that road we have travelled. In the next section we shall turn our attention to concerns that arise even at the early twists and turns in that road, at points where, even though the target AI system is not at the partially scoped-out level of full autonomy, nevertheless we have ceded control to that system in some potentially sensitive or safety-critical, in-the-wild scenario.

Relinquishing Control

The commercial peer-to-peer ride-sharing business, Uber, began testing self-driving cars on the roads of Arizona in February 2017. In March 2019, in Tempe, an Uber-owned self-driving car, travelling in autonomous mode (although with a safety driver on board), struck and killed a pedestrian crossing the road at an unauthorized point. The preliminary report from the US National Transportation Safety Board suggested that after detecting the victim six seconds before impact, the controlling software struggled with ambiguity in the perceptual input, first identifying the pedestrian as an unknown object, then as a vehicle, and then as a bicycle. (She was pushing a bicycle at the time.) About one second before impact, the vehicle made the decision that emergency braking was required, but no emergency auto-braking system was available. This was not a malfunction. The engineers had been concerned that a self-driving car with an active autonomous emergency braking system would be at risk of behaving in unexpected, erratic, and thus potentially dangerous, ways, as a result of that system repeatedly being triggered unnecessarily by 'false-positives' (such as mistaking a pedestrian standing harmlessly on the sidewalk for one about to jump into the road). Moreover, Uber had turned off the car's off-the-production-line automatic emergency braking system so that there would be no conflicts between the two kinds of technology. Following the tragic incident in Arizona, Uber immediately implemented a temporary suspension of its self-driving car operations on public roads, in order to revisit its safety protocols.[18]

---

[17] Niker et al., 'Updating our Selves'.

[18] See https://www.theverge.com/2018/3/19/17139518/uber-self-driving-car-fatal-crash-tempe-arizona, https://www.theverge.com/2018/7/3/17530232/self-driving-ai-winter-full-autonomy-waymo-tesla-uber, https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/, and Uber's video 'Self-Driving Cars Return to Pittsburgh Roads,' reporting on

The foregoing example graphically exposes a rather obvious, but nevertheless worth-stating, dilemma regarding self-driving cars. On the one hand, the whole point of such vehicles is that they, well, drive themselves, which includes making identifications, categorizations, and decisions about what the environmental circumstances are, as well as determining what actions are appropriate. To the extent that we resist ceding this sort of control to the technology – to the extent that, for example, the vehicle is required to seek input from a human operative, whether on-board or remote, before it categorizes or acts – it simply isn't autonomous, in any reasonable sense of the term, and that not only defeats the object of the exercise, it prevents us from reaping the benefits of the technological advances in play. And, of course, there is plenty of evidence that runs counter to the Arizona tragedy – evidence that we might expect to be tabled by certain interested parties – citing the overall safety record of self-driving cars, alongside statistics that emphasize the prevalence of human error in road accidents.[19] On the other hand, to the extent that we do cede control to the technology, we inherit a range of safety-critical risks that pose some difficult ethical problems, as well as technical and legal challenges. For example, one of our instincts when things go wrong is to wonder who, if anyone, should be blamed. But, in the case of self-driving cars, that's not a straightforward matter. The car itself cannot be held responsible (given the lower-grade kind of autonomy it enjoys, it's simply not a blameworthy a moral agent), so maybe our ethical attention should be focused on the owning company, the designers, developers or engineers, or the safety driver (where there is one – the autonomous vehicle gold standard is surely to do away with such individuals altogether). For present purposes, the point here is not to choose among the candidates for responsibility – no doubt all kinds of context-dependent complexities mean that no universal principle or policy will work – but rather to register the higher-order point that relinquishing control or not relinquishing control look like all the available options, and each has its drawbacks. What do we do?

Before saying something by way of a response, we should remind ourselves that self-driving cars are not the only on-the-cards technological innovations that raise ethical questions in the vicinity of our milder form of autonomy. We could raise a similar or related dilemma regarding robot surgeons. On average, such systems will quite likely perform more accurate surgical movements while navigating and reasoning in enormous, multi-dimensional, patient-related data spaces in a manner that is safer and speedier than human surgeons. If this prediction were to be confirmed, it would provide positive evidence that we should cede control to such systems. After all, surely we all want a healthier population maintained by more efficient medical delivery. But then it's hard to eliminate the now-familiar nagging concerns about moral responsibility and legal accountability, and so our dilemma returns.

Things might seem rather graver in another context for decision-making by mildly autonomous AI, a context in which although our highlighted ethical dilemma could certainly be stated in the abstract, the cynics among us might wonder whether it constitutes a genuine socio-political choice, given where the power in our societies ultimately lies. Thus consider autonomous weapons systems – weapons systems that, 'once activated, [will] select and

---

'months of reflection and improvement' following the Arizona incident https://www.youtube.com/watch?v=0E5IQJj_oKY. All last accessed June 30 2019.

[19] See e.g. the 2018 safety report by Google-owned Waymo, https://storage.googleapis.com/sdc-prod/v1/safety-report/Safety%20Report%202018.pdf, and the aforementioned Uber 'Self-Driving Cars Return to Pittsburgh Roads' video, both last accessed June 30 2019

engage targets without further intervention by a human operator' (US Department of Defence directive 2012, updated 2017[20]). This sort of autonomous AI will be charged with deciding routinely (not just in emergency situations) whether to take human lives. Predictably, then, the development and deployment of such systems have been subject to widespread criticism, leading to demands for a proper international framework for ethical design and regulation (see, for example, 2017's 'Open Letter to the United Nations Convention on Certain Conventional Weapons', signed by the leading technology entrepreneur Elon Musk and over 100 other CEOs of technology companies, calling for the UN structures to find a way to protect us all from the dangers of lethal autonomous weapons systems[21]).

In the academic and public debate, a range of arguments against autonomous weapons systems have been lodged. These include, but are not limited to, the following:

> Extant and imminent instances of such weapons will not be sophisticated enough to allow those systems to follow international humanitarian law – the legal principles of armed conflict designed to protect civilians which turn on delicate and complex, judgment-laden notions such as a distinction between combatants and non-combatants, proportionality in the use of force, and a sense of what is necessary from a military perspective.[22]
>
> Accountability is compromised, in that it is unclear who to blame for any unnecessary casualties resulting from the decisions of autonomous weapons, and more specifically it becomes harder to regard military personnel as morally or legally responsible for the relevant war crimes.[23] (Cf. the similar worry raised earlier in the case of self-driving cars.)
>
> Because an inanimate AI system will be incapable of genuinely respecting the value of, or understanding the loss of, a human life, allowing such a machine to end a human life is an affront to that person's dignity.[24]

---

[20] Quoted by Amanda Sharkey, 'Autonomous Weapons Systems, Killer Robots and Human Dignity,' *Ethics and Information Technology* 21:2 (2019): 75-87.
[21] https://futureoflife.org/autonomous-weapons-open-letter-2017, last accessed June 30 2019.
[22] Among many others, see: Peter Asaro, 'How Just could a Robot War be?' in Philip Brey, Adam Briggle and Katinka Waelbers, eds., *Current Issues in Computing and Philosophy* (Ios Press, 2008), 50-64; Noel E. Sharkey. 'Death Strikes from the Sky: the Calculus of Proportionality,' *IEEE Science and Society* Spring 2009: 16-19; Noel E. Sharkey 'Killing Made Easy: from Joystics to Politics, in Patrick Lin, George Bekey, and Keith Abney, eds., *Robot Ethics: The Ethical and Social Implications of Robotics* (Cambridge, Mass.: MIT Press, 2012), 111-128; Santoni de Sio and van den Hoven, 'Meaningful Human Control over Autonomous Systems'. For a more optimistic assessment of what autonomous weapons systems might achieve in this area, see Ronald C. Arkin, 'The Case for Ethical Autonomy in Unmanned Systems,' *Journal of Military Ethics*, 9:4 (2010), 332-341.
[23] Again, among many others, see: Robert Sparrow, 'Killer Robots,' *Journal of Applied Philosophy* 24:1 (2007): 62-77; Sharkey 'Killing Made Easy'; Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, A/HRC/23/47* (New York: United Nations, 2013).
[24] Yet again, among many others, see: Bonnie Docherty, *Shaking the Foundations: The Human Rights Implications of Killer Robots*. Human Rights Watch website. (2014) https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-

Once again, then, when confronted by the advent of smart machines that are able to make and execute safety-critical, and sometimes life-critical, decisions for us – perhaps, in spite of us – the relinquishing of control to such machines raises acute ethical challenges. But this time around, the thought that society in general might actually have the power to refuse to allow the military to relinquish control to the autonomous AI systems in question may be essentially chimerical. This would resolve the dilemma accompanying the decision over whether or not to relinquish control, but at an obvious and alarming cost.

Returning to self-driving cars, one response to the ethical problems posed has been to launch a massive on-line research project investigating what people across the world think an autonomous vehicle should do when faced with moral choices.[25] The basis for this research was a well-trodden philosophical thought experiment known as the trolley problem.[26] In this scenario, you are confronted by a runaway trolley and positioned in front of a lever for redirecting that trolley onto a side track. You are presented with, and must select between, different outcomes. For example, it could be set up like this: you could (a) pull the lever to save the lives of five people trapped on the trolley, but you will thereby cause the death of one person trapped on the side track, or (b) not pull the lever and let the five people die, meaning that the single person survives. The permutations, in terms of numbers and who the people are – relations, politicians, children, rich, poor and so on – are limitless, and this has made the trolley problem a popular philosophical tool for exploring moral decision-making. Back in the land of AI, it's not hard to see how the trolley becomes a self-driving car and the lever becomes its programming, hence the empirical study in question.

Here is not the place to explore precisely how the data from the study came out, although it is worth noting that while some universal trends did emerge (e.g. save humans over animals), the participants' judgments were often culture-specific. What we are concerned with here is a more general point. The data gathered would arguably enable the designers of autonomous vehicles to predict what particular communities' responses might be to accidents involving such vehicles. Thus moral decision-making by autonomous vehicles might be tailored to the culture-specific sensitivities at work in a particular region of operation. That sounds like a potentially useful thing to do: self-driving car companies already adapt their vehicles to different (e.g. more or less aggressive) 'driving cultures'. But even if this looks like some sort of progress, critics of autonomous vehicles who are closer to the technical coal face might well be moved to complain that the complex moral trade-offs that trolley-problem-style scenarios introduce are well beyond the capacities of today's self-driving cars, which (those

---

implications-killer-robots, last accessed July 01 2019; Christof Heyns, 'Autonomous Weapons in Armed Conflict and the Right to a Dignified Life: An African Perspective,' *South African Journal on Human Rights*, 33:1 (2017): 46–71. For discussion, see Sharkey, 'Autonomous Weapons Systems, Killer Robots and Human Dignity.'

[25] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan, 'The Moral Machine Experiment' *Nature* 563 (2018): 59–64.

[26] For the classic formulation of the trolley problem, see Philippa Foot, 'The Problem of Abortion and the Doctrine of the Double Effect,' *Oxford Review* 5 (1967): 5-15. For a philosophical discussion of the trolley problem in relation to self-driving cars, see Patrick Lin, 'Why Ethics Matters for Autonomous Cars,' in Markus Maurer, Chris Gerdes, Barbara Lenz, and Hermann Winner, eds., *Autonomous Driving: Technical, Legal and Social Aspects* (Berlin: Springer, 2016), 69-85.

critics will argue) have yet to overcome more basic categorization challenges, as indicated by the Uber vehicle's ultimately tragic struggle to disambiguate its perceptual input (see earlier). The same species of complaint will be lodged against current autonomous weapons systems, thereby bolstering the claim that they are unable to navigate the laws of conflict. Here the critic will be tempted to make reference to an actual AI machine learning system that allegedly misclassified enemy and friendly tanks due to a contingent and irrelevant property of the training set, namely that the training images of enemy tanks mostly featured cloudy skies, while those of friendly tanks mostly featured cloud-free skies. The result was a system that learnt to track the distinction between cloudy and non-cloudy skies, a distinction that, beyond the training set, was not reliably correlated with the difference between enemy and friendly tanks.[27]

In order for us to feel comfortable about relinquishing control to AI systems, it seems necessary (although not sufficient) that the kinds of examples just cited are containable as eliminable edge-cases. And when one is confronted by the recent, undeniably impressive advances in AI, and especially in machine learning, optimism might seem to be the order of the day. Indeed, one might easily come to believe that the road to autonomy is paved with a combination of deep learning and big data.

Deep learning networks typically deploy multi-layered cascades of nonlinear processing units alongside (supervised or unsupervised) machine learning algorithms to perform pattern analysis and classification tasks, by deriving higher level features from lower level features to build hierarchical representations spanning different levels of abstraction. As Metz reports, such systems are 'already pushing their way into real-world applications. Some help drive services inside Google and other Internet giants, helping to identify faces in photos, recognize commands spoken into smartphones, and so much more'.[28] They have famously learnt to play challenging intellectual games to high levels of proficiency, culminating in Google's AlphaGo, a deep-learning-based system for playing the game Go that, in March 2016, recorded a 4-1 victory over Lee Sedol, one of the highest ranked human players in the world. In addition, they are being used to complete life-critical assignments such as detecting earthquakes and predicting heart disease. And, crucially for the present discussion, deep learning networks are central to the control mechanisms that the autonomous AI industries see as pivotal to the eventual success of their products, especially when combined with huge data sets that may be analyzed and navigated by the networks in question to track and reveal task-useful distinctions, patterns, and trends.

So, what is the problem? One issue to note is that, in spite of all the justified enthusiasm about deep learning, there remain barriers to be overcome. For example, and stated in terms of a general tendency, there is a clear sense in which although such networks perform extremely well on specific tasks, no single network performs well across multiple tasks, even within the same general domain. Thus consider a network that must learn multiple classic Atari video games. As a team from Google's DeepMind has shown, it is possible to use the same algorithm, network architecture and hyperparameters to learn 49 such games, retraining

---

[27] Eliezer Yudowsky, 'Artificial Intelligence as a Positive and Negative Factor in Global Risk,' in Nick Bostrom and Milan M. Cirkovic, eds., *Global Catastrophic Risks* (Oxford: OUP, 2006), 308–345.

[28] Cade Metz, 'Google's AI wins Fifth and Final Game against Go Genius Lee Sedol,' *Wired* https://www.wired.com/2016/03/googles-ai-wins-fifth-final-game-go-genius-lee-sedol/, published online March 15 2016, last accessed July 01 2019.

the system from scratch for each new game.[29] What is not yet possible, however, is either for one network to learn all the different games in serial while retaining all its competence, because the process of learning the games one at a time eventually results in the catastrophic forgetting of previous games, or for one network to learn all the different games in parallel, because the different rule-sets interfere with each other. Of course, with a recognition of these limitations in place, there are strategies under development, such as a progressive chaining technique in which separate deep learning systems pass on relevant information to each other to scaffold learning, although this approach eventually runs aground on the intractability of the increasingly large model.[30] The point for us, however, is that it is arguable whether the AI systems on our roads and battlefields, and in our operating theatres, possess the kinds of generalization capacities that they will need, if we are to relinquish control to them.

Moreover – and now we are in the vicinity of the kinds of categorization errors noted earlier – Szegedy et al. have influentially demonstrated that deep learning neural networks are systematically prone to so-called *adversarial exemplars*.[31] Let's consider one of Szegedy et al.'s own examples, a network that had successfully learnt to categorize images into two groups – 'cars' and 'not cars'. The researchers proceeded to systematically generate a range of minutely altered images of cars. The deformations were very small changes made at the pixel-level, meaning that, to the unaided human eye, the new images looked identical to other images to which the network had been exposed, and which it had learnt to categorize correctly as cars. The in-advance prediction would surely have been that the network would correctly classify these altered images as cars. Surprisingly, however, it classified them as non-cars, hence the status of those images as *adversarial* exemplars. Of course, armed with the knowledge that adversarial exemplars exist, designers can systematically generate such items and include them in their networks' training sets. But, especially given finite time constraints, there is surely a danger that the effect of this will be akin to flattening out a lump under a carpet. The lump will simply reappear somewhere else.

The overarching worry, then, is this. Deep learning networks, especially when navigating huge data sets, will no doubt perform ever more impressive feats of reasoning in complex and ethically sensitive domains. Thus we will find ourselves increasingly tempted to cede control to them. But those same networks will sometimes divide up the world in ways that do not coincide with our ways of dividing up the world, meaning that some of their decision-making will be divergent from ours and presumably opaque to us. (What was it about those few pixels that stopped that image being classifiable as a car?) This is troubling, because we have seen that a capacity for reliable categorization – more specifically, the consistent partitioning

---

[29] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis, 'Human-Level Control through Deep Reinforcement Learning,' *Nature* 518 (2015): 529–533.

[30] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell, 'Progressive Neural Networks' (2016) arXiv:1606.04671.

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, 'Intriguing Properties of Neural Networks' (2013) arXiv:1312.6199

of the world into the categories that are ethically relevant for us (e.g. combatants and non-combatants) – is a necessary ability for any AI that is to enjoy even our milder kind of autonomy. The potential existence of unknown adversarial exemplars in the problem spaces in question, as those spaces are partitioned by deep learning networks, should at least make us pause to reflect on how close present AI systems are to meeting this constraint.

A Final Twist

The point at which we relinquish control to AI is the point at which questions regarding our lack of a grip on precisely how certain contemporary AI architectures see the world, and thus on exactly what an autonomous intelligent machine deploying such an architecture in a safety-critical context characterized by uncertainty might do, become prompts for nervous apprehension. The precise path to the alleviation of that concern is not yet clear, but let's finish with a brief, admittedly speculative suggestion that connects the two perspectives on autonomy that have been in view during this chapter.

In many of the ethically challenging scenarios canvassed in the case of autonomous weapons and self-driving cars, one part of the solution may be a machine that has knowledge of the consequences of its actions for sentient beings and is able to reflect on those consequences.[32] This capacity for assessment will be even more likely to prevent unknowing harm if it is deployed by an artificial agent that is able to arrive at its own 'life-plan' and then adaptively modify that plan in light of experiences and evidence. In other words, imbuing AI with the kind of ability that is required for our more demanding, full-strength variety of autonomy may be one way of addressing the concerns that accompany our less demanding, milder variety. Of course, there's a gigantic elephant in the room: what's needed is a fully autonomous artificial agent whose 'life-plan' is shaped not by psychopathic tendencies, but by a demonstrable understanding of, and empathy for, humankind. Some commentators remain sceptical about any such possibility.[33] However, there is a case to be made that, without that achievement in place, autonomy in the realm of the artificial, even in its milder register, is likely to remain a matter of controversy and anxiety.[34]

Bibliography

Colin Allen, Gary Varner, and Jason Zinser, 'Prolegomena to any Future Artificial Moral Agent,' *Journal of Experimental Theoretical Artificial Intelligence* 12 (2000): 251-61.

Ronald C. Arkin, 'The Case for Ethical Autonomy in Unmanned Systems,' *Journal of Military Ethics*, 9:4 (2010), 332-341.

[32] Colin Allen, Gary Varner, and Jason Zinser, 'Prolegomena to any Future Artificial Moral Agent,' *Journal of Experimental Theoretical Artificial Intelligence* 12 (2000): 251-61.
[33] See e.g. Sharkey, 'Autonomous Weapons Systems, Killer Robots and Human Dignity'.
[34] Some short passages of text in this chapter were adapted from Michael Wheeler, 'The Reappearing Tool: Transparency, Smart Technology, and the Extended Mind', *AI and Society*. Published online February 07 2018, https://doi.org/10.1007/s00146-018-0824-x. Many thanks to my student Laurie McMillan who taught to be more optimistic about the possibility of a benign, fully autonomous AI.

Margaret A. Boden, 'Autonomy and Artificiality,' in Margaret A. Boden, ed., *The Philosophy of Artificial Life* (Oxford: OUP, 1996), 95-107.

Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: OUP, 2014).

Daniel C. Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge, Mass.: MIT Press, 1984).

Patrick Lin, 'Why Ethics Matters for Autonomous Cars,' in Markus Maurer, Chris Gerdes, Barbara Lenz, and Hermann Winner, eds., *Autonomous Driving: Technical, Legal and Social Aspects* (Berlin: Springer, 2016), 69-85.

David A. Mindell, *Our Robots, Ourselves: Robotics and the Myths of Autonomy* (New York: Penguin, 2015)

Noel E. Sharkey 'Killing Made Easy: from Joystics to Politics, in Patrick Lin, George Bekey, and Keith Abney, eds., *Robot Ethics: The Ethical and Social Implications of Robotics* (Cambridge, Mass.: MIT Press, 2012), 111-128

Robert Sparrow, 'Killer Robots,' *Journal of Applied Philosophy* 24:1 (2007): 62-77

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, 'Intriguing Properties of Neural Networks' (2013) arXiv:1312.6199