[Review of] Chesterman's We, The Robots

Book Review

We, The Robots: Regulating Artificial Intelligence and the Limits of Law by Simon Chesterman (2021) Cambridge University Press, 289 pp, ISBN 9781316517680

John Zerilli*

Please cite this book review as:

John Zerilli, 'Book Review: *We, The Robots: Regulating Artificial Intelligence and the Limits of Law* by Simon Chesterman' (2022) 44(3) Sydney Law Review 499.

The preface to Simon Chesterman's *We, The Robots*[1] signals its intended readership: those concerned with regulating the activities of artificial intelligence ('AI'). But with a subtitle like '*Regulating Artificial Intelligence and the Limits of Law*', the book was always going to entice the technologically savvier members of the legal profession — practitioners, scholars, judges, etc — and so not just the regulators. Unfortunately, however, practitioners and scholars, and possibly even the regulators themselves, are likely to hanker for more direction than the author provides. Many of Chesterman's discussions have a whiff of ambivalence about them and conclude at just the point where a keen observer of the subject would like to know more. For example, in winding up a lengthy discussion on negligence,[2] Chesterman states that 'for the purposes of tort liability [the process by which AI systems make decisions] raises the question of whether an autonomous system's behaviour could itself constitute a new intervening act that avoids liability'.[3] Tantalisingly, that is just

* Chancellor's Fellow in AI, Data, and the Rule of Law, University of Edinburgh, Scotland; Research Associate, Oxford Institute for Ethics in AI, University of Oxford, England.
Email: john.zerilli@ed.ac.uk; ORCID iD: https://orcid.org/0000-0002-7010-2278.

1 Simon Chesterman, *We, The Robots: Regulating Artificial Intelligence and the Limits of Law* (Cambridge University Press, 2021).

2 Ibid 88–91.

3 Ibid 90–1.

how the discussion began: 'In relation to causation, in some circumstances AI systems … may constitute an intervening act in their own right.'[4]

This is what you get, I suppose, when a book offers a *tour d'horizon* — and make no mistake, the book is a masterful catalogue of practically every issue that has been raised in the past six years of law and technology scholarship. In the fever of cataloguing, however, answers are either not forthcoming, underspecified, or indeterminate. To be fair to Chesterman, perhaps this is because he is ever mindful of his target audience: the civil servant exercised by the practical imperatives of government policy. But occasionally too, things get bundled together that should probably be kept separate. For instance, when talking about the due process requirements of outsourcing to AI, he cites jurisdictions that have banned the use of AI for facial recognition.[5] But because the discussion in this part of the book is meant to connect with what elsewhere in the book he calls 'matters of legitimacy'— functions which it is not immoral to outsource to AI provided that due process measures are in place to ensure human accountability — it is slightly confusing to see outright bans on facial recognition systems being mentioned here. Limits on the use of AI strike me as fitting more naturally in discussions of what instead he calls 'matters of morality' — applications of AI that are inherently, deontically, objectionable, and which ought to be seen as posing 'red lines'.

So much for general appraisal and criticism. I also have a somewhat more specific complaint, and this is that I was not convinced by Chesterman's argument that our systems of civil liability must inevitably produce 'accountability gaps'. To his credit, Chesterman does emphasise a number of times, particularly in the later chapters of the book, that our current civil liability regimes 'will cover the majority, perhaps the vast majority, of AI activities in the private sector',[6] and I fully agree. Curiously, however, the chapter that examines this issue at length (Chapter 4) reads — and concludes — with somewhat less conviction. For instance, he writes that 'the speed, autonomy, and opacity of AI systems will give rise to accountability gaps … future cases will arise where there is a harm not attributable to a person or a company'.[7] As an example he gives: 'the death of a child hit by an unidentified drone … or killed in error by a lethal autonomous weapon'.[8] But in neither of these two cases does it seem to me that we should be in any doubt about the law's resourcefulness, even as it stands.

The first case is equivalent to a hit and run where no one is around to witness it and the defendant remains unknown. That is not a case of the law running out, or an accountability gap, so much as a case of our not knowing to *whom* the law applies — a situation hardly unique to cases involving AI systems. In the second case, in which someone is killed in error by a lethal autonomous weapon system ('LAWS'), again, nothing Chesterman had to say on the topic convinced me that the principles of tort, soundly applied, would produce accountability gaps. On the contrary, I am inclined to think someone can almost always be held liable, even in cases illustrating

---

4 Ibid 88.
5 Ibid 190–1.
6 Ibid 187. See also earlier in the book, eg, at 38.
7 Ibid 112–13.
8 Ibid 113.

the 'problem of many hands', so long as someone is at fault somewhere in the chain of events (and often enough, even when *no* one is at fault in the chain of events). Moreover, I remain to be convinced that it makes sense to think of the interstitial crevices of unsettled law as the sorts of things that give rise to 'accountability gaps'. At least, I am not sure we should worry about them for AI more than we usually would without AI.

Every case will obviously come down to its own facts, and it is trite to point out that sometimes claimants simply do not deserve to win (whether in virtue of their contributory negligence or something else). But after all the evidence is in, are we really meant to doubt that someone somewhere in the chain will be held liable, when they *ought* to be, either through the application of product liability principles, vicarious liability, non-delegable duties, apportionment and contribution principles, the maxim *res ipsa loquitur*, and of course, the principles of causation and remoteness of damage? That last principle, in particular, can do quite a lot of heavy lifting. People worry that because machine learning algorithms *learn* to do things for themselves, including any errors, this somehow raises the prospect of a danger being inherent in the software for which no human can be identified as responsible. Chesterman seems worried by this too, because he notes more than once that the autonomy of an AI might make *it*, as opposed to its *manufacturer*, responsible for harm. But this has long struck me as a non-starter. Even the intervening actions of third parties do not break the causal nexus between tortfeasor and claimant harm so long as the intermediary's intervention was (roughly) of such a kind as to be reasonably foreseeable in all the circumstances. So why should it be different when the intermediary is an artificial agent — indeed one designed or developed by the defendant for commercial gain?

Take the example of an autonomous vacuum cleaner. As the programmer, you want it to avoid bumping into furniture. So, the reward function might be something like, 'avoid the sensors at the front of the vacuum cleaner coming within a certain proximity of objects'. To the householder's dismay, the system learns to maximise its reward in a most unorthodox way, by simply travelling *backwards*. In this manner, the vacuum cleaner fully maximises its rewards, despite bumping into furniture left, right, and centre, simply because its sensors are positioned at the front of the device! To my mind, this is just the *kind* of thing that could go wrong with a machine-learning-driven vacuum cleaner, and which falls unquestionably within the field of its manufacturer's reasonable foresight.

I do not doubt that incremental adjustments here and there will be required to our civil liability regimes, such as an amendment allowing software to be considered a 'product' under the *Consumer Protection Act 1987* (UK). But these are changes one would expect in the ordinary course of legal evolution anyway. Indeed, they are already in the wings: we do not need AI to educe these developments, although it may well precipitate them.

Standing back from all this and reflecting for a moment on the common law, perhaps the more pertinent question is whether the legerdemain of our judges *should* be relied upon to accommodate technological innovation. One drawback of squeezing all we can out of existing legal doctrines is complexity — the United Kingdom's common law of privacy bears witness to the messiness that has been

tolerated out of deference to extant legal categories. But arguably this is not half as bad as what happens in rights discourse, where there seems to be genuine difficulty in applying the legal equivalent of Ockham's Razor to declarations of rights (the difficulty that, in pressing as much consequence as one can out of, say, the right to life, one brings about an unprincipled and potentially self-defeating inflation of the right, to say nothing of conceptual confusion). In matters of doctrine and doctrinal evolution, by contrast, a kind of theoretical parsimony is arguably exactly what is called for. To take only one example, the tort of injurious falsehood started off as an action against allegations of false title to land — hence its early name, 'slander of title' — but its scope soon extended to encompass all manner of aspersions cast on a claimant's goods and business dealings, to the point where it could even be brought for plainly defamatory imputations, as well as for a kind of passing-off ('reverse passing-off'). Conceptual confusion did not inevitably result.

Chesterman is on firmer ground when he notes the difficulties attending any attempt to prosecute war crimes committed through LAWSs. But this raises other issues. Chesterman thinks that 'some decisions over life and death require that a human soul grapple with them'[9] because it is important for humans to be accountable for them.[10] Indeed, he thinks this consideration has precedence over any argument seeking to justify the use of LAWSs on the basis of their superior performance. If that is the background assumption — that someone *must* be accountable for mishaps involving LAWSs come what may — then yes, difficulties in tracing responsibility along the chain of command will be a decisive consideration in any decision to deploy them. But what if that is the wrong assumption? Wouldn't the fact that LAWSs might eventually be much less prone to misidentify targets then count for more than being able to pin the blame on someone? Given the choice between a high probability of being killed by someone that has wrestled with their conscience, and a low probability of being killed by a piece of kit, is it not rational to plump for the latter? Perhaps in time, we might even come to see accidental death by LAWSs as akin to death by natural causes — acts of God, earthquakes, or volcanic eruptions, say — for which no one need be blamed.

For sheer breadth of coverage, Chesterman cannot be faulted. As I said before, there is scarcely an issue that has been discussed among the cognoscenti of law and technology over the past few years that does not receive even a touch of Chesterman's prodigious learning. But if I were to sum up my estimate of the book it would be that, for all its deft integration of material, it lacks a satisfying, cohesive theoretical vision to contain what Tennyson, speaking of the common law, once described as 'that wilderness of single instances'.[11]

---

9 Ibid.

10 Ibid 104.

11 Alfred Lord Tennyson, 'Aylmer's Field' in *The Works of Alfred Lord Tennyson* (Wordsworth Editions, 2008).