Machine Learning and the Law: Five Theses

Thomas Burri
Assistant professor of International Law and European Law at the University of St. Gallen, Privatdozent, Dr. iur. (Zurich), lic. iur. (Basel), LLM (College of Europe, Bruges), admitted to the Bar of Zurich
thomas.burri@unisg.ch, www.thomas-burri.com

Abstract

This paper proposes five theses with regard to machine learning and the law. The idea is to offer some food for thought and highlight elements of machine learning systems that are relevant to the law. Since the field is new for lawyers, the paper necessarily is preliminary in nature. While some parts may prove controversial, the main idea is to stimulate discussion, rather than to state absolute truths.

1 Thesis 1: While machine learning systems pose challenges to the law of liability, a combination of legal considerations and technical measures is capable of addressing them.

The behaviour and actions of machine learning systems are not fully foreseeable in all situations, even when the algorithm directing the learning is known. A machine learning system's behaviour is based on the patterns and correlations it discovers in datasets. These patterns and correlations by nature are not known in advance, else no learning would be required. That the behaviour is not fully foreseeable poses a challenge to civil and criminal liability regimes, since the law usually holds users, producers, programmers, etc. only liable, when a behaviour was foreseeable. (When one causes harm intentionally by means of a machine learning system, e.g. by programming a malicious algorithm or influencing a data set so that the system causes damage, liability is relatively straightforward.) If behaviour was not foreseeable, there is no one to blame for harm done by a machine learning system. This may result in a responsibility gap.

Certain theoretical considerations and technical measures increase the law's capacity to deal with unforeseeable behaviour of machine learning systems. (i) One relevant consideration is the perspective of the victim suffering harm at the hand of a machine learning system. Philosophy teaches us that for the victim suffering harm it is irrelevant whether, from the perspective of the perpetrator, it had been foreseeable that the harm would be caused (see Birnbacher, 2016). For the victim harm is harm, done to its dignity, no matter how foreseeable it had been. Whether it had been a human driver who texted while driving or an autonomous car that did not recognize a pattern – the victim the car hits suffers the same harm. Foreseeability in the first case does not render the harm more acceptable for the victim than in the second case. From this perspective, foreseeability is irrelevant. (ii) The behaviour of humans in concrete situations is not fully foreseeable, either. Especially in situations calling for rapid reactions humans' reactions may differ widely. (iii) For the purpose of civil liability, insurance reduces the relevance of the unforeseeability of machine learning systems' behaviour. Insurance for conventional cars, for instance, need not be fundamentally changed to transpose it to autonomous cars; inter-insurance offset works without that fault or blame need to be assigned. (iv) Criminal liability for negligence is likely no longer necessary. In the future, testing of machine learning systems combined with certification by the state will eliminate the possibility of negligence. The more likely physical harm, though,

the more rigorous both testing and certification will have to be. (Criminal law in general does not become superfluous; rather, its scope will be limited to intentional acts and omissions – which will be hard enough to distinguish from negligent behaviour.) (v) Foreseeability should in general be distinguished from expectations inbuilt in the law. The law expects certain behaviour in specific situations, e.g. that a driver brakes when a collision becomes unavoidable. In a similar vein as with foreseeability, both humans and machine learning systems may on occasion prove (in)capable of meeting such expectations. However, machine learning systems will likely be more capable of meeting expectations originally formulated with humans in mind. An autonomous car may be capable of hitting the brake more rapidly than a human driver. This in turn may feed back into expectations, shaping them over time. (vi) In certain situations, it may be helpful to take likely human behaviour as a point of reference, if only to establish a bottom line. It may be asked what a human being would have done in the stead of a machine learning system which caused damage. Would he or she have possibly overlooked the truck the car hit? If the answer is negative, the producer must be held liable, regardless of unforeseeability; if it is in the affirmative, the solution should depend on whether testing had been done lege arte. (vii) Others have shown that full transparency of the underlying algorithm is not necessary to establish liability; there are technical alternatives, notably cryptographic commitments combined with zero-knowledge proofs (see Kroll et al., 2016). (viii) It is not just foreseeability that is limited with machine learning systems; probably the reasons why a machine learning system behaved in a certain way in a specific situation cannot always be fully explained ex post either, or at least not without unreasonable efforts. On the one hand, though, linear models – or for that matter humans – do not always fare better when it comes to ex post explanations (see Lipton, 2016). On the other hand, the currently emerging right to have 'decisions' taken by machines explained should be interpreted accordingly. A basic explanation – such as the explanation that incoming sunlight blinded a sensor in a specific case or more generally an explanation by means of maps and pictures – should suffice.

2 Thesis 2: Machine learning systems force law- and policymakers to define the essence of humanity.

The potential of machine learning systems and their rapid development forces law- and policymakers to confront the question which functions, decisions, roles, etc., if any, necessarily need to remain in the hands of humans at all cost and all times. Is there any essence of humanity which should not be transferred to machines under any circumstances?

Clearly, driving a car is not essential for humanity, at least not for the overwhelming majority of human beings. Conversely, it seems almost as clear that killing human beings is part of this essence. Machines should not learn how to kill humans, or at the very least humans should retain significant control over the final decision in a concrete instance to kill a human being. This is what the discussions in Geneva on autonomous weapons systems are about. However, perhaps we would not be equally unapologetic if the setting was not armed conflict but civil life. What about a machine capable of euthanasia? Or one that waits at the end of death row, or at the beginning of life? At least on the face of it, in these highly controversial situations there may be certain advantages in transferring the final act or decision away from a third human person involved, such as the doctor or the executioner, to a machine.

Further hard decisions loom. Voting in democracy and the process accompanying it are certainly essential. Accordingly, they should be insulated from the influence and control of artificial intelligence, be it of the machine learning or of another kind. However, illegitimate 'influence' and 'control' may be hard to distinguish from legitimate functions. In contrast, personal care of patients and the elderly is probably less essential. At least the rapidly expanding supplementation of human care by robots indicates so. Finally, sitting on boards of companies appears less essential than sitting as jurors in criminal law cases and judges more generally. With all these cases, it should be kept in mind that the essence of humanity might not be the same for all humans.

3 Thesis 3: Machine learning systems end up making laws.

The law made to apply to machine learning systems should be distinguished from the law

made to govern machine learning. On the one hand, there is the law – data protection law, civil law, criminal law, etc. – that determines what machine learning systems lawfully may (not) do; on the other hand, there is the law that determines how a machine learning system behaves. The first is about objective application of the law to a system, the second is about making machine learning systems work by means of code and programming. Ideally, the two will merge at one point, namely when the law is programmed into machine learning systems, ultimately allowing them to respect the law. To reach that point the law will need to be adapted to some extent in order to fit the capabilities of machine learning systems. The law, thus, will not only influence machine learning systems, but they will also influence the law vice versa.

Yet, there might also be a subtler evolution leading to a new kind of law. 'Soft law' may emerge from the behavioural interaction among machine learning systems. Consider the interaction among human car drivers which determines what each of them will do in specific situations. Eye contact among human drivers, for instance, may be enough to determine who gives way; or, more generally, specific modalities may arise around roundabouts car commuters use daily. Similar patterns could come to govern some of the behaviour of autonomous cars as well. Certain behavioural patterns, and even rules, may emerge with machine learning systems, especially when they interact in large numbers. A similar phenomenon has notably been observed with robots in swarms (see Rubenstein et al., 2014). However, such patterns and rules will arise unpredictably. Observers are, accordingly, reduced to simply stating their emergence – which will be a new position for lawmakers used to establish rules.

4 Thesis 4: Randomization is a way out of ethical dilemmata confronting machine learning systems.

Machine learning systems force humans to take early, explicit decisions with regard to certain ethical dilemmata which they have hitherto avoided until the very last moment. Witness only the old 'trolley-problem' which is being vigorously discussed anew with regard to autonomous cars, because decisions need to be taken earlier and programmed into systems. This early confrontation with dilemmata may have the added benefit of increasing ethical clarity (see Kolmar and Booms, 2016). However, despite all clairvoyance some dilemmata may prove hard to solve. And while sensors and accurate prediction may prevent certain ethical dilemmata from arising concretely in the first place, some of them will inevitably have to be solved. It seems that in these truly hard cases – namely true dilemmata in concrete cases where ethics fails to identify a clearly preferable option – a randomly taken decision should determine the way forward. Is it not the case that when only two equally bad options are available, it is fairest and most acceptable to all involved when the choice is taken by chance? Making randomization public in advance may further increase fairness. Computers are good at randomization and can apply it instantly, in contrast to humans. Besides, what else are humans doing when it comes to taking ethically sensitive decisions rapidly on the spot than applying a sort of imperfect randomized decision? Finally, randomization would certainly be preferable over sacrificing the many potential advantages of machine learning systems. A handful of hard cases, which occur rarely, should not be allowed to drag down the whole system.

5 Thesis 5: Machine learning systems and the law share an affinity for structured environments.

Machine learning systems work best in structured environments. Computer or board games usually present such environments. Hence machine learning systems are capable of devising successful strategies. The internet is structured, allowing search engines to develop strategies to display good results. Offices, as structured spaces, are ideal for vacuum cleaning robots. In contrast, natural, 'real world' environments are less structured, making machine learning more challenging. Local roads, for instance, are unstructured and therefore not as amenable to machine learning cars as highways. A kindergarten is highly unstructured. Machines are therefore unlikely to make good kindergarten teachers.

Describing an environment as highly structured is merely a colloquial way of saying that it lends itself well to statistical analysis and inference, upon which most machine learning is

based. The law may be considered to be part of the 'environment' and thus a variable of 'structure' to be incorporated in machine learning. But this is not the only possible perspective. The law may also be considered to be extrinsic, a lens, so to speak, to look through. In this perspective, the law shares machine learning systems' affinity for structured environments. Highly structured environments are thus easily regulated. Games usually are based on few, simple rules. Only a few rules are necessary to allow highway traffic to flow. In contrast, when the environment is less structured, the law is challenged more. Local traffic is subject to numerous rules, including stop signs, traffic lights, roundabouts, differentiated speed limits, pedestrian crossings, etc. – not to mention the countless subtle rules drivers rely upon when communicating spontaneously. Any attempt to lay down the rules governing kindergartens, beyond highly general and abstract norms, would be pointless; too many rules would be needed, too much discretion would be involved.

Given this affinity, law- and policymakers should keep an eye out for highly structured environments. This is where machine learning will likely be applied in the near future. The constitution of the law itself may even point the direction. Where the law operates with open norms, abstract concepts, and flexible discretion and where lawful behaviour depends on circumstances, machine learning is less likely to be deployed soon.

References

[1] Birnbacher, Dieter, 'Are autonomous weapons systems a threat to human dignity?', in Bhuta, Nehal; Beck, Susanne; Geiß, Robin; Liu, Hin-Yan; and Kreß, Claus (ed.), Autonomous Weapons Systems – Law, Ethics, Policy, Cambridge, Cambridge University Press, 2016, p. 105-122.

[2] Kolmar, Martin; Booms, Martin: Keine Algorithmen für ethische Fragen, <http://www.nzz.ch/meinung/kommentare/keine-algorithmen-fuer-ethische-fragen-ld.4483>, last visit: 26 January 2016.

[3] Kroll, Joshua A.; Huey, Joanna; Barocas, Solon; Felten, Edward W.; Reidenberg, Joel R.; Robinson, David G.; Yu, Harlan, 'Accountable Algorithms', (2016) 165 University of Pennsylvania Law Review (2) 60.

[4] Lipton, Zachary C., 'The Mythos of Model Interpretability', (2016) ArXiv:1606.03490v1 (10 June 2016) 96.

[5] Rubenstein, Michael; Cornejo, Alejandro; Nagpal, Radhika, 'Programmable self-assembly in a thousand-robot swarm', (2014) 345 Science (6198) 795.