# Machine Learning, Morality and Law

Professor Seumas Miller semiller@csu.edu.au
Australian Graduate School of Policing and Security Studies at Charles Sturt University (Canberra)
4TU Centre for Ethics and Technology at
Delft University of Technology (The Hague)
Uehiro Centre for Practical Ethics at the
University of Oxford

# Sources

- Seumas Miller and Ian Gordon *Investigative Ethics: Ethics for Police Detectives and Criminal Investigators* (Wiley-Blackwell, 2014)
- Seumas Miller *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016) Chapter 10 'Autonomous Weapons and Moral Responsibility'

# Machine Learning

- Spam removal by pre-programmed rules: (i) detects property e.g. “lottery winner” and applies rule ‘discard emails containing “lottery winner”’.
- Spam removal by machine learning:

(i) detects property e.g. “lottery winner” and applies rule ‘discard emails containing “lottery winner”’;

(ii) discovers that further property, “Nigeria”, is statistically correlated with existing detected spam;

(iii) generates adjusted rule ‘discard emails containing “lottery” and/or “Nigeria”’

# Machine Learning

- Machine/computer detects physical properties, e.g. noises or marks “lottery winner”, but does not understand meaning of sentence “lottery winner”.
- “lottery winner” is a proxy for fraudulent emails i.e. a species of spam
- Statistical correlation in the past between “lottery winner” and fraudulent emails
- If fraudsters know that the algorithms the machine/computer is using to discard their emails, then they can defeat the algorithm; i.e. past does not determine the future

# Machine Learning and Law

- Predicting future legal outcomes of cases based on past outcomes assumes:

(i) large data set of past cases;

(ii) new cases have similar features to past ones

- Determinations of likelihood of success in application for legal aid are based on outcomes of past cases and weighting of criteria used in these past cases
- Algorithms can change law:

Algorithms yield knowledge of, for instance, malpractices in past cases; knowledge of past malpractice can and ought to inform current practice of doctors; failing to act on this new knowledge might form basis for current malpractice lawsuits

# Machine Learning and Law: Limitations - Particularity

- Contested complex criminal cases might be less amenable, given particularity of case
- Robert Black was serial rapist/murderer in UK but only circumstantial evidence in each murder case; prediction in each case would have been 'not guilty'
- Evidential link between cases: modus operandi of Black
- 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.
- NB: Point is not merely that there was a pattern i.e. the MO – did not need machine learning to discover pattern in a handful of murder/rape case; rather a discretionary legal decision was made to allow an evidential relationship between different cases to be useable in a single discrete case.

# Machine Learning and Law: Limitations - Proxies

- Machines sensitive to physical proxies e.g. wearing uniform and bearing arms is proxy for combatant, i.e. morally legitimate target
- Morally legitimate target is not conceptually equivalent to its proxy; referent and its proxy might come apart
- Machine necessarily tracks - and kills - proxy e.g. anyone wearing uniform & bearing arms, and does not kill non-proxies, e.g. anyone not wearing uniform & bearing arms
- Moral agents tracks and kills referents of concept of morally legitimate target and does not track and kill non-referents;
- Moral agent uses proxy only as a defeasible rule of thumb; therefore, unlike machine, does not necessarily track – and kill - innocent civilians wearing uniform & bearing arms and may well track and kill combatants not wearing uniforms & bearing arms.

# Machine Learning and Law: Rational Choice Theory

- Game-theory/rational choice is controversial qua normative theory as well as qua descriptive theory of rationality
- Outcomes of actors engaged in dispute resolution operating in accordance with rational choice principles may deliver collective irrationality, e.g. tragedy of commons, or manifestly unfair outcome (Nobel Laureate Elinor Ostrum quipped "Mancur Oslen's 'Logic of Collective Action' should have been called 'Logic of Collective Inaction')
- Advice based on assumption of compliance with rational choice principles coupled with knowledge of past outcomes might not deliver fair outcome whereas a highly particular process of morally informed interpersonal rational deliberation focussed on collective end/good may well do so.

# Machine Learning and Moral Responsibility

- Moral -and, therefore, legal?- responsibility to understand algorithm under some adequate description of algorithm, e.g. properties a, b, c with evidential weights w, w*, w** determined legal outcomes in x% of past cases.
- Independent test of veracity of result of application of algorithm (e.g. problem of discriminatory algorithms), given result is a legal or moral outcome, e.g. justice, as opposed to a physical one, e.g. plane or car crashes.
- Discriminatory algorithms, e.g. profiling, under-representation in data of those who never had a loan results in their deemed less creditworthy by algorithm
- Collective responsibility for legal outcomes of designers, producers and operators of machine learning tool, legislators, users (e.g. lawyers, judges)

# Responsibility Gap

- Human operators of autonomous weapons ought to be in-the-loop (human makes final decision to fire weapon) or at least on-the-loop (human can at any time shut down weapon once it is activated), but ought not to be out-of-loop (human does do not make final decision to fire weapon and cannot shut down weapon once it is activated);
- Out-of-the-loop weapons i.e. autonomous weapons, do not remove moral responsibility of human operators, designers, commanders – rather they demonstrate irresponsibility
- So autonomous weapons do not demonstrate moral responsibility gap; only abnegation of human responsibility
- Autonomous cars, auto pilot ought to meet safety standards and have driver/pilot on-the-loop

# Machine Learning, Morality and Law: Models?

- Chess is not a good model for morally informed reasoning or for law; but chess is tailor-made for machine learning
- Chess: (i) rules entirely specifiable in physical terms and governing all possible moves; (ii) fully determinate goals e.g. checkmate; (iii) multiple means specifiable in advance, albeit very large number of possible combinations of moves.
- Morality: (i) motive: act for the sake of moral rule i.e. not mere compliance;
- (ii) following a rule is not same as following a proxy rule – see slide 7 above;
- (iii) choosing ends because one cares about ends, e.g. justice;
- (iv) ends not capable of being fully specified in advance, e.g. happy life, just outcome in conflict resolution
- Law: Pulled in two different directions but not reducible to either model of morally informed reasoning or that of chess/machine learning
- (1) Law ought to considerable extent reflect morality; law is akin to rules of morality
- (2) Law ought to be specified clearly and completely; law is akin to rules of chess/machine learning

# Moral Principles for Conduct of War: International Humanitarian Law

- Principle of discrimination: do not deliberately kill civilians (as opposed to combatants)
- Principle of military necessity: only kill persons if militarily necessary e.g. to win war
- Principle of proportionality: do not kill a disproportionately large number of persons, e.g. do not bomb an enemy position in a village to kill two enemy soldiers if it puts at risk lives of 100 villagers, do not use nuclear weapon (Hiroshima?) to win war

# Programming Legally Enshrined Moral Rules?

- Ronald Arkin has argued that legally enshrined moral principles, such as military necessity, proportionality and discrimination, can be reduced to rules, and these rules can be programmed in to computers
- Therefore, machine trained robots can, at least in principle, conduct war in accordance with international law (IHL)

# 1st Consideration: Non-reducibility of Moral to Physical

- 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.
- 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, this is extremely doubtful especially relatively vague and quite general ius in bello principles

# 2nd Consideration: Military Necessity Not Determined by Rules

- Application of principle of military necessity involves reasonably reliable, contextually dependent judgments at various collective and individual levels, and across different theatres of war
- Given nested character of individual and collective ends, their necessarily underspecified evolving content, and need to be responsive to actions, e.g. counter-measures, of enemy combatants and leaders, there is constant interplay between the various collective and individual levels and different theatres of war
- For example strategic commanders at headquarters and (i) combatants in a firefight in context of a ongoing battle, and; (ii) air attack in context of an ongoing second battle.
- No rule or set of rules  can determine in advance what is militarily necessary
- NB: Commander programming in to computer his/her prior judgment as to what is militarily necessary is NOT equivalent to computer operating in accordance with principle of military necessity (or proxy thereof)

# 3rd Consideration: Military Necessity and Proportionality

- Proportionality is relation between moral weight attached to loss of innocent civilian life and moral weight attached to military necessity
- Given what counts as militarily necessary cannot be determined in advance, there is conceptual but non-algorithmic relationship between principles of necessity and proportionality
- Judgments of proportionality require irreducibly normative process of morally informed deliberation
- Machine learning provides the results of an ongoing descriptive calculative process based on data consisting of past morally informed military decisions.
- Machine learning does not eliminate normativity/morality; rather prior or past moral decisions are embedded via proxies into its algorithms
- Problem: past moral mistakes infect current moral decision-making