Machine Learning, Ethics and Law

Miller, Seumas

DOI
10.3127/ajis.v23i0.1893

Publication date
2019

Document Version
Final published version

Published in
Australasian Journal of Information Systems

Citation (APA)
Miller, S. (2019). Machine Learning, Ethics and Law. *Australasian Journal of Information Systems*, *23*, 1-13.
https://doi.org/10.3127/ajis.v23i0.1893

Machine Learning, Ethics and Law

Seumas Miller
Charles Sturt University
TU Delft
University of Oxford
semiller@csu.edu.au

Abstract

Recent revelations concerning data firm Cambridge Analytica's illegitimate use of the data of millions of Facebook users highlights the ethical and, relatedly, legal issues arising from the use of machine learning techniques. Cambridge Analytica is, or was – the revelations brought about its demise - a firm that used machine learning processes to try to influence elections in the US and elsewhere by, for instance, targeting 'vulnerable' voters in marginal seats with political advertising. Of course, there is nothing new about political candidates and parties employing firms to engage in political advertising on their behalf, but if a data firm has access to the personal information of millions of voters, and is skilled in the use of machine learning techniques, then it can develop detailed, fine-grained voter profiles that enable political actors to reach a whole new level of manipulative influence over voters.

My focus in this paper is not with the highly publicised ethical and legal issues arising from Cambridge Analytic's activities but rather with some important ethical issues arising from the use of machine learning techniques that have not received the attention and analysis that they deserve. I focus on three areas in which machine learning techniques are used or, it is claimed, should be used, and which give rise to problems at the interface of law and ethics (or law and morality, I use the terms "ethics" and "morality" interchangeably). The three areas are profiling and predictive policing (Saunders et al. 2016), legal adjudication (Zeleznikow, 2017), and machines' compliance with legally enshrined moral principles (Arkin 2010). I note that here, as elsewhere, new and emerging technologies are developing rapidly making it difficult to predict what might or might not be able to be achieved in the future. For this reason, I have adopted the conservative stance of restricting my ethical analysis to existing machine learning techniques and applications rather than those that are the object of speculation or even informed extrapolation (Mittelstadt et al. 2015). This has the consequence that what I might regard as a limitation of machine learning techniques, e.g. in respect of predicting novel outcomes or of accommodating moral principles, might be thought by others to be merely a limitation of currently available techniques. After all, has not the history of AI recently shown the naysayers to have been proved wrong? Certainly, AI has seen some impressive results, including the construction of computers that can defeat human experts in complex games, such as chess and Go (Silver et al. 2017), and others that can do a better job than human medical experts at identifying the malignancy of moles and the like (Esteva et al. 2017). However, since by definition future machine learning techniques and applications are not yet with us the general claim that current limitations will be overcome cannot at this time be confirmed or disconfirmed on the basis of empirical evidence.

Keywords: applied ethics; machine-learning; law

1 Profiling and Predictive Policing

Profiling of potential suspects and geographical locations for crimes has been around for some time (Schauer 2003). For instance, law enforcement resources have not only been directed at persons 'known to the police', such as persons with past convictions, but at persons who have features that correlate with the perpetration of crimes of a certain type, e.g. a gambling habit, bad debts, a recent divorce etc. might correlate with certain forms of white-collar crime. Again, police resources can be directed to crime hotspots identified not simply on the basis of past crimes committed at that location, but on the basis of possession of features statistically correlated with crimes of the relevant type and, therefore, believed to make a location crime-prone, e.g. the location has a number of tourist attractions and is close to a main road and to train and bus stations. However, the advent of big data and machine-learning has given significant additional impetus to the use of profiling in law enforcement and, thereby, to so-called predictive policing (Saunders et al 2016).

Offender profiling can involve making inferences from characteristics of offences to characteristics of offenders, and thereby identify the perpetrator of a particular crime or set of crimes. Of course, in a general sense, police investigators have always been engaged in this kind of inference making. However, contemporary policing is able to use a range of theories, research methods and statistical tools to enhance traditional police practice, including machine learning techniques.

Many police services have access to crime analysts and specialist police staff who analyse crime according to specific criteria, for example, modus operandi, behaviour exhibited by offender, and links to other crimes and offenders. Moreover, there are academics researching serious crime who make their findings available to police investigators, e.g. psychological research on different categories of offender.

Historically, the crimes that have involved the use of offender profiling have mainly been rape and motiveless, or sexually motivated, murder cases. However, profiling is not restricted to these crimes. Indeed, high volume crime, such as burglaries and car theft, are in many respects more amenable to profiling techniques and, in particular, to profiling techniques using big data and machine learning, than murder or rape. For one thing, the data bases of murder and rape cases are comparatively small. The comparatively small number of terrorist cases in data bases is also an impediment to the use of profiling techniques dependent on big data and machine learning in this area of law enforcement.

The profiling process is greatly facilitated by databases of offenders (rapists, paedophiles, etc.) in which the data held is detailed and specific, e.g. it describes their modus operandi and provides psychological/behavioural profiles. Here, as elsewhere, the integration or linking of databases in different jurisdictions is also extremely helpful especially, as in the US, where there are a large number of jurisdictions many of which are very small.

Profiling in law enforcement has two main types. The first type involves developing a profile or set of characteristics of a category of persons, for instance, a typical police officer at risk of corruption, based on generalizations from the past behaviour of persons belonging to this category. So a set of characteristics might be developed, including such things as numbers of complaints, associating with known criminals, gambling habits, substance abuse, financial problems, and operating in a high-risk area such as drug law enforcement, and consolidated so as to constitute the profile of a typical police officer engaged in corruption. Once the profile is constructed, the internal affairs investigators can, at least in theory, monitor individuals with that profile in their preventative anti-corruption strategies.

However, big data and machine learning can take this first kind of profiling much further. Obviously, as already mentioned, the size of the data base is of great importance in the development of profiles; so the establishment of detailed electronic databases and the integration of these data bases across multiple jurisdictions greatly facilitates profiling. Crucially, however, machine learning techniques enable profiles to be significantly refined. For machine learning techniques can detect additional correlations and, thereby, generate new features of the pre-existing profile. Importantly, these new resulting 'features' are not known or intuited by analysists in advance – unlike the pre-programmed features mentioned above in traditional profiling – and, indeed, even once known they might not be intuitively obvious. Rather it is a matter of there being an algorithmically-based correlation with all those persons who fit the original profile.

The second type of profiling involves developing a profile or set of characteristics of the person who has committed the crime being investigated but whose identity is unknown; for instance, the serial rapist – whoever that person is - who raped Mary, Betty, and Jane. The profile is based on generalizations from the modus operandi, past behaviour, and so on of this offender and of like offenders. As with the first type of profiling, machine learning techniques can be utilized to refine the profile. Once the profile is constructed the investigators can access their databases, or otherwise look for the particular person that fits the profile, and do so for the purpose of solving the crime in question.

Recent developments in communications and information technology, including the creation and integration of large databases, high-speed, long-distance accessing and communication of content, for instance, via the Internet, and, more recently, the use of machine learning techniques, have enormously facilitated profiling in both of these two forms. However, the existence and possibility of widespread profiling in law enforcement has raised a range of ethical problems. One issue or set of issues concerns privacy, as we saw above in relation to Cambridge Analytica. Do the databases upon which machine learning techniques are applied consist of personal information or confidential information to which those using these techniques do not have a right or are otherwise not morally or legally entitled to access? There are, of course, complications here in relation to what counts as personal information. The content of telephone calls, email etc. is typically regarded as personal or confidential information, but what of metadata; data concerning the caller/called, duration of call etc.? Some have argued that this is not personal or confidential any more than the sender and receiver's name on a parcel sent through the postal service are items of personal data. However, a large bank of metadata extracted from, for example, from a person's phone calls, emails etc., can enable a detailed picture of that person's associates, movements and so on; a picture sufficiently detailed to count as an infringement of their privacy. Again, is the purpose to which the machine learning applications are a means a legitimate purpose? Presumably, the purposes involved in predictive policing are for the most part legitimate law enforcement purposes, such as reducing crime, whereas Cambridge Analytica's purposes – and certainly the purpose of a foreign power to influence an election outcome – are not.

Let us, then, assume that profiling undertaken by law enforcement is in the service of legitimate purposes and that the data upon which the profiling depends is not in any obvious way personal or confidential in character, e.g. the data is either publicly available or is appropriately anonymized personal data.[1] Are there other ethical concerns? Arguably, there is a further ethical concern or, at least, need for moral justification, in respect of profiling in our first sense, that is, where there is no specific (actual or reasonably suspected) past, imminent or planned crime under investigation. For example, in the case of police officers who are not known to have committed any crime, what is the justification for monitoring their behavior – other than for ordinary work performance purposes? Are police any different from ordinary citizens in this regard? It might be argued that, given their position of trust and the fact that they have extensive powers of arrest and use of lethal force not possessed by ordinary citizens, that such monitoring is morally justified (Kleinig 1996; Miller and Gordon 2014: 201-223). Moreover, their occupational role as a police officer is one freely chosen. However, these arguments justifying the profiling of police officers are not available in the case of ordinary citizens, even if profiling of citizens might serve the ultimate purpose of reducing crime. Speaking generally, there ought to be knowledge that a crime has been, or is about to be, committed or is being planned, and reasonable suspicion that citizen Smith has or will commit the crime in question (absent police intervention), prior to any process of monitoring. The fact that Smith has committed such crimes in the past might in some cases constitute reasonable suspicion, as might the fact that Smith had the ability and opportunity to commit an already committed, or about to be committed, crime. However, in the absence of police focus on any specific (actual or reasonably) past, imminent or planned crime, (e.g. the murder of Jones), or ongoing spate of crimes, (e.g. recent burglaries in a specific location), surely the mere fact that some citizens might have the profile of individuals who tend to commit a certain crime type is not a sufficient justification for generating such profiles, determining who among the citizenry has these profiles and monitoring these citizens. Such invasive law enforcement practices are evidently inconsistent with fundamental principles underpinning liberal democracy and, in particular, the individual's right to freedom from state interference absent prior evidence of violation of its laws. In a liberal democratic state, it is generally accepted, the state has no right to seek evidence of wrongdoing on the part of a citizen whose actions have not otherwise reasonably raised suspicion of unlawful behavior. At the very least in a democracy these law enforcement practices would need to be consented to by the citizenry via the democratic process.

These ethical problems with profiling are compounded in the case of the use of machine learning techniques by the so-called 'black box' issue. In the case of the profiles generated by machine learning the algorithmic based correlations may not be known or understood by either the citizens or law enforcement. So citizen Smith might be being monitored without either Smith or the police monitoring him knowing at any point what a key part of the justification for this monitoring is, i.e. what features he possesses that make him, in effect, an object of suspicion.

Frederick Schauer's work can be viewed, at least in part, as an attempt to provide an antidote to many of the fears that have arisen in relation to the practice of profiling. By Schauer's lights as an advocate of the morality of decisions based on generalizations, profiling is already an acceptable and ubiquitous practice. For example, he discusses the case of Sokolow who was searched by customs officials at an airport and found to possess drugs (Schauer 2003: 172). He was searched because he fitted the profile of a drug courier and this was taken to constitute reasonable grounds for suspicion. Sokolow argued in court that fitting a profile did not

[1] There are potential ethical problems with the latter which would take me too far afield to go into here.

constitute reasonable grounds for suspicion, but he lost the case. Schauer claims that, "the issue is not about profiling at all, for profiling is inevitable" and "profiling is largely unobjectionable" (Schauer 2003: 174).

Whatever the value of Schauer's discussion of profiling understood in a general sense, it is unconvincing in relation to profiling that utilizes machine-learning techniques. For one thing, as just mentioned, the justification for the elements of the profile might not be known to law enforcement. For another, on many influential accounts of judgment Schauer distorts the nature and function of *discretionary* judgment; discretionary judgment is not, many theorists would hold, simply personal, unscientific profiling. Rather, discretionary judgment is called for in relation to matters that have an inherent particularity (see next section for instances of this). By the lights of some theorists (Harre and Madden 1975), such judgements include ones involved in establishing causation in the law. On this kind of view, causation is sharply distinguished from correlation by virtue of the inherent particularity of the former; causal powers are powers possessed by powerful particulars rather than mere regularities in behaviour. Accordingly, the role of discretionary judgment may well be to supplement profiling – including profiling utilizing machine learning techniques - in order to rule out or rule in instances where the profiling generates manifestly absurd outcomes, e.g. citizen Jones fits the profile but has a rock-solid alibi. Finally, it could be argued that Schauer's conception does not give due weight to the dangers of profiling, including that involving machine learning techniques, e.g. the danger of discriminatory algorithms. For instance, racial profiling may entrench existing racist attitudes and may generate over-policing leading to police-community tension which in turns obstructs law enforcement.

2 Machine Learning and Legal Adjudication

Another area in which machine learning techniques are being used is in the legal quagmire of divorce proceedings (Zeleznikow 2017). Although separation and divorce can be amicable, they can also be monumentally expensive as a result of legal fees. New machine learning software now attempts to predict the settlement outcome, based on a huge case history and relatively clear-cut legal criteria for settlement. If the protagonists accept the prediction, a low-cost agreement can be achieved.

Accordingly, the utilization of machine learning techniques in areas of the law involving a high volume of similar types of case and relatively clear-cut legal rules, such as divorce proceedings, may well be hugely beneficial. However, here as elsewhere, there is a need for caution. Consider the following somewhat non-standard case.[2] Rachel had risen to vice-president of a profitable company, leaving not enough time for looking after her 6 year old son. Now she was hit a double whammy. Her partner had found a new lover and wanted a split, while her company had merged with a larger company and she was out of a job. She would get back again, but it would take a while to find something of comparable level. But she was happy to have more time for her son, and, finances would not be too bad for a few years, given a reasonable settlement with custody. Unfortunately, the software predicted she would not get custody!

As noted above, predicting future legal outcomes of cases based on past outcomes assumes, firstly, a large data set of past cases and, secondly, that new cases have similar features to past

[2] I owe this example to Terry Bossomaier.

ones. Determinations of 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. However, in doing so predictions in current cases in which adjudications do not repeat past errors might have turned out to be false predictions and, thereby, mislead those who have acted upon these predictions. Moreover, the possibility of correcting these errors might well be lost if, for example, the prediction is simply accepted at face value and acted on and, in particular, acted on without going through a thorough process in which all the various arguments, evidence and so on are aired prior to a considered adjudication.

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. Consider the legal adjudication in the case of the serial murderer and rapist, Robert Black. The case warrants detailed description since the point of using it as an example is to demonstrate its inherent particularity. [3]

In July 1990 a six-year-old girl, Mandy Wilson was abducted as she walked in her street in Stow (a town in Scotland close to the English border). By a stroke of immense good fortune, a neighbour saw her walk towards a parked white van and he could see her feet under the open passenger door beside those of a man. The girl's feet vanished, the van drove off; the witness took the registration number and immediately called the police. The witness was describing the event to the girl's father (a police officer) when the van reappeared and was immediately stopped by the police. The father found his daughter bound, gagged and stuffed in a sleeping bag behind the driver's seat. She was terrified and had already been sexually assaulted.

The driver of the van was Robert Black, a delivery driver who travelled throughout the UK. Black was arrested, charged, pleaded guilty and was sentenced to life imprisonment.

Black now became the main suspect in the Maxwell, Hogg and Harper murders, there being evident similarity between the cases in terms of MO (modus operandi) and other factors; Black however declined to speak about the abductions and murders of the three girls. The investigators began a thorough and scrupulous examination of Black's movements and lifestyle between 1982 and 1990, where a key focus was on his job as a delivery driver. Using work records, including wage records and fuel receipts, they built up a picture of his movements and were able to place him in the vicinity of each abduction at the appropriate time and also where the bodies had been found. The investigators also discovered an attempted abduction of a 15-year-old girl, which had failed because she had fought back and a friend had gone to her assistance. The witnesses' descriptions of the assailant were an exact match to Black.

There was no forensic evidence and no admission by Black – the case was built on the above-described circumstantial evidence linking the various murders to one another and to Black - but in April 1992 the Crown Prosecution Service elected to prosecute.

Importantly, for our purposes here, this was a unique case in UK criminal law. The defence argued there was no direct evidence to establish that Black had committed the offences and argued that each murder should be treated separately. But the court allowed the murders to

[3] This description of the case is taken from Miller and Gordon (2014, 127-132).

be presented as a series and allowed evidence from the earlier case relating to the abduction of Mandy Wilson and the attempted abduction of the 15 year old. Black appeared at Newcastle upon Tyne Crown Court in April 1994, when the prosecution detailed the striking similarities between the murder cases and also the attempted abduction and the actual abduction of Mandy Wilson. The series of murders exhibited a common modus operandi (MO) followed by Black and Black was linked to each of them in terms of his movements

Black was convicted of all the charges before the Court and sentenced to life imprisonment on each one; there was a minimum term of 35 years recommended for each of the murders.

In October 2011 Black was further convicted of the abduction and murder in 1981, in County Down Northern Ireland, of Jennifer Cardy, 9 years. This offence also relied on painstaking investigation of Black's work records and included testimony from Detective Chief Superintendent (retired) Roger Orr, SIO in the Mandy Wilson abduction, where he described Black's actions in that offence and in the Maxwell/Hogg/Harper murders. The prosecution alleged the evidence given by Mr Orr amounted to "a signature for Robert Black and that the case of Jennifer also bears that signature". It is strongly believed that Black may be responsible for a further twelve abduction and murders of young girls, in the UK, France and Holland. To date he has declined to assist any police enquiry.

To summarise: Robert Black was a serial rapist and murderer but there was only circumstantial evidence in each of the murder cases. Accordingly, the prediction in each case considered on its own– including predictions using machine learning techniques - would have been 'not guilty'. However, there was an evidential link between each of these cases: the modus operandi of Black. The prosecution argued that there was a distinctive 'signature' MO in each case and that this MO was used in the abduction case for which he was convicted as well as the murder cases for which there was insufficient evidence absent recourse to the signature MO. The point to be stressed here is not that there was a pattern i.e. the MO, although the existence of a pattern was a necessary condition for the legal outcome. Nor is it that machine learning was not necessary to establish this pattern, although obviously machine learning was not required to discover a pattern in a handful of murder-rape cases. Rather the point to be stressed is that a discretionary, and inherently particular, legal decision was made; a decision that allowed for the first time an evidential relationship between different cases to be useable in a single discrete case. In short, the Robert Black case was at the time unique and unique in a manner that made it unable to be predicted on the basis of adjudications in past cases of serial murder and rape. Accordingly, the legal adjudication in the Black case would have been immune to prediction on the basis of machine learning techniques.

The more general point is that appropriate legal adjudications in complex cases may have an inherent particularity that renders them immune to prediction on the basis machine learning techniques. Therefore, there are evidently 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, whether in the form of punishing the innocent or failing to punish the guilty.

3 Machines and Compliance with Legally Enshrined Moral Principles

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.[4]

Ron Arkin has been undertaking research with a view to building machines that could fight wars in accordance with the rules of war (Arkin 2010). The actual conduct of war is governed by moral principles enshrined in international law (the so-called jus in bello of just war theory), notably the principles of (1) military necessity, (2) proportionality and (3) discrimination. Accordingly, Arkin has sought to demonstrate how weaponized autonomous robots could not only fight wars, but in doing so comply with the principles of military necessity, proportionality and discrimination. As will become evident, these principles are quite unlike the precisely defined rules - and precisely defined and limited contexts of application of rules - in games, such as chess and Go, in which, as mentioned above, machines have had notable success in recent times. Thus the rules of chess precisely define what moves the chess pieces, (e.g. pawns, bishops) can and cannot make, and precisely define also the context of application, i.e. the chessboard and the configured chess pieces (e.g. pawns, bishops etc.). They also precisely define what counts as winning, e.g. check-mating the king. For the moment I note that by contrast with the rule of chess, the *ius in bello* moral principles of war are not well-defined, (e.g. what counts as a disproportionate use of force?) and have to be applied in multiple, diverse and shifting military contexts, (e.g. conventional theatres of war and counter-terrorism operations, war at sea and war in the jungle) that are typically not precisely defined or delimited, (e.g. terrorism in civilian areas of failed states). Moreover, what counts as winning a war – and, therefore, what the definition of military necessity is (see below) - is not precisely defined either, e.g. withdrawal of enemy forces from the defending nation’s territory, devastating the enemy’s cities with atomic bombs etc. Importantly, unlike in the typical case of the application of the law by police officers, these principles apply at the collective level, as opposed to merely at the individual level. So the context of any or, at least, most applications of these principles are *multi-levelled* and applies to organizations and not merely to individuals. What do I mean by the collective level(s)?

The principle of military necessity, in particular, but also the principles of proportionality and discrimination, apply at various collective levels, such as at the level of a battle or at the level of an ongoing war fought by a military organization. Accordingly, it might be militarily necessary to bomb a munitions dump in order to win a battle and it might be, in turn, militarily necessary to win the battle in order to win the war. The contrast here is with the individual level, i.e. the level of an individual combatant’s lethal action considered as one-off self-contained action. At the individual level, it might be necessary for Private Smith to shoot dead an approaching enemy combatant who will otherwise kill Jones. Accordingly, the context for the application of these moral principles is a multi-level (individual and collective) context. Let me explain further.

In essence, the principle of military necessity ultimately pertains to the long term, necessarily underspecified collective end of winning the war which generates in turn a nested, dynamic, series of short and medium term collective ends, such as winning particular battles or

[4] An earlier version of a number of the arguments in this section appeared in Miller (2015).

firefights. These short and medium term collective ends are means to the long term collective end of winning the war, albeit means in need of further specification, adjustment or even abandonment in the light of the responses to them of the enemy armed forces. Accordingly, the principle of military necessity is to be understood, firstly, in short/medium/long term means/end, i.e. diachronic, terms. Something is necessary in this sense if, comparatively speaking, it is both an efficient and effective means to an end and there is no obviously superior means available. If it is the *only* means, then it is *strictly* necessary. However, this is frequently not the case and so to this extent 'necessity' is correspondingly less strict. Secondly, the strength of the necessity to deploy a given quantum of lethal military force in (say) the context of a battle turns in large part on the moral weight to be accorded to the winning of that battle in light of its likely contribution to the ultimate (necessarily underspecified) collective end of winning the war (and, of course, the (somewhat indeterminate) moral weight to be attached to the latter). In the case of a crucial battle in the context of a war of collective self-defence, the military necessity to deploy a large quantum of lethal military force might be both strong (there is much at stake) and strict (it is the only available means). Thirdly, and notwithstanding the cooperative interdependence of their military actions, within a given armed force the principle of necessity is applied by multiple different military commanders within a given armed force, and each applies (or fails to apply) the principle on a given occasion on the basis of his own discretionary judgment, e.g. a bomber commander deciding whether to bomb a munitions factory in the vicinity of civilians, the US President deciding to drop atomic bombs on Hiroshima and Nagasaki. What of the principles of proportionality and discrimination?

Roughly speaking, the principle of discrimination forbids deliberately targeting innocent civilians and, also, foreseeably and avoidably putting their lives at unnecessary risk. The latter clause brings the principle of military necessity into play; a risk to civilians is unnecessary if the use of lethal military force which constitutes this risk is not militarily necessary. So the principles of military necessity and discrimination are conceptually (albeit, evidently non-algorithmically) interdependent. An important consequence of this is that the context for the application of the principle of discrimination is also multi-level; it applies at both the individual and collective levels. Moreover, since these levels are interconnected by virtue of nested collective ends, the application of the principle of discrimination necessarily involves taking into account the risks to civilians at these various levels and (possibly) adjudicating between them. For example, pursuing tactic A (aerial bombing) to realize the collective end of winning a battle might lead to many more civilian casualties in this present battle than pursuing tactic B (taking and holding ground without aerial bombing). However, pursuing A might be a more efficient and effective means of decisively winning the battle (because, say, of the much heavier enemy casualties inflicted prior to their retreat) and might, therefore, reduce the number of future civilian casualties in future battles joined in further pursuit of the collective end of winning war. What of the principle of proportionality?

The principle of proportionately arises in contexts in which both the principle of military necessity and the principle of discrimination are applicable. Roughly speaking, it requires that that the quantum of (unintended) civilian deaths resulting from the deployment of lethal military force should not be disproportionate to the strategic (and derivable moral) weight of collective military ends to be realized by that deployment. As such, the principle of proportionality is conceptually (but again, evidently non-algorithmically) interdependent with both the principle of military necessity and the principle of discrimination. Accordingly, the context for the application of the principle of proportionality is also multi-level; it applies at both the individual and collective levels.

What of machine compliance with the legally enshrined moral principle of military necessity, discrimination and proportionality? The weaponized autonomous robots in question can detect and respond to features of their environment and in many cases they have impressive storage/retrieval, calculative capacities and, importantly for our purposes here, the ability to learn new rules and finesse prior rules based on past correlations.

However, the argument at this point is based on the assumptions, firstly, that moral principles, such as military necessity, proportionality and discrimination, can be reduced to sharply defined rules, contexts of application and outcomes to be aimed at and, secondly, that these rules, contexts of application and outcomes to be aimed at are such that they can be programmed in to computers.

An initial problem is that while such robots are sensitive to physical features of the environment they are not sensitive to moral properties. After all, 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. Therefore, computers cannot act for the sake of moral ends or principles understood as moral in character, such as the principle of discrimination. 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'. Each moral principle needs to be expressible in a sharply defined rule couched in purely physical descriptive terms. Given the non-reducibility of the moral to the physical or, at least, the lack of reliable, precise, detailed correlations (e.g. innocent civilian is a vague notion and terrorists consistently seek to thwart attempts to identify them as combatants), this is extremely doubtful especially in respect of relatively vague and quite general principles, such as the principle of discrimination.

A second objection pertains to the ends in play and, in particular, the ends implicit in the principle of military necessity. Putting civilian lives at risk may be justified if it is militarily necessary i.e. in the service of the end of winning the war. But, as already mentioned, the content of the notion of winning the war is underspecified. Winning the war is not well-defined in the manner of winning the chess game. Rather, as is the case with most large-scale human enterprises taking place over extended periods the ultimate and proximate collective ends are underspecified and only become fully specified in the course of undertaking the enterprise; moreover, being collective ends, the cooperating individual who pursue them do so under somewhat different descriptions and the process of further specification is itself collaborative. Accordingly, the collective end of a war or a battle cannot readily be programmed into a computer in the manner in which a geographical located destination can be programmed into a self-driving car.

A more holistic objection to Arkin's project arises from combining a number of the above-identified difficulties attaching to compliance, including compliance assisted by machine learning techniques, with all three of these conceptually connected *ius in bello* principles. Let us remind ourselves what these difficulties are: (1) the *necessarily underspecified* nature of the collective ends constitutive of waging war prior to their realization; (2) the moral importance of realizing these collective ends, individually and in aggregate – an importance derivable ultimately and in large part from the moral benefits of successful collective self-defence and the moral costs of failure; (3) the *conceptual (but evidently non-algorithmic) interdependence* of the principles of military necessity, discrimination and proportionality; (4) the *interplay* between the applications of these principles at the various different collective levels and at the individual level, and; (5) the application of these principles in the *context of an evolving, dynamic and diachronic process of responses to, and counter-measures against, the enemy*. In short, it is not simply a matter of programming in each precisely defined rule, precisely defined outcome to be aimed at and a precisely defined set of contexts of application, and then 'pushing the start button'. Nor it is even a matter of specifying a set of independent rules (since the principles in question are interdependent), together with a meta-rule specifying which rule is to take precedence over which when there is a conflict, and mechanically applying the resultant set of rules in the relevant contexts.

Moreover, while machine learning techniques may well enable a computerized robot to learn precisely defined rules, such as chess rules, on the basis of a data set comprised of instances of past compliance with those rules, and learn and further refine successful tactics and strategies on the basis of relevant chess data sets, matters are somewhat different when it comes to the application of the *ius in bello* principles in war. As has already been stressed, these moral principles, outcomes to be aimed at, and contexts are far from being precisely defined. However, there are further problems. Where are the large data sets comprised of instances of compliance and non-compliance to these principles to be found? Perhaps from past wars? Unfortunately, past wars are extraordinarily diverse in respect of the weapons and tactics used, and the outcomes aimed at; moreover, the actual outcomes of the use of these weapons and tactics in the service of these aims varied enormously depending on the diverse physical, psychological and social contexts in which these conflicts took place. Further, and in part because of these differences in weaponry, tactics, aims and contexts, every war is significantly different from every other war with respect to the application by combatants and their leaders of moral principles. And, course, human beings interpret moral principles in different ways, can choose to apply moral principles to different degrees or, indeed, to ignore them. For instance, the principle of discrimination is applied with much greater diligence by some sides in some wars than by other sides or in other wars. The upshot of this is that there will need to be a heavy reliance on the *human beings in charge of this R&D project* in morally and legally compliant weaponized autonomous robots to interpret the *ius in bello* moral principles, to devise proxy physical rules, to identify or design data sets of incidents involving compliance and non-compliance, and so on. In short, what looked initially to be a project in the natural sciences governed by strict principles of deduction, induction etc., has morphed into a project that is heavily reliant on the prior moral judgements of the researchers themselves.

Notwithstanding the above, let us assume that suitable proxies for the *ius in bello* principles are found and that an appropriate data base is available to facilitate machine learning processes. Let us now turn more directly, and briefly, to some of the moral challenges confronted by our (supposedly) morally and legally compliant weaponized autonomous robots.

Appropriate application of the principle of military necessity requires reasonably reliable, morally informed, radically contextually dependent, judgments at the various collective levels, as well as at the individual level. However, given the nested character of the individual and collective ends in play, their necessarily underspecified content, and the need to be responsive to the actions, including counter-measures, of enemy combatants and their leaders, there is a constant interplay between the various collective and individual levels. Further, the various applications of the principles of necessity, proportionality and discrimination are interdependent, e.g. the application of the principle of proportionality depends on considerations of military necessity and vice-versa. Accordingly, there is a need to adjudicate not only between the means to given ends, but also with respect to the moral weight to be accorded different competing ends at different levels. For example, the individual end to advance to assist a comrade-in-arms coming under heavy fire might compete with the collective end of one's platoon or company to make a tactical retreat to avoid heavy losses. Again, the collective military end to win firefights might be facilitated by relatively permissive rules of engagement (ROE), but perhaps this end competes with the collective end to avoid large-scale casualties among civilians and the latter end is facilitated by a relatively restrictive ROE. Further, at the macro-collective level, the collective end of the military leadership to win an internecine war might compete with the collective end of the political leadership not to inflict losses of a magnitude that would undermine the prospects for a sustainable peace.

So reasonably reliable, morally informed, multi-level contextually dependent judgments with respect to the use of lethal force must apply a principle of military necessity in a context in which there are: (i) other competing, but interdependent, moral principles, e.g. proportionality and discrimination; (ii) a complex and dynamic structure of nested individual and collective ends (many of which are necessarily underspecified) at multiple levels, and; (iii) variable possible responses on the part of 'the enemy' to any putative use of force. The idea that there a set of requirements for the reasonably rational and appropriately morally informed judgments with respect to the use of lethal force in each and every (or, at least, most) relevant situation in war, and that these requirements could be rendered into algorithmic form and programmed into a computer is to say the least doubtful and, in any case, is at this stage just that: an idea.

4 Conclusion

In this paper I have focussed on three areas in which machine learning techniques are used, or it is claimed should be used, and which give rise to problems at the interface of law and ethics. The three areas are, profiling and predictive policing, legal adjudication, and machines' compliance with legally enshrined moral principles. In general terms, I conclude that while machine learning techniques have considerable actual and potential benefits, they also have limitations, and actual and potential ethical downsides. The use of machine learning techniques in profiling by law enforcement agencies is potentially inconsistent with the individual rights of citizens in liberal democracies not to be subject to unwarranted interference by the state. The limitations of machine learning techniques are evident, for example, in relation to compliance with legally enshrined, moral principles in the investigation and adjudication of complex crimes, and in the conduct of complex, morally charged enterprises such as war.

References

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

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S., 2017. "Dermatologist-level classification of skin cancer with deep neural networks". *Nature, 542*(7639), 115–118.

Kleinig, John, 1996. *The Ethics of Policing*. New York: Cambridge University Press.

Harre, Rom, and Madden, E. H., 1975. *Causal Necessity: A Theory of Natural Necessity*. Rowman and Littlefield.

Miller, Seumas and Gordon, Ian, 2014. *Investigative Ethics: Ethics for Detectives and Criminal Investigators*. Oxford: Wiley-Blackwell.

Miller, Seumas, 2015. "Robopocolypse?: Autonomous Weapons, Military Necessity and Collective Moral Responsibility" in (ed.) Jai Galliott and M. Lotze *Super Soldiers: The Ethical, Legal and Social Implications*. London: Routledge. pp. 153-166.

Mittelstadt, Daniel, Carsten Stahl, Bernd, and Fairweather, N. Ben, 2015. "How to Shape a Better Future? Epistemic Difficulties for Ethical Assessment and the Anticipatory Governance of Emerging Technologies" *Ethical Theory and Moral Practice*, 18: 1027-1047.

Saunders, J., Hunt, P., and Hollywood, J. S., 2016. "Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot" *Journal of Experimental Criminology*, *12*(3), 347–371.

Schauer, Frederick, 2003. *Profiles, Probabilities and Stereotypes*. Mass.: Belknap Press.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hassabis, D. 2017. "Mastering the game of Go without human knowledge". *Nature*, *550*(7676), 354.

Zeleznikow, John, 2017. "Can Artificial Intelligence and On-line Dispute Resolution Enhance Efficiency and Effectiveness in Courts" *International Journal for Court Administration*, 8(2). www.papers.ssrn.com.