Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Defence, Security and Emergency Response

This input paper was prepared by Seumas Miller

Suggested Citation

Miller, S (2018). Defence, Security and Emergency Response. Input paper for the Horizon Scanning Project "The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing" on behalf of the Australian Council of Learned Academies, www.acola.org.

The views and opinions expressed in this report are those of the author and do not necessarily reflect the opinions of ACOLA.

Submission to: Professor Toby Walsh, Chair, Horizon Scanning Report on AI for Australia's Chief Scientist and Australian Commonwealth Science Council

From: Professor Seumas Miller

Date: 29th July 2018

Thank you for the opportunity to provide input to your report. Given time constraints, my input will be relatively brief and somewhat narrowly focused on machine learning. However, I note that machine learning as a so-called general-purpose technology is widely regarded as being at the forefront of projected economic, health, educational, criminal justice and other benefits arising from developments in AI – as well as causing, at least potentially, burdens of various kinds. The benefits include ones as diverse as economic efficiencies in logistics, improvements in disease diagnosis and fraud detection. The potential burdens include job losses and loss of human control.

In this submission I restrict myself to outlining a number of key moral or ethical issues (I use these terms interchangeably) rather than providing a set of concrete recommendations (albeit I do offer a few suggestions). However, in doing so I stress that recent developments in machine learning give rise to a host of complex moral problems at a number of levels; problems that obviously require for their solution a great deal of empirical and regulatory input as well as moral analysis. Indeed, the complexity is such that any submission on the ethical problems at this stage is, I suggest, more likely to consist of an introduction to the problems rather than a detailed set of recommendations for fixing them. The obvious exception to this is a recommendation to fund interdisciplinary research into these problems and do so sooner rather than later.

We might usefully categorize the ethical or moral problems in question in terms of whether:
(a) They are a feature of at least some machine learning techniques as such, e.g. even expert users do not adequately understand how the machines in these cases arrive at their results (the so-called 'black box' problem); (b) They arise from deficiencies in, or other problems pertaining to, the data bases which the machine learning techniques in question rely on, e.g. false data, data bases of personal information access to which has not been consented to by the owners; (c) They arise from the potential uses – notably, morally unacceptable uses - to which machine learning techniques might be put, e.g. Cambridge Analytica's use of machine learning techniques to intervene in electoral processes in the US and elsewhere. In light of this threefold categorization it is evident that the responses to moral problems arising from developments in machine learning would include not only the provision of technical solutions (e.g. in the case of black boxes) but also new legislation (e.g. privacy legislation such as EU's General Data Protection Regulation (GDPR)) or even institutional redesign (e.g. in relation to Facebook's business model). I note the potentially international character of the latter responses.

1. Machine Learning and Distributive Justice

As is usually the case with the advent of significant new technologies, moral problems have arisen not only with respect to the nature and quantum of the benefits and burdens thought to flow from machine learning, but also the distribution of these benefits and burdens to disadvantaged groups, in particular, under existing and emerging local, national and global institutional arrangements.

The use of machine-learning techniques based on large data bases of personal information for marketing and advertising, in particular, by a small number of powerful, profit driven, economic actors, e.g. Facebook, Amazon, interacting with a very large number of relatively weak and uninformed consumers may ultimately, it has been suggested (no doubt oversimplifying), have the following effects. Firstly, it might turbo-charge existing processes of commercialization and commodification of human activity (e.g. Facebook 'friendships' as a saleable commodity to advertisers) to the ultimate detriment of societies (understood as moral communities rather than merely as cohorts of consumers). Secondly, it might serve to widen the gap between rich and poor (e.g. in part via job losses in traditional sectors). Thirdly, it might enhance the power of multi-national market-based actors (especially given the large proportion of R&D expenditure on machine learning undertaken in the private sector and by large multinational companies) over democratically elected governments seeking to redistribute those profits to benefit the wider society.

This is, of course, not to disparage the actual and potential benefits of the use of machine-learning techniques in a wide range of sectors (as noted above); nor does it require a response directed at machine learning techniques per se (to invoke my above-mentioned threefold categorization). However, here as elsewhere, the question arises as to whether this new general purpose technology will principally be used to meet important human needs, (e.g. assist in the provision of analyzed bulk data on which to base public policies to combat poverty, the design of robots to assist the disabled,) as opposed to satisfying (and, for that matter, creating) relatively frivolous human desires, (e.g. profile based 'targeted' advertising of fashion accessories and marketing of wannabe celebrities). More generally, is the decision-making in relation to this and related questions to be left entirely to the operations of the market or is there is to be some form of government intervention to ensure morally sustainable outcomes, including a just distribution of the benefits of machine learning to the disadvantaged? I note that an acceptable answer to this question, or rather set of questions, relies heavily on empirical evidence rather than the ideological claims of vested interests (whether from the political left or the political right).

2. Machine Learning and Privacy

Machine learning techniques are currently being widely used by large companies, notably Facebook, Google, Amazon etc., in accordance with a business model based on consumers accessing a variety of communication and related services in return for providing their personal information for use in advertising and marketing. Evidently, this institutional arrangement brings with it communicative benefits (among others) but also privacy burdens (among others) – burdens in the form of privacy rights infringements which may or may not (e.g. are these rights being overstated to the economic detriment of the EU?) be appropriately addressed by recent EU legislation (GDPR) and projected legislation (e-Privacy Regulation).

Notional responses (with differential impact) in relation to the privacy issue include:
(i)Undermining (in effect) the business model of Facebook etc., e.g. by way of regulation or of introducing stringent individual privacy rights (leading to loss of customers and investment);

(ii) Establishing a publicly funded (e.g. by way of license fees) public sector competitor to Facebook, Google etc. along the lines of the Australian Broadcasting Corporation; (iii) Enhanced accountability mechanisms, e.g. in relation to ensuring compliance with privacy regulations.

3. Machine Learning, National Security and Individual Rights

Will the widespread use of machine-learning techniques in the context of bulk data collection by governments, notably in China but also, in the wake of the Snowden revelations about the activities of the National Security Agency (NSA), in the US, enable legitimate national security needs to be better met (e.g. in the context of global jihadist terrorism and inter-state cyber-conflict) or undermine individual rights, (e.g. as in the case of (to take an admittedlyextreme example) the data based 'surveillance society' in which the Uighurs in Xinjiang increasingly live)? In the case of the US and other liberal democracies there are political, legal and other deliberative processes that are well underway to try to 'balance' security requirements against individual rights (or otherwise resolve the tension). However, these issues are far from resolved when it comes to the collection of bulk data comprised of personal information and the attendant use of machine-learning techniques. Here, as elsewhere, purported solutions have a moral dimension, e.g. Do individuals have a right to control all so-called personal information including their metadata – as some have argued for in relation to forthcoming EU e-Privacy regulation? If so, will this not unduly hamper law enforcement agencies in, for instance, their counter-terrorism efforts? One answer to the latter problem might be to stop short of providing individuals with the legal right to control all their metadata but rather to restrict – except in relation to serious crimes and under judicial warrant - the aggregation and analysis of a wide range of the metadata of any given individual since it is by virtue of the aggregated meta-data of an individual that a profile of that individual can be constructed enabling the privacy infringing tracking of the individual to take place. This would be consistent with the bulk collection of metadata by security agencies in so far as the data in question was appropriately anonymized.

Further, as the recent Cambridge Analytica scandal has demonstrated, moral problems with national security implications can arise as a result of cooperation between state actors and market actors. Facebook and Cambridge Analytica are both market actors. Yet bulk data stored by Facebook and machine learning techniques deployed by Cambridge Analytica played a central role in Russia's targeting of 'vulnerable' US voters in marginal seats in particular with, for instance, so-called 'fake news' for the purpose of undermining US democratic processes. Naturally, it is agreed on all hands (corrupt and/or authoritarian leaders exempted) that such undermining of democratic processes is morally unacceptable. Potential regulatory measure to deal with this problem include ones in respect of campaign advertising, e.g. bans on micro targeted political advertising, mandatory transparency by way of public registers of the source of any political messages being disseminated and, at a more general level, deeming Facebook and other social media platforms to be publishers or to have similar responsibilities and legal liabilities to those of publishers in respect of certain types of content communicated by way of their platform.

4. Machine Learning and Legal Adjudication

Machine learning techniques can be used in order to help predict the likely legal outcomes of cases based on past outcomes in similar cases, e.g. for the purposes of settling out of court to avoid lengthy and expensive court cases or in order to facilitate plea bargaining. However, these uses of machine learning techniques typically assume, firstly, a large data set of past cases and, secondly, that new cases have similar features to past ones. Consider, for instance, the potentially highly successful area of predicting the outcomes of divorce proceedings and, thereby, saving disputants much money and drastically reducing court case loads. Determinations of the likelihood of success in divorce proceedings are based on outcomes of past cases and weighting of criteria used in these past cases. However, past cases involve judicial errors, e.g. on the part of solicitors, barristers and magistrates. Accordingly, these errors, especially if frequently made, can now enter into the predictive process. If so, predictions of the likely outcomes of current cases in which the adjudications do not repeat, or would not have repeated (if they had taken place), past errors might turn out, or might have turned out, to be false predictions. Therefore, those who have acted upon these predictions will have been misled.

Moreover, complex, contested criminal cases are much less amenable to machine learning techniques than simple, high volume, legal adjudications, given the inherent particularity of many of these cases. Appropriate legal adjudications in such cases may have an inherent particularity that renders them immune to prediction on the basis of machine learning techniques. If so, there are limitations to the utilization of machine learning techniques in legal adjudication and attempts to exceed these limitations may well lead, not simply to error, but to injustice, e.g. punishing the innocent. Similar points can be made in relation to adjudications with respect to sentencing and the granting of parole, depending on the kinds of adjudications in question, what is morally at stake, the quality and size of the available data bases, the degree of reliance on the output generated by the machine learning techniques used etc.

5. Autonomous Weapons

So-called autonomous robots are now a fact of life. Moreover, they exist in part by virtue of recent developments in machine learning techniques. Autonomous robots are able to perform many tasks far more efficiently than humans, e.g. tasks performed in factory assembly lines, auto-pilots, driverless cars; moreover, they can perform tasks dangerous for humans to perform, e.g. defuse bombs. However, autonomous robots can also be weaponized.

New and emerging (so-called) autonomous robotic weapons can replace some military roles performed by humans and enhance others. Consider, for example, the Samsung stationary robot which functions as a sentry in the demilitarized zone between North and South Korea. Once programmed and activated, it has the capability to track, identify and fire its machine guns at human targets without the further intervention of a human operator. Predator drones are used in Afghanistan and the tribal areas of Pakistan to kill suspected terrorists. While the ones currently in use are not autonomous weapons they could be given this capability in which case, once programmed and activated, they could track, identify and destroy human and other targets without the further intervention of a human operator.

Autonomous weapons are weapons system which, once programed and activated by a human operator, can – and, if used, do in fact – identify, track and deliver lethal force without further intervention by a human operator. By 'programmed' I mean, at least, that the individual target or type of target has been selected and programmed into the weapons system. By 'activated' I mean, at least, that the process culminating in the already programmed weapon delivering lethal force has been initiated. This weaponry includes weapons used in non-targeted killing, such as autonomous anti-aircraft weapons systems used against multiple attacking aircraft or, more futuristically, against swarm technology (for example multiple lethal miniature attack drones operating as a swarm so as to inhibit effective defensive measures); and ones used or, at least, capable of being used in targeted killing (for example a predator drone with face-recognition technology and no human operator to confirm a match).

We need to distinguish between so-called 'human in-the-loop', 'human on-the-loop' and 'human out-of-the-loop' weaponry. It is only human out-of-the-loop weapons that are autonomous in the required sense. In the case of human-in-the-loop weapons the final delivery of lethal force (for example by a predator drone), cannot be done without the decision to do so by the human operator. In the case of human on-the-loop weapons, the final delivery of lethal force can be done without the decision to do so by the human operator; however, the human operator can override the weapon system's triggering mechanism. In the case of human out-of-the-loop weapons, the human operator cannot override the weapon system's triggering mechanism; so, once the weapon system is programmed and activated there is, and cannot be, any further human intervention.

The lethal use of a human-in-the-loop weapon is a standard case of killing by a human combatant and as such is, at least in principle, morally permissible. Moreover, other things being equal, the combatant is morally responsible for the killing. The lethal use of a human-on-the-loop weapon is also in principle morally permissible. What of human-out-of-the-loop weapons? Human out-of-the-loop weapons – so-called 'killer-robots' – are not morally responsible for any killings they cause. Consider the case of a human in-the-loop or human-on-the-loop weapon. Assume that the programmer/activator of the weapon and the operator of the weapon at the point of delivery are two different human agents. If so, then other things being equal they are jointly morally responsible for the killing done by the weapon (whether it be of a uniquely identified individual or an individual qua member of a class). No-one thinks the weapon is morally or other than causally responsible for the killing. Now assume this weapon is converted to a human out-of-the-loop weapon by the human programmer-activator. Surely this human programmer-activator now has full individual moral responsibility for the killing. To be sure there is no human intervention in the causal process after programming-activation. But the weapon has not been magically transformed from an entity only with causal responsibility to one which now has moral or other than causal responsibility for the killing.

Human-*out*-of-the-loop weapons can be designed to have an override function and/or an on/off switch controlled by a human operator. Moreover, in the light of our above example and like cases, in general autonomous weapons ought to have an override function and/or on/off switch. Indeed, to fail to do so would be tantamount to an abnegation of moral responsibility.

6. Machine Learning, Moral Principles and Human Autonomy

Evidently, the introduction of autonomous cars on the streets of Australia and elsewhere is imminent. Autonomous cars need to be able to comply with the road rules, e.g. stop at red lights and zebra crossings, but they also, indeed simultaneously in the case of many road rules, need to be able to comply with moral principles enshrined in laws, e.g. avoid running over pedestrians (as happened recently in the well-publicised case of a woman killed by a self-driving car in Arizona). This raises the issue of the possibility of machines complying with moral principles and, in particular, legally enshrined moral principles.

Perhaps the most dramatic example of this is research being undertaken with a view to building machines that could fight wars in accordance with the legally enshrined moral rules of war and, in particular, the moral principles enshrined in international law namely the principles of (1) military necessity, (2) proportionality and (3) discrimination. Can moral principles, such as military necessity, proportionality and discrimination be programmed in to computers? The problem here is that while such robots are sensitive to physical features of the environment and can pursue a variety of goals they are not sensitive to moral properties. Computers do not care about anyone or anything (including themselves), and cannot recognise moral properties, such as courage, moral innocence, moral responsibility, sympathy or justice as such; nor do they recognise moral ends qua moral ends. Therefore, computers cannot act for the sake of moral ends or principles qua moral ends or principles. A robot can refrain from killing something because it is programmed not to kill things of that kind in the circumstances in question but not because it recognises what is morally right from what is morally wrong and acts from the motive of doing what is morally right. Given the non-reducibility of moral concepts and properties to physical ones, at best computers can be programmed to comply with some non-moral physical proxy for moral requirements. The proxy for ‘Do not intentionally kill morally innocent human beings’ might be ‘Do not fire at bipeds if they are not carrying a weapon or they are not wearing a uniform of the following description’. Given the non-reducibility of the moral to the physical or, at least, the lack of reliable, precise, detailed correlations, this is extremely doubtful in other than highly circumscribed contexts.

The highly circumscribed contexts in question are, firstly, ones in which the set of possible actions is bounded (e.g. there are no entirely novel, game-changing, scientific breakthroughs, such as AI, to radically change the set of possible actions) and the outcomes of these actions, if performed, are predictable. Secondly, they are contexts in which the moral principles to be applied and the ends to be pursued are clear-cut. Thirdly, there is agreement on a pre-determined moral decision-making procedure to resolve conflicts in the application of these principles and the moral weight to be accorded to these ends. However, war is not highly circumscribed in these respects. For one thing, the principles of discrimination, military necessity and proportionality (as it ought to be applied in wars) and, for that matter, the moral weight to be attached to the ends for which war are fought, are not clear cut. For another, there is no moral agreement on a pre-determined moral decision-making procedure to resolve conflicts in the application of these principles and the moral weight to be accorded to these ends. Indeed, it is a matter of controversy whether the notion of a pre-determined moral decision-making procedure is ultimately coherent other than in the form of a set of heuristic devices tailored to particular, narrow, decision-making contexts. For instance, there could be such a heuristic pre-determined decision-making procedure for safe driving on highways (see discussion below on autonomous cars).

What of autonomous cars? Whereas cars cannot be programmed to recognise moral properties as such they can certainly be programmed to avoid crashing into one another or into other physical objects, including human beings. Hence the use of autonomous trucks without human drivers to transport iron ore in remote mining sites in Australia. An important question here concerns the use of autonomous cars in much less predictable driving circumstances, e.g. near schools. Consider the moral question: Should a car-driver break hard to avoid hitting a child but not to avoid hitting a dog in circumstances in which breaking hard will likely cause a rear end collision with consequent injury both to himself and the driver of the car which crashes into him? Answer: Yes (depending on the likelihood of serious driver injury and…). If so, can this degree of moral complexity be programmed into a computer and, in all relevant driving conditions, and can children be reliably distinguished from dogs, wallabies, robots etc.? Perhaps so; or perhaps autonomous cars can be appropriately designed/programmed to do better than humans in respect of compliance with pre-determined moral decision procedures and in distinguishing children from dogs under all or most driving conditions etc. However, given what is at stake (i.e. serious injury, life itself), the unpredictability of events, computer failure and, more generally, the potential inadequacy of pre-determined moral decision procedures (see above) should a human 'driver' of a car, at least in many driving circumstances, be on-the-loop – as opposed to in-the-loop? That is, should the human driver be able to override the 'decisions' of the autonomous car? If there is no human in or on-the-loop, who should be held morally responsible and/or legally liable when death/injuries are caused and, if so, on what grounds, e.g. the human driver for choosing not to be on-the-loop (if this choice is indeed available), the manufacturers because (in part) they have deep pockets, the designer/programmer for selecting/inputting the particular moral decision-making procedure expressed in the algorithms (according to which, for instance, (controversially) killing a dog is less morally preferable than allowing a human to suffer serious injury), no-one? On the other hand, if there is a human on-the-loop, does this not significantly reduce the advantages of autonomous cars? In short, it is not clear that all the moral issues in respect of autonomous cars have been resolved.