Machine Ethics: the design and governance of ethical AI and autonomous systems

Alan Winfield, Katina Michael, Jeremy Pitt and Vanessa Evers

1. Introduction

The so-called 4th industrial revolution and its economic and societal implications is no longer solely an academic concern, but has become a matter for political as well as public debate. Characterised as the convergence of robotics, AI, autonomous systems and information technology – or cyber-physical systems – the fourth industrial revolution was the focus of the World Economic Forum, at Davos, in 2016 [1]. Also in 2016 the US White House initiated a series of public workshops on artificial intelligence (AI) and the creation of an interagency working group, and the European Parliament committee for legal affairs published a draft report with recommendations to the Commission, on Civil Law Rules on Robotics.

Since 2016 there has been a proliferation of ethical principles in AI; seven sets of principles were published in 2017 alone – including the influential Future of Life Institute's Asilomar principles for beneficial AI. Nor has there been a shortage of industry initiatives, perhaps the most notable is the Partnership in AI; founded in 2016 by representatives of Apple, Amazon, DeepMind and Google, Facebook, IBM, and Microsoft, the partnership now has more than 50 members. Most significant however has been the number of national AI initiatives announced [2]. Since Canada published its national AI strategy early in 2017 more than 20 other countries (and groupings including the EU) have announced similar initiatives. The largest and most comprehensive is undoubtedly China's Next Generation AI development plan, launched in July 2017.

Notably all of these initiatives express the need for serious consideration of the ethical and societal implications of robotics and artificial intelligence. Robot and AI ethics has been transformed from a niche area of concern of a few engineers, philosophers and law academics, to an international debate. For these reasons we believe this special issue – focussed on the ethics of intelligent autonomous systems – is not only timely but necessary.

The primary focus of this special issue is machine ethics, that is the question of how autonomous systems can be imbued with ethical values. Ethical autonomous systems are needed because, inevitably, near future systems are moral agents; consider driverless cars, or medical diagnosis AIs, both of which will need to make choices with ethical consequences. This special issue includes papers that describe both implicit ethical agents, that is machines designed to avoid unethical outcomes, and explicit ethical agents: machines which either encode or learn ethics and determine actions based on those ethics. Of course ethical machines are socio-technical systems thus, as a secondary focus, this issue includes papers that explore the educational, societal and regulatory implications of machine ethics, including the question of ethical governance. Ethical governance is needed in order to develop standards and processes that

allow us to transparently and robustly assure the safety of ethical autonomous systems and hence build public trust and confidence.

2. The landscape of robot and AI ethics

The field of robot and AI ethics is broadly divided into two main branches. By far the largest of these is concerned with the vitally important question of how human developers, manufacturers and operators should behave in order to minimize the ethical harms that can arise from robots and AIs in society, either because of poor (unethical) design, inappropriate application, or misuse. This branch is concerned with the ethical application of robots and AIs and is generally referred to as either AI ethics or robot ethics, and has already led to the development of ethical principles [3,4], standards [5], and proposals for good practice [6,7].

The smaller branch of AI ethics is concerned with the question of how robots and AIs can themselves behave ethically. Referred to as ethical AI, ethical robots or – more generally – *machine ethics*, the field spans both philosophy and engineering. Philosophers are concerned with questions such as "could a machine be ethical, and if so which ethics should determine its behaviour?" alongside larger questions such as "should society delegate moral responsibility to its machines?" while engineers are interested in solving the (significant) technical problem of how to build an ethical machine. The two disciplines are not disconnected; philosophers are interested in the outcomes of practical machine ethics not least because – if successful – they lend urgency to the moral questions around ethical machines in society. Equally, engineers engaged in designing ethical machines need philosophers to advise on the definition of appropriate ethical rules and values for these machines.

Of course the idea that robots should not only be safe but also actively capable of preventing humans from coming to harm has a long history in science fiction. In his short story *Runaround*, Asimov [8] expressed such a principle in his now well-known Laws of Robotics. Although no-one has seriously proposed that real-world robots should be 'three-laws safe', work in machine ethics has advanced the proposition that future robots should be more than just safe. In their influential book *Moral Machines – teaching robots right from wrong*, Wallach and Allen [9] set out the philosophical foundations of machine ethics, coining the term Artificial Moral Agent (AMA); Wallach and Allen write:

> "If multipurpose machines are to be trusted, operating untethered from their designers or owners and programmed to respond flexibly in real or virtual world environments, there must be confidence that their behaviour satisfies appropriate norms. This goes beyond traditional product safety ... if an autonomous system is to minimise harm, *it must also be 'cognisant' of possible harmful consequences of its actions, and it must select its actions in the light of this 'knowledge'*, even if such terms are only metaphorically applied to machines" (italics added).

Since Čapek's 1920 play *Rossum's Universal Robots*, and the stories of Asimov, Frayn [10] and many others, science fiction narratives have played a key role in the exploration of artificial morality [11]. In the study of machine ethics we can now, for the first time, begin to investigate artificial morality in fact.

3. The state of the art in Machine Ethics

Machine ethics is a new field. Although the antecedents of machine ethics can be found in computer ethics [12], the earliest works on machine ethics were published less than 20 years ago. Those works are, not surprisingly, theoretical and philosophical. The field of machine ethics was *de facto* established by Allen *et al* [13,14], Asaro [15], Moor [16], Powers [17], Anderson and Anderson [18,19] and Wallach and Allen [9].

The earliest proposal for a practical *ethical governor* for robots – a mechanism for moderating or inhibiting a robot's behavior to prevent it from acting unethically – was from Arkin [20][1], although not tested on real robots. At the time of writing the number of experimental demonstrations of ethical robots remains very small indeed; to the best of our knowledge there have been only five such demonstrations to date: (i) the GenEth system of Anderson and Anderson [22], (ii) the Asimovian ethical robots of Winfield *et al* [23] and Vanderelst and Winfield [24], (iii) Bringsjord *et al*'s Akratic robot [25], (iv) the "sorry I can't do that" robot of Briggs and Scheutz [26] and (v) the Intervening robot mediator in healthcare of Shim, Arkin and Pettinatti [27]. Papers which review and update the approaches of (i) and (ii) are included in this issue; we now briefly review (iii - v).

Based on earlier work proposing a general 'logistic' method for engineering ethically correct robots [28], Bringsjord's Akratic[2] robot is tested in scenarios in which it is charged with guarding a prisoner of war and must choose between retaliating with violence to an attack (thus satisfying a self-defence goal) or refraining from retaliation [25]. The robot's 'ethical substrate model' incorporates logic that specifies which actions are forbidden under certain circumstances, or which are permitted or obligatory; importantly that logic, a deontic cognitive event calculus (DCEC), can be formally verified.

Briggs and Scheutz [26] demonstrate a robot capable of declining to carry out an instruction, if it reasons that to do so would harm itself. The robot may more properly be described as implicitly ethical since it is designed only to avoid unethical outcomes to itself, but the cognitive machinery it uses to avoid such outcomes is, like Bringsjord et al's robot [25], based on reasoning about obligation and permissibility.

---

[1] The notion of a governor in autonomous systems dates back to the birth of cybernetics (from *Kyvernitis* – to govern), see [21].
[2] From the Greek *akrasia* referring to when a person acts in contradiction to their better judgement.

In the work of Shim *et al* [27] a robot equipped with an ethical governor monitors the interaction between a patient and healthcare advisor. The robot is able to sense (i) the loudness of the patient's speech, (ii) when the patient stands up to leave, and (iii) when the advisor asks the patient to take some medicine. In four test scenarios: 'yelling', 'quiet', 'stay-in-the-room' and 'safety-first', the robot senses the patient advisor interaction and uses both speech and gesture to intervene. In the safety-first scenario the robot might alert the advisor with "the previous records say he had a reaction (to this medicine). I think it's not safe".

Allen *et al* [14] identified three approaches to machine ethics: top-down, bottom-up and a hybrid of top-down and bottom-up. The top-down approach requires training the machine to be able to recognise and correctly respond to morally challenging situations. The bottom up approach instead constrains the actions of the machine in accordance with pre-defined rules or norms. Of the five experimental demonstrations listed here only the first: Anderson and Anderson's GenEth system, adopts a top-down approach. The other four – although very different in detail – are all examples of bottom-up constraint-based approaches.

All five trials are of robots with very limited ethics in constrained laboratory settings; they each demonstrate proof of concept but are far from practical application in real-world settings. Nevertheless we can be confident that it is possible, at least in principle, to build minimally ethical robots. Later in this article we shall consider the question of the degree of ethical agency of these robots.

All of the work above has, not surprisingly, focussed on the core processes – the cognitive architectures – for making ethical choices. But many technical challenges remain: how, for instance, do we design a robot or AI that is "cognisant of possible harmful consequences of its actions"? This requires that a machine can reliably recognise when and how its actions have ethical salience; for a robot interacting with a human that means that it needs to perceive that the human is at risk – and the nature of that risk. The accurate perception of both risk and context is non trivial even in controlled environments – and even more challenging in real-world settings. Context might also change the risk calculation; an elderly person who is frail is clearly at greater risk that one who is physically fit, and the presence of other humans (a nurse for instance) clearly also changes the context for a care robot.

If and when ethical machines are ready for real–world application we would need to be sure that their ethical decision-making processes are both guaranteed and safeguarded against misuse. We would also need robust frameworks for ethical governance, including technical standards for ethical machines alongside processes of accident investigation. And bigger societal questions would need to be addressed around the extent to which we are prepared to delegate moral responsibility to our machines.

4. Defining machine ethics

Although not providing a singular definition of machine ethics Moor's influential 2006 paper: *The nature, importance, and difficulty of machine ethics* [16], defines the field by articulating four categories of ethical agency. These are:

- Ethical impact agents: any machine that can be evaluated for its ethical consequences.
- Implicit ethical agents: machines designed to avoid unethical outcomes.
- Explicit ethical agents: machines that can reason about ethics.
- Full ethical agents: machines that can make explicit moral judgments and justify them.

Anderson and Anderson [18] suggest that the goal of machine ethics "is to create a machine that is guided by an acceptable ethical principle or set of principles in the decisions it makes about possible courses of action it could take". In *Machine Ethics* [19] Anderson and Anderson elaborate upon this definition: "machine ethics is concerned with giving machines ethical principles, or a procedure for discovering a way to resolve they ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making".

In this special issue we follow Anderson and Anderson's 2006 definition. To paraphrase: *an ethical machine is guided by an ethical rule, or set of rules, in deciding how to act in a given situation*. It follows that we are concerned here with *autonomous* machines: either software AIs, or their physically embodied counterpart, robots, which determine how to respond to input without direct human control. Although labelled as autonomous such systems are generally subject to human supervision or monitoring (and intervention if necessary) in a way that is more properly described as 'supervised autonomy'; but what is important is that low-level decisions are made by the system rather than a supervising human. In this special issue we are concerned with ethical systems that fall into both of Moor's second and third categories of implicit and explicit ethical agents.

All of the five demonstrations of ethical robots referenced in the previous section are certainly implicit and, arguably, explicit ethical agents, since they all have either learned or defined ethical rules and the cognitive machinery to take those rules into account when deciding how to act in a given situation and – if necessary – proactively acting or intervening to prevent harm; a process we can reasonably describe as ethical reasoning (albeit of a limited kind). It is important to note that Moor's definitions of ethical agency, and in particular the distinction between implicit and explicit agency, remain controversial [29], as does the question of which systems properly qualify as explicitly ethical machines [30].

We would argue that *all* non-trivial examples of real-world AIs and robots are ethical impact agents as defined by Moor. The ethical impact of an online store's AI in offering suggestions for purchases may be slight, but there can be no doubt

that search engines and AIs that determine which social media posts and advertisements appear on your home page can have significant societal and political impact during, for instance, election campaigns [31].

Consider AIs that recommend loans, determine welfare payments, or recommend prison sentences. We only need to reflect on the impact of an incorrect decision to see that these are systems with ethical impact. Indeed bias in AIs trained with uncurated data sets is now a well-known problem [32]. Even chatbots, which may appear inconsequential, can fall victim to gaming by users and quickly become deeply offensive [33, 34]. Connected AI-toys designed to interact and converse with children are even more worrying. More obvious examples of systems with clear ethical impact include medical diagnosis AIs, assisted-living (care) or companion robots, and driverless cars. All of these systems have the potential for negative ethical impact, either as a result of poor design or simply a lack of forethought about how the system might be used [35]. And few systems, regardless of how well designed, are immune to malicious use [36] It follows that all AI and robotics systems would benefit from ethical risk assessment [5] within the wider framework of Responsible Innovation [37].

We would go further and propose that *all* robots and AIs of the kinds mentioned here should be designed to avoid negative ethical impacts; in other words they should be designed to be implicit ethical agents within Moor's schema. Frameworks for the design and test of implicit ethical agents are now emerging; the IEEE Standards Association document *Ethically Aligned Design* [6] and associated standards currently in draft such as IEEE P7000 *Model Process for Addressing Ethical Concerns During System Design,* together provide a methodology for imbuing ethical values into intelligent systems. Those values are then expressed implicitly in such a way that the system meets the highest standards of, for instance, transparency and explainability (P7001), privacy (P7002) and is – as far as possible – free of bias (P7003). Adamson *et al* outline these and other IEEE ethics initiatives in this special issue [47].

5. How ethical are current explicitly ethical machines?

One objection frequently levelled at the ethical machines demonstrated thus far is that they simply extend the envelope of safe behavior, and that they should not be described as ethical, but simply as 'safety plus' machines.

To counter this objection let us consider a simple thought experiment: you are walking on the street and notice a child who is not looking where she's going (perhaps engrossed in her smart phone); you see that she is in imminent danger of walking into a large hole in the pavement. Suppose you act to prevent her falling into the hole. Most bystanders would regard your action as that of a good person – in more extreme circumstances you might be lauded a hero. Of course your action *does* keep the child safe, but it is also a moral act. Your behavior is consequentialist ethics in action because you have (a) anticipated the consequences of her inattention and (b) acted to prevent a calamity. Your act is ethical because it is an expression of care for another human's safety and well-being. To claim that ethical robots merely exhibit safety plus behavior is to miss

the key point that it is the *intention* behind an act that makes it ethical. Of course, robots and AIs don't have intentions – even explicitly ethical robots – but their designers do, and an ethical robot is an instantiation of those good intentions.

Moral philosophers are also, and perhaps not surprisingly, doubtful over claims that any machines can be described as ethical. Some argue that morality is the exclusive preserve of humans. Of course humans are not the only animals that demonstrate altruism, but we are almost certainly the only species capable of both consciously reflecting upon and justifying the morality of our actions.

One moral philosopher [38] offered the following opinion on the minimally ethical robots described in [23]:

> "The obvious point that any moral philosopher is going to make is that you are assuming that an essentially consequentialist approach to ethics is the correct one. My personal view, and I would guess the view of most moral philosophers, is that any plausible moral theory is going to have to pay at least some attention to the consequences of an action in assessing its rightness, even if it doesn't claim that consequences are all that matter, or that rightness is entirely instantiated in consequences. So on the assumption that consequences have at least some significance in our moral deliberations, you can claim that your robot is capable of attending to one kind of moral consideration, even if you don't make the much stronger claim that it is capable of choosing the right action all things considered."

It does therefore appear to be reasonable to make the limited claim that the ethical robots demonstrated in the laboratory are at least "capable of attending to one kind of moral consideration".

We need to make the distinction between what might in principle be achievable in the near future and the far future. For machines to be as good as humans at moral reasoning they would need, to use Moor's terminology, to be full ethical agents [16]. The best we have demonstrated to date is a handful of proof-of-concept explicitly ethical agents and even those, as we have commented, are only minimally ethical. In no sense are such agents better at moral reasoning than humans.

Does the fact that we have arguably reached the third category in Moor's scheme (explicit ethical agents) mean that full ethical agents are on the horizon? The answer again must be 'no'. The scale of Moor's scheme is not linear. It is a relatively small step from ethical impact agents to implicit ethical agents, then a very much bigger and more difficult step to explicitly ethical agents, which we are only just beginning to witness. But then there is a huge gulf to full ethical agents, since they would almost certainly need something approaching human equivalent AI; full explicit ethical agents, by Moor's definition [16], require consciousness, intentionality, and free will. Indeed, we think it is appropriate to think of explicitly ethical machines as examples of narrow AI; full ethical agents would almost certainly require artificial general intelligence (AGI).

The Ethical Turing test

Several have proposed a test for ethical machines somewhat akin to the Turing Test. Allen *et al* [13] for instance outline a 'comparative Moral Turing Test' (cMTT) in which a human interrogator is presented with pairs of examples of morally significant actions of a human and an ethical machine. The interrogator is asked to judge which of each pair of actions is less moral, and if the less moral actions are, more often, human decisions, then the machine is judged to pass the test. As Allen *et al* point out, there are several problems with this test; one is that human behavior is often less than perfectly ethical, so the cMTT may be setting the moral threshold too low – they note that "decisions that result in harm to others are likely to be much less tolerated in a machine than in another human being".

Anderson and Anderson [22] propose an alternative 'Ethical Turing Test' (ETT) in which a panel of ethicists are presented with the machine's ethical decisions across a range of application domains. Each ethicist is asked whether they agree or disagree with those decisions – if a significant number are in agreement (i.e. the ethicist would have made the same choice in the same situation) – then the machine is judged to pass the test. In a run of this test with a panel of 5 ethicists, described in [22], the level of agreement ranges from 75% to 100%. Anderson and Anderson note that the most contested domain of 'treatment reconsideration' – "should the health care worker (or robot) accept the patient's decision regarding treatment as final, or attempt to change their mind?" – was also the most ethically sensitive for humans. A finding that supports our view that ethical decisions difficult for humans will be equally challenging for machines and their designers.

It is important to note that the idea of an ethical Turing Test remains controversial. Arnold and Scheutz [39] for instance argue against such a test. One of their objections is, essentially, that the ethical Turing Test treats the ethical agent as a black box by considering only its decisions and not the reasoning behind those decisions. Instead Arnold and Scheutz advocate verification, which seeks "predictable, transparent, and justifiable decision-making and action"– a position with which we strongly agree.

The Trolley problem and learned ethical principles

Consider the distinction between a machine that takes decisions according to predetermined principles of ethics and one that learns those principles of ethics from observed decisions. In fact, the latter is closer to the 'scientific' study of (rather than the application of) ethics: the objective is to identify principles which serve to explain judgements of morality (distinctions between good and bad, or right and wrong) in terms of the reasoning used to justify them. This, rather than training, is the real motivation behind studying abstract situations such as the trolley problem [40]. These problems provide data (in the form of a set of moral judgements) from which a general principle of ethics can be derived. As with any good theory, this principle should have predictive leverage: so, if the

problem or its parameters are changed, it should be possible to test whether the principle still holds (or not).

It is one thing to propose principles of ethics, implement them in a machine, and use them in a restricted context (the same context from which the principle was derived, perhaps); although then there is the difficult question of whether it is appropriate to use such machines for decision-support in other contexts. It is another thing to apply Machine Learning algorithms to 'learn' the principle(s) from the data; in which case, we need to be absolutely certain that the dataset has not been biased and that the explanatory principle so learned really does have predictive leverage when applied to a different context. For sure, if the machine learns a 'wrong' or 'inadequate' principle, or even just a 'simple' principle, then there will be problems if we try to apply it in other situations.

Intuitively, the situation appears to be analogous to the problem of distributive justice, where in any given situation there are a number of possible ways of distributing rewards or punishments. These are called legitimate claims [41]. Equally, in any situation requiring a moral judgement, there may well be a number of principles that have stronger or weaker relative relevance. The requirement is to work out which principles apply in any situation, how to accommodate them in case of plurality, and how to reconcile them in case of conflict [42].

6. The morality of building ethical machines

Consider the question: is it ethical to delegate moral responsibility to machines? We routinely delegate task-level responsibilities to our machines; from simple automata such as washing machines to advanced safety-critical systems such as airplane autopilots, we trust a wide range of machines to undertake tasks that used to be exclusively performed by humans. To extend this kind of delegated responsibility to encompass ethics may therefore appear uncontroversial, but – we contend – it is a step that should only be taken with great care.

In considering the morality of building ethical machines we need to differentiate between implicitly ethical machines – those designed to avoid unethical outcomes – and explicitly ethical machines – those that reason about ethics, because the calculus of risk is quite different in each case.

Consider first the category of implicitly ethical machines. We have already argued that *all* intelligent autonomous systems that have the potential to cause harm should be classed as implicitly ethical machines, and designed using processes of ethically aligned design. The risks of not doing so are already apparent as outlined in section 4 above. Building implicitly ethical machines would seem to be an ethical course of action.

Conversely, explicitly ethical machines do bring risk. In this special issue Cave *et al* identify four broad categories of risk: "a) the risk that ethically aligned machines could fail, or be turned into unethical ones; b) the risk that ethically aligned machines might marginalize alternative value systems; c) the risk of

creating artificial moral patients; and d) the risk that our use of moral machines will diminish our own human moral agency" [51]. It is by no means clear that we should build ethical machines, even if we can.

Of course real-world explicitly ethical machines – those that reason about ethics – are not inevitable. There are, to the best of our knowledge, no explicitly ethical agents in real-world use. As we have outlined here the only explicit ethical agents that exist are a handful of proof-of-concept laboratory prototypes. These are minimally ethical machines with limited functionality designed only to test hypotheses about how to build such machines. None are examples of real-world robots – such as autonomous cars with added ethics functionality.

The ubiquitous trolley problem

In the public discourse around the ethics of AI, there is a common assumption that ethical machines – and in particular driverless car autopilots – should be able to resolve ethical dilemmas and choose between two equally undesirable outcomes. This assumption is fuelled by the Trolley Problem [41], of which there are many variations. For driverless cars the problem is often posed as the car having to choose whether to kill one human or several. A recent high profile study asked respondents for their preferences in a range of driverless car scenarios and, perhaps not surprisingly, the study revealed significant cultural variations [43].

From an engineering perspective building a machine that can both reliably perceive and then choose between two unethical outcomes in any real-world environment – let alone in fast moving dynamic situations – is far beyond the state of the art. Yet in the public discourse this is rarely made clear, giving rise to unrealistic expectations of the decision-making capabilities of near future driverless cars. The trolley problem is a thought experiment; it undoubtedly has value both as a philosophical tool (as outlined in section 5 above) and at a statistical level – as suggested by Bonnefin *et al* in this special issue [49] – but should not influence designs for driverless cars or any other autonomous systems.

Even if the technical problem of machines able to resolve real-world ethical dilemmas were solved society-wide debate would then be needed to discuss and agree on the rules and protocols for such machines, not least because society as a whole needs to take responsibility for the human causalities of accidents caused by such machines. It is notable that the federal government of Germany has ruled that "In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another" [44].

7. How should we govern ethical machines?

To recap: *all* intelligent autonomous systems should be regarded as ethical impact agents. Of course some may have no ethical impact whatsoever, but it would be foolhardy to assume this. Ideally all systems should be first, subjected

to an ethical risk assessment, of the kind set out in standard BS 8611 *Guide to the ethical design of robots and robotics systems* [5], and second, redesigned to reduce the impact of any ethical risks exposed by that risk assessment. Those systems that are shown, through ethical risk assessment, to have some ethical impact (and we would wager that this would be almost all systems), which are then redesigned to avoid or minimise that impact, move into the category of implicit ethical machines.

It follows that we regard ethical risk assessment as the cornerstone of ethical governance. However, ethical risk assessment on its own is not enough, especially for implicit ethical machines. One of the general ethical principles set out in Ethically Aligned Design [6] concerns transparency; the principle asserts that it should always be possible to find out how and why an autonomous system made a particular decision. This kind of transparency is not a property of autonomous systems by default. It needs to be designed in, alongside sub-systems for securely logging system inputs, outputs and decisions – the robot/AI equivalent of an aircraft flight data recorder [45]. Without transparency discovering the causes of, for instance, a driverless car accident, or misdiagnosis by a medical diagnosis AI, becomes all but impossible. That process of discovery is vital if the faults that caused the accident are to be fixed, and accountability established [46].

Consider now the governance of explicitly ethical machines. We already have very high expectations for the safety and reliability of our machines – especially those that have the potential to cause serious harm if they go wrong – but if and when real-world robots and AIs are explicitly ethical those expectations will be even higher. Because of this additional burden of expectation on explicitly ethical machines, their governance will need to be especially robust.

Explicitly ethical machines require two governance considerations over and above those needed for implicitly ethical machines. The first is in the choice of ethical rules or – if those rules are learned – the choice of training cases. Those ethical rules or training cases need to be carefully scrutinised, ideally by a panel of users, ethicists, lawyers and representatives of civil society. The second concerns the transparency of the ethical decision making process – it is critical that each ethical decision should be logged for later analysis and that it should be possible to understand why the system made those decisions. This is not only in the event that a system fails or causes harm. We believe that explicitly ethical machines should be subject to a 'probationary period', during which *all* ethical decisions are reviewed (perhaps by the same panel that scrutinises the system's ethical rules). Only after this period is successfully concluded would monitoring revert to the baseline of fully investigating actual or near miss accidents. Given that explicitly ethical machines are neither inevitable nor imminent, these considerations do not have the same urgency as those suggested for implicit ethical machines.

It is clear that *all* ethical machines need to be developed and used responsibly [37], and – for those that are ethically-critical – subject to standards and strong regulatory frameworks (many of which do not yet exist). In particular

transparency should extend beyond the ethical machines themselves, to encompass the processes of design and operation, within frameworks of ethical governance [7]. Without such frameworks it is hard to see how ethical machines will be trusted.

8. The papers of this special issue

Adamson *et al* contribute a paper titled: "Designing a Values-Driven Future for Ethical Autonomous and Intelligent Systems" [47] in which they argue that human values must drive our future autonomous systems in a way that both protects and benefits humanity. The paper describes IEEE's work in this area and includes a table of approximately 50 activities within the IEEE related to ethics.

Anderson *et al* contribute a paper entitled "A Value-Driven Eldercare Robot: Virtual and Physical Instantiations of a Case-Supported Principle-Based Behavior Paradigm" [48]. The paper describes both simulated and real-robot implementations of an eldercare robot in which ethical principles are learned, via inductive logical programming, from a set of training examples provided by a project ethicist using GenEth: a general ethical dilemma analyzer.

Bonnefon *et al* provide an insightful point of view article titled: "The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars" [49]. This paper, which brings attention to what the authors call 'the statistical trolley dilemma', is the only paper of the special issue focussed on autonomous vehicles

Bremner *et al* contribute a paper titled: "On Proactive, Transparent and Verifiable Ethical Reasoning for Robots" [50] in which they review and update an approach to the design of ethical robots based on a simulation-based internal model. This model allows the robot to anticipate when another robot – acting as a proxy human – might be at risk of harm and intervene if necessary; the ethical robot's reasoning is both transparent and verifiable.

Cave *et al*, present a paper aptly titled: "Motivations and Risks of Machine Ethics" [51]. In this paper the authors clarify various philosophical issues surrounding the concept of an ethical machine and the aims of machine ethics. The authors argue that while there are good *prima facie* reasons for pursuing machine ethics, there are also potential risks that must be considered and managed.

Ema *et al*, an 11 person research team based in Japan, present the paper "Clarifying Privacy, Property, and Power: Case Study on Value Conflict Between Communities" [52]. Based around a controversial case study on the 'flaming' of fan fiction, which was complicated by the ambiguous legal position of such fan fiction content in Japan, the paper aims to clarify notions of privacy and draw lessons for the ethical governance of (ethical) AI in the presence of value conflicts, through interdisciplinary collaboration.

Robertson *et al* contribute a paper titled: "Engineering based ethical design methodology for embedding ethics in autonomous robots" [53]. The paper

explores the process of robotics and autonomous systems development using a co-design approach to reduce end-user risk with respect to an endoscopic capsule for diagnosis and drug delivery. The contribution of the paper is a method for embedding ethics into the design of a machine in a socio-technical application.

"Understanding Engineers' Drivers and Impediments for Ethical System Development: The Case of Privacy and Security Engineering" is a paper contributed by Spiekermann *et al* [54]. The study surveys 124 engineers in order to understand the drivers and impediments facing ethical systems development with respect to privacy and security engineering. Their findings indicate that while many engineers regard security and privacy as important, they (a) do not enjoy working on them, and (b) struggle with their organizational environment.

In their point-of-view paper "Toward the Agile and Comprehensive International Governance of AI and Robotics" [55], Wallach and Marchant propose an agile ethical/legal model for the international and national governance of AI and robotics, building on their recommendations on the formation of Governance Coordinating Committees (GCCs) for coordinated oversight of emerging technologies.

References

[1] Schwab K. (2017) The fourth industrial revolution. Portfolio Penguin.

[2] Dutton T (2018) An Overview of National AI Strategies, Medium, July 2018 https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

[3] Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., et al. (2017) Principles of robotics: Regulating robots in the real world. Connection Science, 29 (2). pp. 124-129.

[4] Boddington P. (2017) Towards a code of ethics for artificial intelligence. Cham, Switzerland: Springer.

[5] BS8611:2016 (2016) Robots and robotic devices: guide to the ethical design and application of robots and robotic systems. British Standards Institute, London.

[6] IEEE Standards Association (2017) Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems (A/IS), version 2. See https://ethicsinaction.ieee.org/

[7] Winfield AF and Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems, Phil. Trans. Royal Society A, 376, 20180085.

[8] Asimov I. *I, Robot*. Gnome Press, 1950.

[9] Wallach W and Allen C. (2009) Moral Machines: Teaching Robots Right from Wrong. Oxford: Oxford University Press.

[10] Frayn M (1965), The Tin Men, Collins.

[11] Cave S and Dihal K (2019), Hopes and fears for intelligent machines in fiction and reality, Nature Machine Intelligence, 1: 74-78.

[12] Forester T and Morrison P (1994) Computer ethics: cautionary tales and ethical dilemmas in computing, MIT Press.

[13] Allen C, Varner G and Zinser J (2000), Prolegomena to any future artificial moral agent, JETAI 12, 251-261.

[14] Allen, C., Smit, I. and Wallach, W (2005) Artificial morality: Top-down, bottom-up, and hybrid approaches, Ethics and Information Technology 7, 149-155.

[15] Asaro PM (2006) What should we want from a Robot Ethic?, International Review of Information Ethics, 6 (12): 9-16.

[16] Moor JH. (2006) The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems 21:18–21.

[17] Powers TM (2006), Prospects for a Kantian Machine, in IEEE Intelligent Systems, vol. 21, no. 4, pp. 46-51, July-Aug. 2006. doi: 10.1109/MIS.2006.77

[18] Anderson M, and Anderson SL, eds. (2006). Special Issue on Machine Ethics. IEEE Intelligent Systems 21(4) (July/August).

[19] Anderson, M. and Anderson, S.L., eds., (2011) Machine Ethics, Cambridge University Press.

[20] Arkin RC (2009) Governing Lethal Behavior in Autonomous Systems, Routledge.

[21] Adamson G, Kline RR, Michael K and Michael MG (2015), Wiener's Cybernetics Legacy and the Growing Need for the Interdisciplinary Approach, in Proc. IEEE, vol. 103, no. 11, pp. 2208-2214.

[22] Anderson, M. and Anderson, S.L. (2014): GenEth: A General Ethical Dilemma Analyzer, in Proc. 28th AAAI Conference on Artificial Intelligence, 253-261.

[23] Winfield AF, Blum C, and Liu W (2014) Towards an ethical robot: internal models, consequences and ethical action selection. In Lecture Notes in Computer Science, 8717, pp. 85–96. Berlin, Germany: Springer.

[24] Vanderelst, D. and Winfield, A. F. (2018) An architecture for ethical robots inspired by the simulation theory of cognition. Cognitive Systems Research, 48. pp. 56-66.

[25] Bringsjord S, Sundar N, Thero D, and Si M. (2014) Akratic robots and the computational logic thereof. In Proc. IEEE 2014 Int. Symposium on Ethics in Engineering, Science, and Technology, pages 7:1–7:8, Piscataway, NJ, USA.

[26] Briggs G and Scheutz M (2015) "Sorry, I can't do that": Developing mechanisms to appropriately reject directives in Human-Robot Interactions, in Proc AAAI Fall Symposium Series.

[27] Shim J, Arkin RC and Pettinatti M (2017) An Intervening Ethical Governor for a robot mediator in patient-caregiver relationship: Implementation and Evaluation. 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, pp 2936-2942.

[28] Bringsjord S, Arkoudas K, and Bello P. (2006) Toward a general logicist methodology for engineering ethically correct robots. IEEE Intelligent Systems, 21(4):38–44.

[29] Dyrkolbotn S, Pedersen T and Slavkovik M (2018) On the distinction between implicit and explicit ethical agency, in Proc AAAI/ACM conf. on AI, Ethics and Society.

[30] Sharkey A. (2017) Can we program or train robots to be good?, Ethics Inf Tech. doi: 10.1007/s10676-017-9425-5

[31] Helbing D. et al. (2019) Will Democracy Survive Big Data and Artificial Intelligence? In: Helbing D. (eds) Towards Digital Enlightenment. Springer.

[32] Caliskan, A., Bryson, JJ. Narayanan, A. (2017) Semantics derived automatically from language corpora contain human-like biases, Science Vol 356, Issue 1334, pp183-186.

[33] Neff, G. and Nagy, P. (2016) Symbiotic Agency and the Case of Tay. Int. J. of Communication, vol 10, p. 17.

[34] Wolf MJ, Miller KW, and Grodzinsky FS. (2017). Why we should have seen that coming: comments on Microsoft's tay "experiment," and wider implications. ACM SIGCAS Computers and Society 47(3):54-64. doi: 10.1145/3144592.3144598

[35] Charisi V, Habibovic A, Andersson J, Li J, and Evers V (2017) Children's Views on Identification and Intention Communication of Self-driving Vehicles, In Proc. 2017 ACM Conf. on Interaction Design and Children, pp 399-404, ACM.

[36] Brundage, M., Avin, S., Clark, J. Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A. et al (2018) The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, arXiv preprint arXiv:1802.07228

[37] Jirotka M, Grimpe B, Stahl B, Eden G, and Hartswood M. (2017) Responsible research and innovation in the digital age. Commun. ACM 60, 62–68.

[38] Reilly-Cooper R (2015) Commentary on Towards an ethical robot: internal models, consequences and ethical action selection. Personal communication.

[39] Arnold, T. and Scheutz, M. (2016) Against the moral Turing test: accountable design and the moral reasoning of autonomous systems, Ethics Inf. Technol. 18: 103. https://doi.org/10.1007/s10676-016-9389-x

[40] Thomson JJ (1985), The Trolley Problem, Yale Law Journal, Vol 94, p1395.

[41] Rescher N (1966). Distributive Justice. Bobbs-Merrill.

[42] Pitt J, Busquets D, and Macbeth S (2014) Distributive Justice for Self-Organised Common-Pool Resource Management, TAAS 9(3): 14:1-14:39, 2014.

[43] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon J-F and Rahwan I (2018) The Moral Machine experiment, Nature, vol 563, pp59–64

[44] BMVI (2017) Ethics Commission: Automated and Connected Driving, German Federal Ministry of Transport, June 2017.

[45] Winfield AF and Jirotka M (2017) The case for an ethical black box. Lecture Notes in Computer Science, 10454, pp 262–273, Springer.

[46] Wachter S, Mittelstadt B and Floridi L (2017), Transparent, explainable, and accountable AI for robotics, Science Robotics 2(6).

[47] Adamson G, Havens JC and Chatila R (2019) Designing a Value-Driven Future for Ethical Autonomous and Intelligent Systems," Proceedings of the IEEE, doi: 10.1109/JPROC.2018.2884923

[48] Anderson M, Anderson SL and Berenz V (2019), A Value-Driven Eldercare Robot: Virtual and Physical Instantiations of a Case-Supported Principle-Based Behavior Paradigm," Proceedings of the IEEE, doi: 10.1109/JPROC.2018.2840045

[49] Bonnefon J-F, Shariff A and Rahwan I (2019) The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars, Proc. IEEE.

[50] Bremner P, Dennis LA, Fisher M and Winfield AF (2019), On Proactive, Transparent and Verifiable Ethical Reasoning for Robots, Proc IEEE.

[51] Cave S, Nyrup R, Vold K and Weller A (2019), Motivations and Risks of Machine Ethics, Proceedings of the IEEE, doi: 10.1109/JPROC.2018.2865996

[52] Ema A et al. (2019), Clarifying Privacy, Property, and Power: Case Study on Value Conflict Between Communities, in Proceedings of the IEEE, doi: 10.1109/JPROC.2018.2837045

[53] Robertson LJ, Abbas R, Alici G, Munoz A and Michael K (2019), Engineering-Based Design Methodology for Embedding Ethics in Autonomous Robots, in Proceedings of the IEEE. doi: 10.1109/JPROC.2018.2889678

[54] Spiekermann S, Korunovska J and Langheinrich M (2019), Inside the Organization: Why Privacy and Security Engineering Is a Challenge for Engineers, in Proceedings of the IEEE. doi: 10.1109/JPROC.2018.2866769

[55] Wallach W and Marchant G (2019) Toward the Agile and Comprehensive International Governance of AI and Robotics, Proc IEEE.