Table of Contents

Editorial: Minds and Machines Special Edition

Ethics and Artificial Intelligence

Edited by Paula Boddington, Peter Millican, and Michael Wooldridge

Editor’s Introduction

We are delighted to introduce this collection of essays tackling various questions concerning ethics and artificial intelligence. Interest in this profoundly important area is gaining momentum, powered both by the increasingly impressive achievements in artificial intelligence, and also a corresponding growth in specific projects and initiatives asking hard questions about the ethical implications of this new technology. The initial impetus for this special edition indeed came from work on a project funded by the Future of Life Institute’s programme of research grants for Beneficial AI, and a workshop in Ethics and Artificial Intelligence that we organised at IJCAI16, held at New York in July, 2016.

Answers to ethical questions depend upon answers to questions about minds and agency, including questions about responsibility and intention, and about the place of humanity in the moral universe. Different normative moral theories take different approaches to such issues, and so in addressing practical ethical questions, we must inevitably to some extent presuppose positions regarding the nature of minds, and the nature of machines. Addressing ethical issues in AI requires that we ask questions also about how humans relate to AI, and what the implications are of replacing, supplementing or enhancing human thought, experience and action, with machines. So we need to ask questions about differences and similarities between human information processing and behaviour, and those of machines, and we need to drill down and ask some fundamental questions about the relationship between ethics and agency. What is autonomy, how is it understood, how does it differ between humans and (various) machines, what problems does machine autonomy create and how much should these be feared? The question of human control over machines is critical.

There are thus some very broad general questions that need to be asked, but also questions which are specific to certain contexts, such as in economics, and in law. For example in the case of law, we need to examine various specific branches of law, because these relate differently to questions of human agency, intention, and free will, and hence will raise quite different questions for adapting laws to machine agency. Yet even in the specific case, there will be general questions about mind and agency that need to be asked, or to which answers are being presumed, alongside the ethical questions. In examining specific contexts, we can also ask ourselves to what extent we can

extrapolate from specific contexts to learn general lessons about ethical questions in AI. There are so many layers of questions, and areas of uncertainty, that one tempting approach is to try to reduce this uncertainty; or to try to examine ways of proceeding despite uncertainty and despite areas of dispute. Some of the papers in this special edition attempt variations of such strategies.

In their paper, *Reframing AI Discourse*, Mario Verdicchio and Deborah Johnson address the question of how we understand AI and how we conduct debate concerning AI. They argue that there is a certain degree of responsibility on AI researchers for communications that affect the public understanding of the issues, and they emphasise the importance of the societal and human embeddedness of AI, its design and its deployment. They argue that the term 'autonomy' is liable to create misunderstandings and fear, and call for clarification. They argue that AI systems should be seen as parts of 'sociotechnical ensembles', these being broader systems that 'are combinations of artefacts, human behaviour, social arrangements and meaning'. This is not a mere issue of terminology, the authors explain, but an attempt to address a serious 'sociotechnical' blindspot affecting many aspects of our discussions about AI. Autonomy of systems is a matter of degree, varying with the level of human intervention required and the scope of operation. There are also levels of unpredictability, stemming from different kinds of ignorance, and this unpredictability is a major driver of public fear. Verdicchio and Johnson argue that public fear stems in large part from attributing to machines the same kind of autonomy that is attributed to human agents, with consequent worries about how to control such machines. The authors consider some projected future disaster scenarios for AI, expressing scepticism about the degree of concern that some have voiced. They suggest that those expressing concerns may suffer from a sociotechnical blindness concerning the human role that problematic futuristic scenarios require, for example by implicitly endowing a future 'superintelligence' with almost magical powers to act in the real world without human cooperation.

Their position opens up many further questions. Empirical research into public attitudes could delve more deeply into the source of any public disquiet concerning AI. Sociotechnical blindness could be one amongst a number of factors contributing to fears about AI, depending on the domain. For instance, there is currently widespread awareness of, and concern about, the vagaries and possible injustices of algorithm-led decision making, a concern which would seem to be well-founded, and not to rest upon any mistakenly enriched notion of machine autonomy. However this might be, the issue of how the public understands autonomous machines is clearly extremely important for advancing ethical debate.

Kyle Bogosian considers the question of how we might insert ethical behaviour in intelligent machines, given the range of moral uncertainty that exists. His paper, *Implementation of Moral Uncertainty in Intelligent Machines*, addresses uncertainties of the sort that cannot be resolved simply by addressing the questions of clarity in language and communication that Verdicchio and Johnson raise. His approach focuses upon how moral reasoning might be programmed into machines themselves, perhaps therefore bypassing the point that Verdicchio and Johnson made about examining whole sociotechnical ensembles of humans-plus-machines. Alternatively, one might understand Bogosian's approach as one which includes consideration of human input, given his close examination of methods for dealing with uncertainties arising from the variety in moral theories of the different human actors involved.

One major approach to addressing the control problem in AI is via the notion that machines themselves might be programmed to act ethically. But as Bogasian points out, this putative solution has to grapple with the large range of moral uncertainty that exists – both about which general moral theory is correct, but also regarding the correct response to specific moral problems. Indeed,

the layers of uncertainty are not limited just to these questions that Bogosian here asks. There are uncertainties about how moral theory might be applied to practice, how we might go about testing moral theories, whether there is such a thing as moral truth, what constitutes moral justification, and so on. Moral disagreement in itself seems an obvious problem, but Bogosian goes beyond this by spelling out different sources of the difficulty, arguing that compromise between theories might be hard, but then suggesting a possible way forward. He favours an approach championed by William MacAskill, who has argued for a complex voting procedure amongst the various moral theories and solutions that are considered by the moral agent (taking account, *inter alia*, of the agent's level of credence in each theory. Bogosian outlines this approach, and defends it against various objections.

There are questions to be asked about the background assumptions that tacitly support the kind of approach that MacAskill advocates (in great detail) and which Bogosian endorses. The general approach here seems to assume that there is a correct moral theory that we're all searching for, and that the different theories so far advocated each have some probability of being correct. This models itself rather closely on one particular account of the search for scientific truth, and will be hotly contested by many moral theorists. Bogosian attempts to defend MacAskill against the objection that the problem of credences will simply generate an infinite regress of disagreement, admitting that there will be some errors but that there is no alternative. However, one such alternative – of a familiar kind – is to ensure that there is always a human-in-the-loop. Bogosian asks if methods of deciding between rival moral theories are 'fair', but this in itself may presuppose a shared understanding of fairness, and at times, it would appear that in advocating a decision-making procedure for choosing between moral theories, we are no longer in the realm of the moral, but of the political, where the question is not so much how to decide what the best moral answer is, *per se*, but how to adjudicate between the claims of those advocating for different theories. One might also comment, that MacAskill's general mathematically minded methodology, whilst *prima facie* likely to appeal to those involved in AI, also lends itself more readily to those of a consequentialist bent, whose whole approach to moral questions is more open to pragmatic compromise than those of a more absolutist cast of mind.

In determining how to characterise, and how to address, the ethical questions that AI presents to us, we need to consider in great detail how the functioning and use of autonomous agents compares to the functioning of human agents. In their paper, *Ethical Issues for Autonomous Trading Agents*, Michael Wellman and Uday Rajan tackle the issue of the use of autonomous agents in a specific area, that of financial markets. They point out that pressing ethical and regulatory questions are already being raised in this area, and argue that focus on specifics may be useful not just for addressing the questions of the ethical behaviour of autonomous agents in the financial markets, but also for helping for generating broader conclusions concerning both the regulation of autonomous agents and more general control problems in AI. Financial markets present us with a complex and evolving area where there are already challenges in devising adequate regulations to address such issues as the boundaries between legitimate innovation and market manipulation. So this is a situation where there are already existing questions about how to determine the normative bounds of *human* behaviour – questions arising from a technically complex and global financial system, operating at electronic speed. This does indeed make autonomous trading agents a useful case for consideration, since ethical questions concerning artificial intelligence often occur against the background of situations where incremental and deeply embedded technological changes have been occurring for some time. In the case of financial markets the regulatory and ethical challenges of AI are introduced into a tightly regulated normative landscape which is itself relatively novel and evolving as the increase in computing power has significantly shifted the realities of financial trading.

There are already difficulties in financial regulation concerning the boundaries between intentional and non-intentional outcomes. How are we to regulate AI when its very autonomy means that its behaviour is hard to predict, and it is also hard to know whether this behaviour should be classed as intentional or not – indeed, what would even count as intentional? Answering such questions involves not only deep understanding of AI, but of the basis for attributing intention to humans and how this impacts upon the normative assessment of actions.

Moreover, notwithstanding the authors' suggestion that this case could provide a model for considering AI more generally, in drawing such conclusions we would have to consider the underlying ethical framework. The need for regulation of financial markets is immense, for a number of obvious reasons; but consideration of how this should be done also raises more fundamental questions about the basis of the economic system, including economic modelling, forecasting, and the calculation of economic value. Wellman and Rajan carefully explain the notion of arbitrage and the regulatory and ethical issues that it raises for autonomous trading agents, within a framework in which the existing basic economic and financial world is presumed. Questioning the framework itself may be beyond the scope of these discussions. The operation of artificial intelligence however, may sometimes act to push a system to such stresses that fundamental questions about its basic underlying values are inevitably exposed. One reason for this may be the inbuilt brakes to the operation of a system provided by human biology and the limits on the speed at which we can think and act, compared with the huge power and speed of modern computers, especially when operating with vast amounts of data and in multi-agent systems which can be replicated and operated at massive scale. Most of our laws and regulations have been designed for human agency and intent, operating within biologically mediated epistemic and speed constraints. And it is a human motivation – loss aversion – which provides a strong incentive for addressing these concerns, given the possibly large financial penalties of poorly operating autonomous trading agents.

Another incentive for addressing the regulation of autonomous systems is the concern that failure to do so might hamper their development. Ugo Pagallo addresses the issue of regulation of robotics in his article, *When Morals Ain't Enough: Robots, Ethics, and the Rules of the Law*. This paper nicely complements the other papers in this volume, for Pagallo addresses questions that remain even if there could be agreement on what is right and wrong, what is good and bad, about a robot's behaviour. In order to consider normative questions about *machine* behaviour in the legal sphere, it seems appropriate first to look carefully at how we ascribe legal responsibility to *humans*, which varies in different fields of law. In contract and in tort law, the already existing notion of strict liability could be extended to machines relatively straightforwardly, but in criminal law, there are interesting and pressing questions concerning responsibility and guilt that remain open with regard to the behaviour of machines. And whatever our moral views, in law, criminal liability can only be attributed on the basis of explicit legal norms. Pagallo's paper, then, provides an interesting examination of how differences between the humans and machines, in respect of their behaviour and its causes, might mould our normative responses in the field of legal regulations. Interestingly, Pagallo notes how thinking about whether or not machines might be endowed with free will raise questions which might well shape our whole current legal system, not simply those aspects concerning future possible AI. Again, we see how the challenge of AI is pressing us to address extremely fundamental questions about mind and agency, and their relationship to ethics.

Pagallo goes on to examine the nature of the risk of unpredictability of robots and the threat this poses to traditional notions such as reasonable foreseeability. He suggests that secondary legal rules could be brought in to manage the development of legal responses to the unpredictability challenge, examining in particular the Japanese government's introduction of special zones for

robotics empirical testing and development. Discussing the contrasting limits of consequentialism and of deontology with regard to intentions, Pagallo suggests that a virtue ethics approach might prove especially suitable for understanding the pragmatic approach of legal testing zones for robotics. Virtue ethics is often hailed as way of navigating areas of technological change, where there is a need to retain reference to the importance of human motivation and agency, yet at the same time, a need to amend existing rules and regulations, as an alternative to the agent-blindness of consequentialism, and the perceived rule-bounded approach of deontology. The key factor here seems to be the flexibility to change that Pagallo attributes to virtue ethics. Virtue ethics has its origin in the work of Aristotle, who developed his ethical thinking at a time lacking our rapid technological progress, and in a society very different from ours. It is a moot question then, how suited virtue ethics is to our current situation of technological and societal uncertainties, although it is an approach that many favour and are currently testing. It remains to be seen, moreover, how legal norms might in fact be developed within robotic testing zones.

We are delighted that these four papers together present a variety of approaches to the pressing ethical and legal questions in artificial intelligence, and recommend them as together providing a valuable contribution to this unfolding and vital debate.