BOOK REVIEW

Robot Ethics 2.0: from Autonomous Cars to Artificial Intelligence, Patrick Lin, Keith Abney and Ryan Jenkins, 2017, Oxford University Press, Oxford, ix + 436pp., £31.50 (hardback) ISBN: 9780190652951

Sitting in my basement, a Zoom workshop about compassion for student trauma while teaching online during a pandemic on one screen, and a livestream of day seventy of well-equipped, highly-technologized, militarized police violence enacted on journalists and peaceful protestors in Portland, Oregon on another. I can't help but feel the weight of the technological matrix in which I find myself already embedded. Much like the early empiricists understood our senses to be the fundamental mediator between us and the world, it makes just as much sense to point to our technologies as the meaningful mediator of our everyday lives. Of course, this idea is hardly new or generally insightful on its own, but when the world shuts down almost overnight and many schools and businesses transfer their operations to peoples' homes, with vast wealth inequality reflected in who must remain in the risky, uncontrolled pandemic world and who gets to sit comfortably back and interact with the world safely from a technology-mediated distance, these questions press more heavily on all of us. As some developed nations deploy controversial tracking technologies which are helping to control their outbreaks while also compiling large quantities of data on their citizens, other countries fail to deploy any social or digital technologies in the service of public health, and so all sorts of inequalities and underlying belief structures are laid bare. Students across the United States wonder how their educations will continue, angry that bodies-in-rooms classes are being replaced with technologically-mediated systems of synchronous and asynchronous videos, and left wondering why in-person learning and socialization feels so much better. And while Lin, Jenkins, and Abney entitled their book *Robot Ethics 2.0*, it is not about just robots; it's a technology ethics book through and through.

There is some indication here of how quickly the world of technology ethics moves; when the first version of this book was published by MIT Press in 2012, self-driving cars were barely worth a passing mention. Yet, when this follow-up book was compiled in 2017, the authors and editors recognized self-driving cars to be the ideal vehicle to examine the state of

both emerging technology ethics and the legal messiness that always follows important questions and judgments about responsibility and blame. They are explicit about this in the introduction:

> As the first robots to be integrated with society at any significant scale, self-driving cars may set the tone for other social robotics, especially if things go wrong. Therefore, we will use robot cars as a crucial case study throughout this book. (p.ix)

But now, in 2020, as the paperback edition of the book makes these twenty-four papers more readily available, it seems like another decade has passed since 2017: many of us, in the US especially but not exclusively, find ourselves locked away in a country that denies the science of a global pandemic, watching the wild effects of disaster capitalism emerge through not just a coronavirus pandemic, but, in the US in particular, through a reckoning with centuries of racism best understood as a public health emergency of police violence enacted largely on black folks. And self-driving cars? We don't hear quite as much about them right now, in spite of the fact that every year a handful of companies report that they intend to have fully-autonomous self-driving cars on the roads the very next year. We're still waiting.

Yet, this text is generally wise in both its promise and delivery of the self-driving car content; it somewhat overstates how much of this you'll find here. Yet this is a tough book to wrangle, in part because it covers just so much ground; its twenty-four chapters are split into four sections, covering the broad themes of legality and coding, trust and deception in design, applications of robotics (in situations as varied as love and war), and finally, broader questions about the future of artificial intelligence (AI) in general. Each of these chapters and sections includes various emphases on specific philosophical theories of ethics, or the current legal landscape for certain kinds of automation. Some of these essays are instant classics (indeed, some were classics before publication, such as Kate Darling's contribution 'Who's Johnny', which circulated widely on the internet for several years before finally finding its publication home here. And yet, some of these essays are instantly… whatever the opposite of classics is. There are a few (only a few) chapters here which are embarrassing for their authors, or which should be. My goal here is to offer an overview, then, of the book in general, pulling out examples of chapters that work brilliantly and would serve both as excellent research touchstones and good teaching materials, without shying away from those chapters that left me wondering how they were ever published in the first place. Let's try and start with the good: there's lots of it here.

It's well-known that technology scholarship goes stale faster than similar work, in part because the technology itself moves quickly and our 'hot takes' one month become absolutely unreadable the next, when the technologies we spent our time dissecting and critiquing prove to have been short-lived and easily forgotten. On the other hand, when we fail to predict certain types of technologies (or creative uses for the more mundane), it can also make the work seem quickly outdated and myopic. *Robot Ethics 2.0* gets a lot of this right. In spite of the heavy reliance in some chapters on self-driving car technology (which may still pan out, in which case the authors gambled and won with this book), there are a lot of big-picture papers here, as well as a lot of papers about ethics and legality that should, in principle, apply beyond the specifics. One of the strongest themes running through the book (as much as themes can apply to so many disparate papers) is that the philosophical approaches, whether working up from theory to application, or from application back down to theory, do matter, and should matter, beyond just armchair speculation. As the chapters focused on legal scholarship point out, this theory stuff needs to have regulatory teeth.

The second chapter, 'Ethics settings for autonomous vehicles' by Jason Millar, is one of the standouts here. The premise gave me anxiety; I was afraid he was going to offer a simple utilitarian approach and set us up for more 'here's how you solve the trolley problem' pontificating, but instead what Millar offers is a demand that, ultimately, we "embrace ethical and regulatory complexity where complexity is required" (p.31). There are no oversimplified answers here; instead of just reporting that ethical theory X leads to Trolley Problem Solution Y, the argument Millar offers leads us to recognize that this is not the right question. There's nuance here. While he starts us in a place where it appears we will be forced into answering what ethics settings we should program into self-driving cars, by the end it's clear that the real question we should be asking is Who gets to make this decision? There's a cheap conceit here that turns up in much fiction: Amazon's 2020 show *Upload* featured a near-future with self-driving cars in which the ethics settings were a user-controlled option. The conceit is used, in part, as a storytelling device to reveal the kinds of people two of the main characters are: one always keeps the ethics settings on 'prioritize occupant' and the other on 'prioritize pedestrian', and we learn who the hero is supposed to be just like that. Millar provides some good insight here: "… it seems that engineers do not have the moral authority to make ethical decisions on behalf of users in hard cases where the stakes are high" (p.25). This seems exactly right, especially in cases of life and

death. We have versions of this now, with doctors needing informed consent around end-of-life care. The potential for ethics settings that require user input can seem a bit silly until we think in these terms. This is one way these chapters gesture beyond themselves; they are not just about ethics setting in autonomous vehicles, but about ethical expertise and regulatory structures that rest ultimately on the right kinds of expertise. Importantly, Millar doesn't make the same mistakes here that some of the other chapters on this topic do when they make broad proclamations about how (for example, Loh and Loh in Chapter 3) in the future we'll be required to complete a questionnaire that would populate a moral profile before we'd be able to get into an autonomous car so that blame can be attributed in the case of an accident. The problem with so many of the arguments framed in this way is that they allege self-driving cars bring something new in kind, not just in degree, which demands special kinds of moral examination. For example, in one chapter, we're told:

> … a general societal discourse is needed in order to raise awareness of these moral dilemma situations and prepare drivers for the responsibility they bear of making moral choices that can have a huge impact on themselves, their co-passengers, and other traffic participants. (p.46)

The call for societal discourse around ethics and technology is very welcome, but there's nothing new here in the potential level of impact or the people involved. Every time we sit in our mundane, not-self-driving cars now, we take on the exact same responsibilities. I may be just as likely to face a real event where I have to choose which way to steer the car and whose life to risk in a crash (most crashes probably fit exactly this model). When we sit inside our high-speed, 1500kg containers of metal and glass, we risk facing the exact same questions, and we don't ask anyone to fill out a moral profile before driving. We demand only that drivers have passed a simple test of rules (rules designed to maximize safety and minimize loss of life!) and then we let them loose.

But beyond the self-driving car chapters, there are some really thoughtful pieces here that deserve more attention than I can possibly give them in this space. In Chapter 6, 'Skilled perception, authenticity, and the case against automation', David Zoller lays out an elegant argument about what is lost from human life and experience when we hand over some of our regular activities to automated systems of various sorts. In this case, automated cars aren't centered, but offer a nice application of the larger principle he proposes. Zoller weaves

Merleau-Ponty and Husserlian phenomenology together with Gibsonian affordances and Noë-style situational experience to claim that skilled perception is more than (and more special than) merely an action to hand over to automation without real consideration of what is lost. The argument isn't technophobic, but rather offers robust and solid theoretical grounding that ought to give pause to anyone who wants to hand too much of our lives over to algorithms. Virtue ethics meets phenomenology to demonstrate the nuances in how even algorithmic bias works (without quite using that language). He urges caution, claiming:

> Given that automating a skilled activity means agreeing that we will exit some niche of perceptual reality, and maybe exit it forever, we should be confident that either (a) there was nothing of importance to see in that niche, or (b) if there was, we know how to synthesize or artificially reconstruct that meaning… (p.86)

It's a welcome call for a more humanistic approach to what we lose and gain with various sorts of automation. This is a chapter I didn't realize I wanted in the book until I read it. It is also one chapter among many that applies to self-driving cars without hinging an entire argument on them: they are a demonstration of the larger principle rather than the reason for the argument's existence.

Another stand-out chapter (7 - Meacham and Studley) focuses on care robots, but rather than centering the robots and their intentions (or lack thereof) or blame and praise for designers, it argues rather cleverly that such robots are morally permissible based on enactive principles and phenomenology. Meaning emerges in the interaction itself, so such robots create a genuine kind of care-based environment without needing to be agents of care themselves. This was another surprising chapter insofar as the overlap between such work as that of De Jaegher and Di Paolo (2007) on participatory sense-making and practical applications of commercial robots is extremely sparse. The argument itself serves as a nice introduction to this material for those who need it, as well as convincingly claiming that these care robots can be placed in delicate spaces with (for example) elderly folks with healthcare needs without worrying that the automation of previously human labor automatically does harm. This premise could support (and deserves) an entire book-length treatment, and it's always exciting to find unexpected arguments like this.

There are more really great, standout chapters here (Vallor and Bekey's discussion of self-learning robots in the final 'future of AI' section is on a par with their usual work, an outstanding, well-grounded dose of reality, strongly backed up by theory. Then there is Kate

Darling's chapter on anthropomorphism and framing in human/robot interactions, a favorite that situates the debate about whether and when robots should take on human characteristics). But there are also some problematic pieces in this book, and the ways in which they are problematic speak to larger assumptions and issues in the AI/robot/technology ethics community more broadly. So it's worth sitting with them for a bit to understand why and how they're so very problematic. And beyond the good and the bad, if there were more space, I wish we could grapple a bit with the missing: robot and technology ethics continue to overlook the role that scholars of disability studies can play here. For instance, a piece such as 'All technology is assistive' by Sara Hendren (2014) would have elevated the possible discourses around this text, and made the book an even better teaching tool than it already is.

But here are the problems. There are (at least) two chapters in this book that felt like reading posts on the sorts of stereotypical websites where the self-declared hyper-rational congregate (Reddit, for example); overflowing with claims that we should ignore all context and focus on what's rational and logical (as if these can be filtered out of the text, context-free). Sexism and racism are not explicitly called 'irrelevant' or 'illusory' here, but are still dismissed in every line in some of these chapters - and this ignores not only the needed context in which many of these arguments are necessarily made, but also who is actively hurt by their being made. Both philosophy as a discipline and technology studies more broadly are in constant need of a reckoning with these problems, which critics have been pointing out for as long as the fields have existed.

There's one chapter (14 - Bołtuć) that engages an argument about sex robots, arguing that if they are functionally equivalent in all ways to a human partner in sex-relevant respects at whatever level of granularity is necessary, then there's no real moral problem with sex robots. There's an argument here, but before readers can critique the problematic metaphysical assumptions, they have to get past the casual sexism, the strange co-opting of the notion of the uncanny valley to laud "human geniuses who come too early in history to fit with the rest of society…" (p.220), and what can only be described as rape apologia that really fails to meet even the lowest bar I could possibly set for an academic book about robot ethics. In the author's words: "Not to care whether one's lover feels anything at all is especially cold" (p.222). This would be an ignorable sentence if not for a footnote to it, which leads the reader to a single line,

thanking someone else for pointing this out. Fair enough; we don't like philosophers taking intuitions for granted, I guess! The author continues:

> People scoff at those who have sex with stoned drug users, other unconscious individuals (even if their bodies are mobile in some way), not to mention dead bodies, even if there was prior consent - we expect the feeling of some sort. (p.222)

People scoff. At rape! I feel as though the author's words should stand on their own, without needing much explicit critique for people to understand why they are inappropriate here, but then, they've been published, so maybe many people do need this to be made explicit. Please don't excuse rape offhandedly in your academic work. Much casual sexism is also embedded within the argument; the author enumerates labels for a variety of different hypothetical sex robots, and offers:

> … an emotionally engaged Church-Turing lover (such as Harmony) would be caring and behave as a human person in love; it might also play the role of an artificial companion (by cooking, taking care of the partner's health and well-being, etc.). (p.221)

You don't need to be told that the Harmony robot being referenced here is a busty blond woman robot. But again, maybe the idea of such a robot also cooking for its 'lover' doesn't immediately strike you as sexist; it should.

Perhaps unsurprisingly, the section of the book dealing with sex robots is the weakest in an otherwise academically-strong and philosophically-interesting book. There is a chapter on 'lovotics', which the authors describe as "human-robot romantic and intimate relationships" (p.194). But the chapter is just a mess. It opens with eleven pages of defensive justification for why lovotics counts as a legitimate subject of study, and then offers semi-technical specifications for a robot. None of this is going to be particularly valuable to a philosopher or technology ethicist trying to understand the ethical landscape of this (genuine) question about the ethics of love and/or sex with robots. When the authors do finally discuss the ethics, eleven pages in, we're treated to a quickfire barrage of paragraphs, disconnected from one another and without serious engagement, offering brief topics that may be related to the question, and may be quite valuable, but without any real sense of the ideas. None of these philosophical arguments is given any real time or consideration because the essay has used its space stomping its feet in a demand for academic legitimacy - which it could have gained simply by offering these short citations in real context with a real argument.

The book closes with a challenging chapter that cannot possibly be sufficiently explored in a review like this, but it deserves some mention. It is a chapter in defense of the Unabomber's ideas, based on both the public writings of Theodore Kaczynski and the author's personal correspondence with him. The chapter starts with a Reddit-style demand to debate the technological critique of the Unabomber without becoming bogged down in the details of mass murder and terrorism, so we're already off to a bit of a rocky start. Despite my own misgivings, I tried to engage with the project in this way, but it's a very hard project to approach with an open mind. Unironically, the chapter refers to 'modern man', and at times reads like the now-infamous Google manifesto, excusing racism and misogyny time and again and pasting a torn image of rationalism over everything. (The author might claim to be only reporting Kaczynski's use of such phrases as 'innate biology', but the author chose to engage with these particular pieces and it's hard not to see a defense of these ideas written into the chapter.) The author explicitly and repeatedly endorses Kaczynski's views, even when they amount to literal eugenics, and there is no way to read this chapter except as an attempt to rehabilitate someone who allowed his absolute failures of critical thinking to murder people. Again, this reads a lot like common internet manifestos, and it really seems like we only have to contend with these ideas academically because this particular manifesto-writer carried out his plans to kill people and so rose to infamy. There's nothing new here; at times the chapter reads a bit like an inelegant retelling of Samuel Butler's satire, *Darwin Among the Machines* (1863). The author appears to reason that if we're too late to disconnect ourselves from the technological society in which we find ourselves, then it follows that we must morally take whatever action is needed to try and alter that society. I wanted this chapter to be the good kind of challenging: I wanted to have to work to argue with the premises and conclusions. But instead, I was given the likes of this:

> It must, for instance, be admitted that the Unabomber provides some worthwhile insights and that Kaczynski's analyses of autonomy and deprivation of control over one's fate are fundamentally accurate. It is difficult to deny that society has evolved to a point where most people work on tasks that have little to no tangible value outside of being part of a much larger and incomprehensively complex and largely technological process. (p.374)

It must not be admitted and it is not difficult to deny. I hoped this chapter would be a strong closing to the book, challenging readers with a sort of taboo idea but backed up with serious arguments. Instead, it was too much like reading the work of every disgruntled poorly-argued

and largely-unsupported manifesto on the internet. Default assumptions like this have to be among the most poisonous ideas in philosophy.

Again, there are some really good (and some less good) arguments in this book. In general, even in the cases where I think the arguments are wrong, they're just that - philosophical arguments presented for evaluation. It's good that they don't all present the same view, and important that they're not all engaged in exactly the same project. The result is a successful project overall, and in general a valuable resource. A few chapters, however, are wrong not just in their philosophical details, but also because of the implicit or explicit appeals to misogyny, or ableism (stop tying reasoning ability to language ability, please), or the straightforward belief that people should be able to write dangerous things and that anyone who disagrees is irrational or doesn't understand. We do understand. We just wish you could be better. The philosophy of technology, particularly its ethics, carries a heavy burden right now. This book is an excellent guide to the reasons why that burden is so pressing, and (in most cases) the book offers a really strong roadmap of where we are, and where we ought to go.

REFERENCES

Butler, Samuel (1863) 'Darwin among the machines', *The Press* (Christchurch, New Zealand) 13 June.

De Jaegher, H. and Di Paolo, E. (2007) 'Participatory sense-making. An enactive approach to social cognition', *Phenomenology and the Cognitive Sciences*, 6, pp.485-507.

Hendren, S. (2014) 'All technology is assistive,' *Wired*, 16 October, available at https://www.wired.com/2014/10/all-technology-is-assistive/ (accessed September 2020).

Robin L. Zebrowski
Departments of Philosophy, Psychology and Computer Science
Beloit College, Beloit WI, United States
zebrowsr@beloit.edu