Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence **by Patrick Lin, Keith Abney, and Ryan Jenkins (eds), Oxford University Press, 2017, ISBN: 9780190652951, 432 pp, £30.00.**

'Happy are those ages when the starry sky is the map of all possible paths—ages whose paths are illuminated by the light of the stars. Everything in such ages is new and yet familiar, full of adventure and yet their own.'[1]

'Any sufficiently advanced technology is indistinguishable from magic.'[2]

[Accepted 2 August 2018]

Reviewer: J Savirimuthu

Robots have moved from the niche enclaves of science fiction and research laboratories into everyday life. How we address the transformative and at times disruptive impact of innovations such as drones, driverless cars, assistive technologies and social robots rightly deserves close and careful scrutiny. As robotic technologies gradually become embedded into society, there has been a noticeable rise in debates about the ethics of emerging technologies, their potential and even concerns about the increasing mechanization of our lives and existential risks posed by artificial intelligence (AI). The collection of essays in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* aims to examine the ethical values embedded in a range of emerging technologies and identifies areas where particular care is warranted in respect of engineering and development decisions which may impair or undermine fundamental ethical norms in light of actual or perceived lack of legal standards and rules. *Robot Ethics 2.0* provides an appropriate resource for study and reflection on a range of philosophical issues that in a number of respects can be traced back to those encountered, for example, during the Enlightenment when the essence of human life and consciousness were debated and which continue to be sharply felt now as serious issues about sentience, consciousness and intelligence have had a renaissance in the age of Big Data and AI. At the core of these debates is the coming together of hardware and software and which forms the subject matter of the collection.

---

[1] Georg Lukacs, *The Theory of the Novel* (MIT Press 1971) 21

[2] Arthur C Clarke, 'Hazards of Prophecy: The Failure of the Imagination' in *Profiles of the Future: An Inquiry into the limits of the possible* (Gollancz 1962) 21

Before proceeding further, it may be appropriate to highlight two caveats to this book review. First, the review is undertaken from the perspective of a lawyer and not as an ethicist or philosopher. The task undertaken in this review is to reflect upon how the contributors view and approach the paradigms of complexity, trust and morality. Second, in view of the fact that governments and intergovernmental organisations alike are already embarking on wide ranging consultations on the very issues and topics covered in *Robot Ethics 2.0*, reflections are provided on what some of the conclusions might imply for expectations of the role of the institution of law in a novel environment. Even though scholars in the field of robotics and AI will bring their specialist knowledge and expertise to bear on the topics covered in *Robot Ethics 2.0*, how "Law" views these developments and can or ought to respond to the ethical imperatives will become a major feature in policymaking and governance initiatives in the future. With these two qualifications in mind, the key point made here is that the expert analysis in *Robot Ethics 2.0* rightly compel us to reassess our preconceptions about ongoing and emerging ethical dilemmas. The vignettes offered by the contributors have a broader dimension in so far as they also confront law with some fundamental doctrinal and constitutional questions about legitimacy, accountability and the rule of law. Specifically, from a legal and regulatory perspective, as standard setting norms and values gradually evolve to keep pace with innovations, the institution of law and lawyers needs to be cognizant of the fact that emerging technologies and AI are also giving rise to new claims, values, meanings and expectations over how information is constituted and repurposed.

Overview

Robot Ethics 2.0 builds on an earlier collected edition of essays on robot ethics published in 2014.[3] This new collection introduces an additional layer of knowledge and understanding to the ethical landscape of robotics in two ways. First, it makes an important contribution to the growing public and scholarly debate by contextualizing the social, ethical and normative choices at stake when robots and AI take centre stage in society. Second, *Robot Ethics 2.0* admirably frames the ethical dilemmas that must be grappled with at the research and product development phase before they are introduced to the general public. Taken together, the essays provide scholars and those new to this field with an accessible and well-researched account of major technological developments and innovations and should serve as a good resource for

[3] Patrick Lin, Keith Abney, Keith and George Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press 2012).

policymaking and furthering debates about robotics and AI in the age of modernity. The research undertaken by the contributors is impressive and complemented by an ability to integrate what might appear as abstract concepts and ethical dilemmas with concrete examples. The book is well structured with four main parts; it guides readers through the normative foundations of moral and legal responsibility, before exploring the relational dynamics of trust in human-robot relations and their specific applications from "Love to War", and concluding with some reflections on the implications of emerging technologies for humanity. The twenty-four chapters in the collection are also preceded with a helpful summary provided by the editors which frame the context and highlight some key questions to be addressed by the contributions. The only aspect that may come as a bit of a surprise with *Robot Ethics 2.0* is that Part I is principally focused on driverless cars. After all, problems of agency and the complications introduced by information to the machine question are not foreign to the philosophy of action and mind.[4] Be that as it may, the discussion, analysis and conclusions reached in Part I do not jar against the topics covered in the remainder of the collection. We can now turn to some key contributions and ideas emerging from the collection of essays.

Moral and Legal Responsibility

Attribution of responsibility has long been the staple of discussions on driverless cars. Part I lays the foundation for a careful examination of the contours for responsibility and provide thoughtful heuristics: the grammar of responsibility, the 'moral uncertainty' behind the grand vision of autonomous technology (Bhargava and Wan Kim), the unknown knowns of the 'ethics of ethics settings' (Millar), the artificiality of creating bright lines between human and machine agency (Loh and Loh) and the models for a precautionary principle in averting the possibility of an AI catastrophe (Gurney). The first four chapters provide a dynamic account of how responsibility could be conceptualised throughout the lifecycle of input and output data, emphasising some of the ethical tensions and ambivalence that lies beneath the choices and values inherent in any programing task or application of models of liability and responsibility to particular settings. It is particularly important to note that these highly refreshing and insightful treatments of responsibility, together with contributions by Zoller and White and Baum, acknowledge that any conception of autonomy is unlikely to veer far from Utilitarian

[4] Rodney Brooks, 'Intelligence Without Representation' (1991) 47 Artificial Intelligence 139-159; Joseph Weizenbaum, 'ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine' (1966) 9(1) Communications of the ACM 36-45.

or Kantian conceptions of responsibility, liability and the good life. It is probably a fair assessment that discussions about attributing liability and responsibility within the context of relations constituted by information seem to be less about divining the sentience of driverless technology and may ultimately resolve into policymakers and manufacturers confronting the complexity of hybrid interactions and determining how these could be managed through design or articulation of moral obligations of various legal actors. The unarticulated hope of many engaged in this domain of policymaking is that economic self-interest will help allay any lingering doubts about the provenance of some of the marketing hype relating to driverless technology.

Trust and Human-Robot Interactions

The relentless pace of innovation in robotics suggest that robots will come to play an important role in society and more than likely perform roles and tasks that have previously carried out by humans. As readers work through the chapters in Part II, they may also wish reflect on what it is that defines us as human beings or conversely, those features which distinguish us from machines. One specific policy question that Part II addresses well is the elucidation of the role and significance of trust in Human-Robot Interactions (HRI). As we near the end of the second decade of this century where encounters with programmed toys, digital assistants, sensor networks and assistive technologies are increasing and becoming commonplace, how trust is to be facilitated is a recurring theme in HRI research. The six chapters in Part II grapple not simply with the problems of trust and deception but also explore how interactions between humans and robots should be negotiated through the development of ethically sensitive designs and applications. This is an important area for research. Without trust, many concerns about an individual's vulnerability or exposure to violations of trust are likely to erect cultural and social barriers to the adoption of robots in diverse contexts and situations. Negotiating the boundaries between likability and trust is only one issue for HRI. The other involves overcoming the uncanny valley. The notoriety of the uncanny valley in HRI stems from feelings of uneasiness, anxiety or mistrust that emerge when robots resemble human feelings and characteristics. In addition to both these issues, there is the question of what principles should guide designers of intelligent robots in promoting the good life and the related task of operationalising these. The contributions do not shirk from addressing critical questions about methodology and the role and limits of 'user centric' approaches to HRI applications. Four chapters emphasise the importance of HRI to specific groups of individuals in society. Meacham and Studley, for

example, argue that the 'internal states of the agents' is not the only consideration. In care settings, they suggest that HRI should not discount the critical importance of giving priority to equally important considerations such as 'attentiveness, competence and responsiveness' (p. 99). Elder picks up this theme when examining the use of assistive technologies for autistic children. The coverage and analysis of social robotics literature and the paradox of friendships is helpful and particularly important to understanding the emotional and therapeutic benefits to be derived from the use of assistive technologies. There is much to be said about Elder's suggestion of drawing on Nussbaum's and Sen's theory of human capability and flourishing as a strategy for averting bias (p. 124). Borenstein, Howard and Wagner by contrast provide a sober reminder that excessive trust or reliance placed in robots may also give rise to moral hazards that may adversely affect individuals who may be unaware of the emotional bonds taking place. Pediatric healthcare is used as a context for examining some of the problems that may result from overtrust. There are some important insights generated which have implications for designers as well as those commissioning their use in settings where emotional attachments are frequently seen as having therapeutic benefits and correspond with the ideal of engaging with patients as human beings with real feelings and needs. The authors also make the point that we should not lose sight of the importance of identifying the types of relationships and expectations that may emerge from choices made about design, physical characteristics and functionality of human like machines. If monitoring trust reposed in robots is an important area of concern, their insights can also be used to reflect on how trust is to be operationalized in HRI and crucially, in areas where some of the risks associated with mistrust can lead to considerable harm for humans as in the context of medical surgery and hazardous environments such as nuclear energy or armed conflict. There is a broader point worth emphasising, namely that in addition to re-assessing the moral hazards for human-machine interpersonal relations, attention also needs to be given to the impact of mistrust on other human-human interactions. We get a useful insight on what level of trust should be placed on humans in highly structured and complex environments in the next chapter. Kirkpatrick, Hahn and Haufler's exploration of the boundaries between trust and reliance is particularly relevant as robot technologies become infused with artificial intelligence (AI) and machine learning capabilities which make assessing risk behaviour and adaptability less than straightforward (p. 149-151). Another problem associated with the role of humans in programming and designing robots is that we may unconsciously imbue machines with bias, values and attributes. The discriminatory or civil libertarian issues have been well documented. At a personal or subliminal level, the unintended consequences of the mimetic process and how these can or should be counteracted is less than

clear. There is no better example of the mimetic process of anthropomorphizing human like machines than the spontaneous outpouring of sympathy to a hitchhiking robot, appropriately named 'hitchBOT', which was vandalized. The final two chapters in this part brings us back to an important question relating to the contradictions whenever we anthropomorphize robots (Kate Darling) and counter-intuitively the need to recognise the value of designing robots to deceive for the greater good (Isaac and Bridewell). Both chapters clarify the range of cultural, ethical and legal considerations that need to be brought to the forefront in conversations about robotic functionalities, their filtering capabilities and how robots can be used to enhance the quality of the lives of individuals and society generally. It is equally important in this regard to not overlook issues of diversity and gender when designing social robots.

Applications: From Love to War.

It is apparent that the technological advances in social robotics, AI and autonomous systems are blurring the boundaries between humans and machines and one of the paradoxes is the unchallenged dominance of the ideology of permissionless innovation and the belief in the invisible hand in guiding development. The newness of the technology or the benefits of autonomy or automation does not invariably mean better or desirable. The topics covered in Part III could be viewed as a reminder of the assumptions that tend to be made when faced with new technologies and innovations. One should not underestimate the serious issues and topics discussed in this Part. Chapters 13 (Cheok, Karunanayaka, and Yann Zhang) and 14 (Bołtuć) consider HRI interactions in a sexual context. The use of empirical research and references to specific robotic applications provide some welcome and balanced analysis of the likely impact of social robotics on cultural norms and expectations as individuals and consumers. While we may not have too much ethical concerns about the use of Roomba helpers or Alexa in domestic settings, is there an ethical line that is crossed when robot sex brothels and voice recognition devices can be used for self-gratification or emotional engagement? What troubles some in society is the commercial dimension and the treatment of individuals, particularly women as objects of sexual gratification. The use of Snapchat and Instagram as spaces for exploring sexual identity and preferences may perhaps explain the anticipated growth of the robot sex technology industry which is reputed to be worth $30 billion. The race is already underway to build the world's first sex robot. If the quest to gain access to emotional or social resources to enhance our human capabilities could be seen as acceptable, albeit straining the boundaries of received social or cultural norms, we cannot exclude from our consideration the values and

power dynamics embedded in affective technologies. The problematizing of HRI in the emotional/sexual domain also draws attention to the susceptibility of individuals to be manipulated, particularly when data driven processes become the proxy for constituting and ordering relations, preferences and values often without the user's awareness (Henschke). Even if we are cognizant of the ethical challenges, these aspirations must be made manifest in the design and construction of robotic applications, a point which Klincewicz spends much time elaborating. This contribution repays careful reading as it raises some fundamental questions about the challenges faced in translating moral theories and counterfactuals into engineering solutions. The broader point that seems to emerge from each of these contributions in this Part may be that engineers and philosophers will need to better understand each other so that steps can be taken to find engineering solutions which correspond with human values and ethical norms. This is a legitimate goal for identifying and developing rules to a point. There is however the added dilemma of defining the ethical landscape for robotics since the normative structures (whether Utilitarian or Kantian) are fluid and creating a hierarchy of values not entirely free from their own questions regarding the hierarchy of rights to be prioritised. Where does one actually start when allocating to robots the range of universal rights for robots? How do we avoid the problem of over-or-under inclusion? Who decides and should the robot be given a discretion? Can we program robots to behave ethically or must there be a human in the loop? These are some the dilemmas discussed in the earlier text[5] and readers may find it helpful to consult this work alongside the present. The chapter by Talbot, Jenkins and Purves is particularly noteworthy as it contributes to the debate on how best to equip robots with ethical evaluation capabilities. One could, as the authors suggest that machines should be programmed with a set of ethical coordinates to guide the decision making so that they can act as consequentialists. Even though the authors make a compelling case for this analytical shift, there seems to be an acceptance that the policy choices to be made will depend on sector-specific Codes or principles such as those being prepared by the IEEE in forging a social consensus The final chapter returns us to subject that has been the public and media spotlight – autonomous weapons and the boundaries between legality and morality. Kahn is convinced that weaponising robots will not only lead to increasing armed conflicts but will also be morally objectionable. Kahn is right in his observation of the relative lack of political enthusiasm to withdraw from the lethal autonomous weapons (LAWS) arms race. The chapter provides a

[5] Lin and others (n 3).

useful account of how public pressure and increased visibility of the objectionable nature of LAWS may lead to self-imposed constraints on governments.

The Future of AI and Robotics.

The chapters in Part IV do not disappoint. LaBossiere proposes that artificial beings should be presumed to enjoy moral status (pp. 303-304). The term 'moral status' is used in the sense of an expectation that any harm or abuse without adequate justification would be regarded as wrong.[6] While humans are regarded as being worthy of moral consideration, claims that moral status should now be extended to artificial intelligence has its detractors. While discussions about the moral status of artificial intelligence may have once have been regarded simply as a philosophical debate, their practical relevance cannot be ignored. Intelligence in the technological domain now involves machines interacting with each other, sensors transmitting information and AI making decisions relayed in real-time in the Internet of Things (IoT). Th machine question manifests itself in calls for AI to be entitled to an ersatz moral status (pp. 303-304). DiGiovanna's chapter extends the rights discourse to enable artificial beings to be entitled to an identity (not in the sense of personal identity of humans). This idea of once barely imagined beings having artificial identities may seem to more appropriate for science fiction than a matter for the law. This may be short sighted. A deeper issue that emerges in the chapter is whether we may have to eventually have re-think our understanding of personhood in light of the gradual convergence of human enhancement and AI.[7] If the trajectory of technological development and epistemic uncertainties regarding personhood are unsettling, the emergence of superintelligent and unsympathetic AI is likely to provoke considerable anxiety regarding the existential risks it poses to humanity.[8] It may be prudent not to underestimate John Good's observation of what intelligence explosion implies for society and humanity:

> Thus the first ultra-intelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is

[6] See John Basl, 'Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines' (2014) 27 Philosophy & Technology 79-96; David Gunkel, 'A Vindication of the Rights of Machines' (2014) 27 Philosophy & Technology 113-132.
[7] Nick Bostrom, *Superintelligence: Paths, dangers, strategies* (OUP 2014).
[8] Stephen Hawking and others, 'Transcendence Looks at the Implications of Artificial Intelligence—But are we Taking AI Seriously Enough?' *The Independent* (1 May 2014).

> curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.[9]

Bostrom's work[10] provides the inspiration for Peterson's chapter. How do we avert the risks associated with super intelligent AIs? This is a challenging question as we simply have no precedents in history to assist us in dealing with runaway superintelligent machines. Petersen is not under any illusions of the complexity of the challenge that lies ahead but explores the possibility of a super ethical AI which may help offset existential threats. Vallor and Bekey's chapter maps the challenges machine learning applications pose for various sectors in the economy and society generally. An emerging concern raised by the chapter is how we should view AI and think about the knowledge produced by machines. (The General Data Protection Regulation[11] addresses this issue through new transparency and accountability rules to ensure checks and balances are provided when humans delegate responsibilities for decision making to machines, which may have significant legal consequences).

Abney's chapter provides an insight into the new information frontier for machines – space. This chapter provides a clear and concise account of the functions that robots can undertake in space. Indeed, many will agree with Abney that robots and AI do not have biological or physical constraints that enable them to better explore and navigate new or difficult terrains in space, undertake critical repairs and even set up bases in new colonies. The final chapter had the title of 'The Unabomber on Robots'; what makes this chapter intriguing is the fact that its author Galliott engaged in direct correspondence with Kaczynski, and an explanatory note (p. 383) addresses the ethical dilemma he faced in deciding to initiate this communication. His broad thesis can be framed in the form of a question that should also be central to any philosophy of technology: What are the emerging forms of marginalisation and disenfranchisement being engendered by our technoindustrial systems? There is a tendency to overly focus on the poster children of emerging technologies and perhaps underestimate the way power structures marginalise communities or influence and shape particular values and

[9] Irving John Good, 'Speculations Concerning the First Ultraintelligent Machine' (1966) 6 Advances in Computers 31–88.
[10] Bostrom (n 7).
[11] Regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

interests. Galliott's chapter reminds that all of us have a responsibility to ensure that we create a democratic and inclusive governance framework.

Conclusion

Robot Ethics 2.0 neatly captures the complexities of the design of emerging technologies and the ethical dilemmas created as a result of their integration into social environments and particular contexts. The role of law in these debates is particularly conspicuous by its perceived inability to keep pace with emerging technologies and concerns. This is only part of the narrative that has now been elevated to a viral status in light of driverless cars, social robots, and threats posed malicious AI. There are other explanations that should not be discounted. Two can be noted. First, there are considerable difficulties in ascribing responsibility through the formulation of clear and precise definitions of rights, duties and obligations in an environment where technological change seems to be relentless. Legislating involves the process of engaging with questions relating to the effectiveness, legitimacy or appropriateness of any legal or regulatory intervention, examination of options as well as identifying limits to extending existing category of rights, duties and harms. The second relates to the need to re-think the dialectical relationship between law and the empirical reality of information flows that involve hardware and software. The moral and legal responsibility associated with driverless cars is emblematic of the challenges in developing "an interpretive community" of rules, which enable the normative choices of the law of negligence, strict liability and product liability to be navigated.[12] Rules are unable to apply themselves and hence require legal actors order their activities and make their choices.[13] The problem in creating a community for rule interpretation is not limited to driverless technology but straddles other innovations where the intelligence explosion has enabled machines to assume human like characteristics and attributes. Our understanding of rules and decision making have long been based on normative frameworks and standards involving human-human interactions in society. How do we transpose or even extend these approaches to human like and non-biological entities like care robots or AI when it is well documented that rules in themselves can be ambiguous, permit discretion and can be indeterminate? The European Civil Law Rules on Robotics, well-intentioned and aspirational in its outlook, can also be seen as a timely reminder of the need to

[12] Julia Black, *Rules and Regulators* (OUP 1996) 214; Robert Baldwin, 'Why Rules Don't Work' (1991) 53 MLR 321.
[13] Black (n 12) 215-216.

reassess how rules function in the age of emerging technologies, reflecting on the strategies for compliance and contexts and the community in which rules are interpreted and applied. The quest to calibrate ethical principles with regulatory strategies legal rules may become one of the defining features of the interplay between the politics of science and the politics of technology. How philosophers position themselves within the sphere of policymaking is only part of the compelling case studies provided in *Robot Ethics 2.0*. It is beyond the scope of this review to address the other major concern - the power of corporate actors in driving innovation and how these are to be regulated. Notwithstanding the considerable opportunities now made possible by corporate actors with global influence, their challenges for policy development and rule making are real. First, permissionless innovation should never be presumed to be the default rule as it may lead to a culture of 'making rules up as we go along'. The fatality involving Uber's driverless car on 18th March 2018[14] raises a number of policy questions that have never been fully addressed in public: if driverless cars are meant to reduce accidents, are there cheaper and more efficient safety measures? Are the risks and burdens to be borne by pedestrians and other road users proportionate? Can driverless cars be justified in Utilitarian or Kantian terms? Second, one consequence of regulatory capture in the context of emerging technologies is that individuals will be treated as laboratory experiments, when sensors don't function as expected or if a robot is hacked or corrupted by malware. To what extent have the general public been involved in participating in decisions involving the transformation of urban spaces and critical infrastructures? However, like *Robot Ethics 2.0* it is important to end on a positive note and embrace the potential of emerging technologies such as robots and AI. Arthur C Clarke's famously regarded imagination as critical to enabling cultures to benefit from technology.[15] We can include within the long list of the output of human imagination, the horseless carriage, the Internet, and AI. While science fiction may be seen as a source for inspiration and magical thinking, the role and evolution of the law already provides a capacious resource for imagination and morality. If we are persistent, law could arguably be regarded as a technology for barely imagined beings and provides us with a compass of where to find them.

Joseph Savirimuthu

Senior lecturer, Liverpool Law School, University of Liverpool

[14] Daisuke Wakabayashi, 'Woman's Death in Arizona Casts A Pall on Driverless Car Testing' *New York Times* (20 March 2018) A1.
[15] Clarke (n 2) 12, 21.