Ethical Governance is essential to building Trust in Robotics and AI Systems

Alan FT Winfield[1] and Marina Jirotka[2]
[1]Bristol Robotics Laboratory, UWE Bristol, UK
[2]Department of Computer Science, University of Oxford, UK

August 24, 2018

Abstract

This paper explores the question of ethical governance for robotics and AI systems. We outline a roadmap – which links a number of elements including ethics, standards, regulation, responsible research and innovation and public engagement – as a framework to guide ethical governance in robotics and AI. We argue that ethical governance is essential to building public trust in robotics and AI, and conclude by proposing five pillars of good ethical governance.

1 Introduction

The aim of this paper is to present a case for a more inclusive, transparent and agile form of governance for Robotics and Artificial Intelligence in order to build and maintain public trust and to ensure that such systems are developed for the public benefit. Building public trust in intelligent autonomous systems (IAS) is essential. Without that trust the economic and societal benefits of IAS will not be realized. In this paper we will lay out a roadmap. The value of a roadmap is twofold; firstly it connects and maps the different elements that each contribute to IAS ethics, and secondly it provides us with a framework to guide ethical governance.

Over the last decade the private sector has made significant investments in the development of robots and AI with autonomous capacities that can interact with humans in order to fulfill roles in work, home, leisure, healthcare, social care and education. These developments potentially offer huge societal benefits. They can save time, reduce human effort to perform tasks and reduce costs. They can also improve well-being through the provision of reliable care assistance for the ageing population, standardisation in service encounters, companionship and affective aids for different user groups, and relieve humans from both dangerous and menial tasks.

Public attitudes toward these new intelligent technologies are generally positive [20, 21], However, concerns have been raised regarding the irresponsible use

and potentially harmful impact of IAS. These concerns are frequently raised in public rhetoric which tends to position robots' future ubiquity as inevitable and construct dystopian scenarios featuring the usurpation of human autonomy, safety and authority. At the same time there are genuine – and possibly well founded – fears around the impact on jobs and mass unemployment [12]. We know there is no 'formula' for building trust, but we also know from experience that technology is, in general, trusted if it brings benefits and is safe and well regulated.

Building such trust in robotics and AI will require a multiplicity of approaches, from those at the level of individual systems and application domains [44] to those at an institutional level [29, 38]. This paper argues that one key (necessary but not sufficient) element in building trust in intelligent autonomous systems is ethical governance. We define ethical governance as a set of processes, procedures, cultures and values designed to ensure the highest standards of behaviour. Ethical governance thus goes beyond simply good (i.e. effective) governance, in that it inculcates ethical behaviours in both individual designers and the organisations in which they work. Normative ethical governance is seen as an important pillar of responsible research and innovation (RI), which "entails an approach, rather than a mechanism, so it seeks to deal with ethical issues as or before they arise in a principled manner rather than waiting until a problem surfaces and dealing with it in an ad hoc way" [42].

Given the increasing pace of innovation [22] new and agile processes of governance are needed. A recent World Economic Forum white paper has suggested that the rapid pace of transformative technological innovation is "reshaping industries, blurring geographical boundaries and challenging existing regulatory frameworks" [10]. Increasingly not only policy makers, but also businesses and innovators feel a duty to engage with policies to address the societal consequences of their innovation; the report calls for a more inclusive and agile form of governance. The WEF, as the international organisation for Public-Private Cooperation, is launching a global initiative on Agile Governance dedicated to reimagining policymaking for the Fourth Industrial Revolution [45]. The Forum defines agile governance as "adaptive, human-centred, inclusive and sustainable policymaking, which acknowledges that policy development is no longer limited to governments but rather is an increasingly multi-stakeholder effort" [54]. It is those non-governmental stakeholders, including individual researchers, research institutions and funders, professional bodies, industry and civil society, who are key to making ethical governance both agile and practical. In practice this means incorporating different kinds of knowledge, including that from citizens, to inform the goals and trajectories of innovation.

This paper is concerned with the ethical governance of both physical robots such as driverless cars or personal assistant robots (for care or in the workplace), and software AIs such as medical diagnosis AIs or personal digital assistants. All of these are intelligent agents with some degree of autonomy, so we refer to them collectively as intelligent autonomous systems (IAS). The last 18 months have seen a proliferation of new ethical principles for robots and AI (especially AI). But principles are not practice and, while it is heartening to witness a

growing awareness of the need for ethics, there is little evidence of good practice in ethical governance. Given that transparency is a core principle of ethical governance one has to be skeptical of any claims that organizations make unless they, for instance, publish the terms of reference and membership of ethical boards, alongside evidence of good ethical practice. The gap between principles and practice is an important theme of this paper and the five pillars of good ethical governance that we propose in this paper are aimed that addressing this gap.

The paper is structured as follows: in section 2 we build a roadmap to show how the components of ethical governance including ethical principles, responsible innovation, standards and regulation are connected. Section 3 then provides a commentary on the roadmap, considering public fears, standards and regulation, safety-critical AI, transparency and moral machines. A brief concluding discussion – including a set of five recommendations for ethical governance – is given in section 4.

2 Building the Roadmap

The core of our roadmap connects research on ethics with emerging standards and regulation. Standards often formalise ethical principles into a structure which could be used either to evaluate the level of compliance or, more usefully perhaps for ethical standards, to provide guidelines for designers on how to reduce the likelihood of ethical harms arising from their product or service. Ethical principles may therefore underpin standards either explicitly or implicitly. Consider safety standards such as ISO13482 [33] – here the underpinning ethical principle is that personal care robots must be safe. In ISO13482 that principle is explicit but in many standards it is not. Process standards such as the ISO 9000 family of quality management standards could, for instance, be said to express the principle that shared best practice benefits all. But standards also sometimes need teeth, i.e. regulation which mandates that systems are certified as compliant with standards, or parts of standards. Most standards are voluntary. There is no requirement to adopt IEEE 802.11 (WiFi), for instance, in a new networked product, but not to do so would clearly be commercially unwise. And those standards that are mandated – often because they relate to safety – are *de facto* directed because a license to operate a system would not be granted until after that system has been shown to be compliant with those standards. Furthermore *soft governance* plays an important role in the adoption of standards: by requiring compliance with standards as a condition of awarding procurement contracts governments can and do influence and direct the adoption of standards – across an entire supply chain – without explicit regulation. Accepting that this is a simplification of a process with many intervening factors, we argue that ethics (or ethical principles) lead to standards, which in turn lead to regulation as shown in Figure 1 and that this characterisation has value in understanding the landscape of ethical governance.

Figure 1 references some foundational ethical frameworks, including the 2006

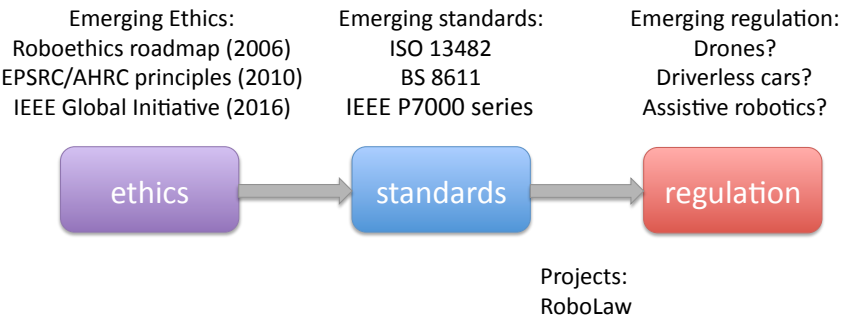

Figure 1. Linking Ethics, Standards and Regulation

Euron Roboethics Roadmap [51] and the EPSRC Principles of Robotics [7]. An informal survey at the end of 2017 [60] discovered that a total of 10 different sets of ethical principles (including Asimov's laws of robotics[1]) had been proposed by December 2017, seven of which appeared in 2017. These are listed in table 1.

There is a good deal of commonality across these principles, notably that IAS should (1) do no harm, including being free of bias and deception, (2) respect human rights and freedoms, including dignity and privacy, while promoting well-being, and (3) be transparent and dependable while ensuring that the locus of responsibility and accountability remains with their human designers or operators. Perhaps the most important observation relates not to the content of these principles but the increasing frequency of their publication: clear evidence for a growing awareness of the urgent need for ethical principles for IAS. But principles are not practice. They are an important and necessary foundation for ethical governance, but only the first step.

[1] Notable because Asimov was undoubtedly the first to establish the principle that robots (and by extension AIs) should be governed by formally stated principles.

Principles	# principles	Year and Refs
Asimov's Laws of Robotics	3	1950
Murphy and Wood's three laws of Responsible Robotics	3	2009 [39]
The EPSRC Principles of Robotics	5	2011 [7]
Future of Life Institute's Asilomar principles for beneficial AI	23	Jan 2017
ACM US Public Policy Council's Principles for Algorithmic Transparency and Accountability	7	Jan 2017
Japanese Society for Artificial Intelligence (JSAI) Ethical Guidelines	9	Feb 2017
Draft principles of The Future Society's Science, Law and Society Initiative	6	Oct 2017
Montreal Declaration for Responsible AI draft principles	7	Nov 2017
IEEE General Principles of Ethical Autonomous and Intelligent Systems	5	Dec 2017 [32]
UNI Global Union Top 10 Principles for Ethical AI	10	Dec 2017

Table 1: Principles of robotics and AI published by Dec 2017.

In Fig. 1 we reference recent standards such as ISO 13482 [33] (Safety requirements for personal care robots) and – perhaps the world's first ethical standard for robotics – BS 8611:2016 [14]. Whereas ISO 13482 is concerned with personal care robots, the scope of BS 8611 extends to all classes and domains of robots and robotic systems.

Ethics and Standards both fit within a wider overarching framework of Responsible Research and Innovation (RI). RI initiatives across policy, academia and legislation emerged over a decade ago and began with an aim to identify and address uncertainties and risks associated with novel areas of science. This has recently expanded to consider Computer Science, Robotics, Informatics and ICT more generally. RI proposes a new process for research and innovation governance [46]. The aim is to ensure that science and innovation are undertaken in the public interest by incorporating methods for encouraging more democratic decision-making through greater inclusion of wider stakeholder communities that might be directly affected by the introduction of novel technologies.

Responsible Innovation both informs and underpins ethics and standards, as shown in Figure 2. Importantly ethical governance is a key pillar of RI. RI also connects directly with ethics through, for instance, public engagement, open science and inclusivity; notably open science has been described as a 'trust technology' [26]. Another key component of RI is the ability to systematically and transparently measure and compare system capabilities, typically with standardised tests or benchmarks [19].

A further key element of RI, especially when systems move into real world

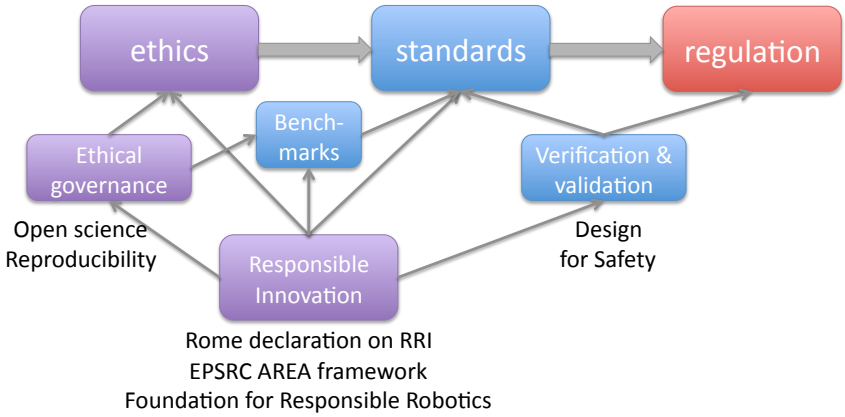

Figure 2. Scaffolded by Responsible Research and Innovation

application is the need for verification and validation, to provide assurance both of safety and fitness for purpose. Verification and validation might be undertaken against published standards, and – for safety critical systems – conformance with those standards may be a legal requirement without which the system would not be certified. Hence verification and validation links to both standards and regulation. Figure 2 references underpinning frameworks for responsible innovation including the 2014 Rome Declaration on Responsible Research and Innovation [43], the EPSRC Anticipate, Reflect, Engage and Act (AREA) framework [24, 46] and the recently established Foundation for Responsible Robotics [62]. Furthermore the AREA framework has been tailored specifically for ICT [27, 34].

In general technology is trusted if it brings benefits while also safe, well regulated and, when accidents happen, subject to robust investigation. One of the reasons we trust airliners, for example, is that we know they are part of a highly regulated industry with an outstanding safety record. The reason commercial aircraft are so safe is not just good design, it is also the tough safety certification processes and, when things do go wrong, robust and publicly visible processes of air accident investigation. It is reasonable to suggest that some robot types, driverless cars for instance, should be regulated through a body similar to the Civil Aviation Authority (CAA), with a driverless car equivalent of the Air Accident Investigation Branch. It is important to note that air accident investigations are social processes of reconstruction that need to be perceived as impartial and robust, and which serve as a form of closure so that aviation does not acquire an enduring taint in the public consciousness. We anticipate very similar roles for investigations into robot accidents [59].

Regulation requires regulatory bodies, linked with public engagement [56, 55] to provide transparency and confidence in the robustness of regulatory processes.

All of which supports the process of building public trust, as shown in Figure 3.

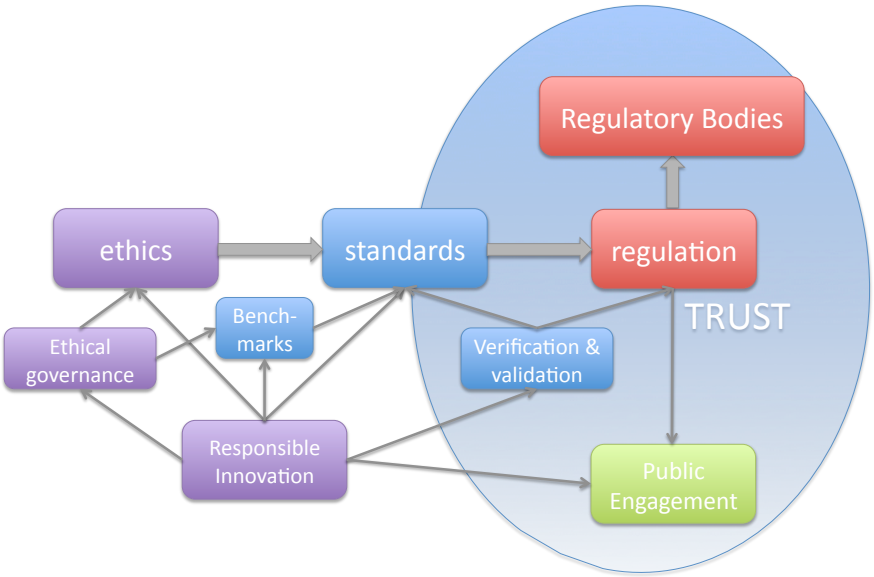

Figure 3. Building Public Trust

3 Commentary on the IAS Ethics Roadmap

In the following sections we provide a deeper commentary on a number of aspects of the roadmap touched on above, including public fears, standards and regulation. We also deepen and extend the present and future context of the roadmap with an introduction to safety-critical AI, the need for transparency, and – looking to the future – the governance issues of moral machines (systems that explictly reasons about ethics).

3.1 Public fears

It is well understood that there are public fears around robotics and artificial intelligence. Many of these fears are undoubtedly misplaced, fuelled perhaps by press and media hype, but some are grounded in genuine worries over how the technology might impact, for instance, jobs or privacy.

The most recent Eurobarometer survey on autonomous systems showed that the proportion of respondents with an overall positive attitude has declined from 70% in the 2012 survey [20] to 64% in 2014 [21]. Notably the 2014 survey showed that the more personal experience people have with robots, the more favourably

they tend to think of them; 82% of respondents have a positive view of robots if they have experience with them, whereas only 60% of respondents have a positive view if they lack robot experience. Also important is that a significant majority (89%) believe that autonomous systems are a form of technology that requires careful management.

A recent survey of decision making in driverless cars reveals distinctly ambivalent attitudes [9] "... participants approved of utilitarian Autonomous Vehicles (AVs) (that is, AVs that sacrifice their passengers for the greater good) and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. The study participants disapprove of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV."

It is clear that public trust in robotics cannot simply be assumed [6, 16, 28]; to do so could risk the kind of public rejection of a new technology seen (in Europe) with Genetically Modified foods in the 1990s [35]. Proactive actions to build public trust are needed including, for example, the creation of a 'machine intelligence commission' as argued by [38]; such a commission would lead public debates, identify risks and make recommendations to parliament, for new regulation or regulatory bodies, for instance, and recommend independent mechanisms for responsible disclosure.

3.2 Standards and Regulation

Work by the British Standards Institute technical subcommittee on Robots and Robotic Devices led to publication – in April 2016 – of BS 8611: *Guide to the ethical design and application of robots and robotic systems* [14]. BS8611 incorporates the EPSRC principles of robotics [7]; it is not a code of practice, but instead gives "guidance on the identification of potential ethical harms and provides guidelines on safe design, protective measures and information for the design and application of robots". BS8611 articulates a broad range of ethical hazards and their mitigation, including societal, application, commercial/financial and environment risks, and provides designers with guidance on how to assess then reduce the risks associated with these ethical hazards. The societal hazards include, for example, loss of trust, deception, privacy and confidentiality, addiction, and employment.

The primary output from the IEEE Standards Association's global ethics initiative [30] is a discussion document called *Ethically Aligned Design* (EAD), now in its second iteration [32]. The work of 13 committees, EAD covers: general (ethical) principles; how to embed values into autonomous intelligent systems; methods to guide ethical design and design; safety and beneficence of artificial general intelligence and artificial superintelligence; personal data and individual access control; reframing autonomous weapons systems; economics and humanitarian issues; law; affective computing; classical ethics in AI; policy; mixed-reality, and well-being. EAD articulates a set of over 100 ethical issues and recommendations. Each committee was asked to recommend issues that should be addressed through a new standard. At the time of writing, 11 stan-

dards working groups are drafting candidate standards to address an ethical concern articulated by one or more of the 13 committees outlined in EAD: the so called IEEE P7000 'human' standards. To give one example IEEE P7001: *Transparency in Autonomous Systems* is drafting a set of measurable, testable levels of transparency for each of several stakeholder groups including users, certification agencies and accident investigators [13].

Significant recent work on surveying the state of robotics regulation was undertaken by the EU project RoboLaw. The primary output of that project is a comprehensive report entitled *Guidelines on Regulating Robotics* [41]. That report reviews both ethical and legal aspects; the legal analysis covers rights, liability and insurance, privacy and legal capacity. The report focuses on driverless cars, surgical robots, robot prostheses and care robots and concludes by stating: "The field of robotics is too broad, and the range of legislative domains affected by robotics too wide, to be able to say that robotics by and large can be accommodated within existing legal frameworks or rather require a *lex robotica*. For some types of applications and some regulatory domains, it might be useful to consider creating new, fine-grained rules that are specifically tailored to the robotics at issue, while for types of robotics, and for many regulatory fields, robotics can likely be regulated well by smart adaptation of existing laws".

3.3 Safety Critical Artificial Intelligence

Initial efforts toward ethics and regulation was focussed on robotics; it is only more recently that attention has turned towards AI ethics. Because robots are physical artifacts they are undoubtedly more readily defined and hence regulated than distributed or cloud-based AIs. This and the already pervasive applications of AI (in search engines, machine translation systems or intelligent personal assistant AIs, for example) strongly suggest that greater urgency needs to be directed toward considering the societal and ethical impact of AI, including the governance and regulation of AI.

A reasonable definition of a modern robot is 'an embodied AI' [57], thus in considering the safety of robots we must also concern ourselves with the AI controlling the robot. The three types of robot indicated in figure 1: drones, driverless cars and assistive robots[2] will all be controlled by an embedded AI[3], of some appropriate degree of sophistication. Yet these are all *safety-critical* systems, the safety of which is fundamentally dependent on those embedded AIs; decisions made by these embedded AIs have real consequences to human safety or well being (in that a failure could cause serious harm or injury). Let us consider two general issues with AI: (1) trust and transparency and (2) verification and validation, both of which come into sharp focus for our three exemplar robot categories of drones, driverless cars and assisted-living robots.

AI systems raise serious questions over trust and transparency:

[2]including care or workplace assistant robots. For a robot taxonomy refer to [57, pp.37-41]
[3]Perhaps augmented by a cloud based AI

- How can we trust the decisions made by an IAS and, more generally, how can the public have confidence in the use of AI systems in decision making?
- If an IAS makes a decision that turns out to be disastrously wrong, how do we investigate the logic by which the decision was made, and who is responsible (noting that the AI cannot itself be responsible)?

Existing safety critical systems are not AI systems, nor do they incorporate AI systems. The reason is that AI systems (and in particular machine learning systems) are largely regarded as impossible to verify for safety critical applications – the reasons for this need to be understood.

- First is the problem of verification of systems that learn. Current verification approaches typically assume that the system being verified will never change its behaviour, but a system that learns does – by definition – change its behaviour, so any verification is likely to be rendered invalid after the system has learned.
- Second is the *black box* problem. Modern AI systems, and especially the ones receiving the greatest attention, so called Deep Learning systems, are based on Artificial Neural Networks (ANNs). A characteristic of ANNs is that after the ANN has been trained with data sets[4], any attempt to examine the internal structure of the ANN in order to understand why and how the ANN makes a particular decision is more or less impossible. The decision making process of an ANN is opaque.

The problem of verification and validation of systems that learn may not be intractable, but is the subject of current research, see for example work on verification and validation of autonomous systems [53]. The black box problem may be intractable for ANNs, but could be avoided by using algorithmic approaches to AI (i.e. that do not use ANNs). Notably a recent report has recommended that "core public agencies ... should no longer use 'black box' AI and algorithmic systems" [15].

3.4 Transparency

One aspect of ethical governance discussed above is 'transparency'. Transparency is an essential property of ethical governance; it would be hard to argue that opaque governance is ethical. Ethical governance in robotics and AI should ideally demonstrate both transparency of *process* and transparency of *product*; the former refers to the transparency of the human processes of Research and Innovation, the latter to the transparency of the robot or AI systems so developed.

Consider now product transparency. This will necessarily mean different things to different stakeholders – the kinds and levels of transparency required

[4] which may be very large, so called 'big data' sets – which themselves pose another problem for verification

by a safety certification agency or an accident investigator will clearly need to be different to those required by the system's user or operator. Ideally systems should be explainable, or even capable of explaining their own actions (to non experts) as well as transparent (to experts). There is a growing literature on transparency, see for instance work on transparency and explainability in robot systems [49, 59, 61], transparency in relation to the EU General Data Protection Regulation (GDPR) [23, 52], and on the limitations of transparency [3].

An important underlying principle is that it should always be possible to find out why an autonomous system made a particular decision (especially if that decision has, or might, cause harm). Given that real-world trials of driverless car autopilots have already resulted in several fatal accidents [47, 48] there is clearly an urgent need for transparency in order to discover how and why those accidents occurred, remedy any technical or operational faults, and establish accountability. A new IEEE standard P7001 *Transparency in Autonomous Systems* is currently under development, which will define measurable, testable levels of transparency and "provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency" [31].

A technology that would provide such transparency, especially to accident investigators, would be the equivalent of an aircraft flight data recorder (FDR). We call this an ethical black box (EBB) [59], both because aircraft FDRs are commonly referred to as black boxes[5], and because such a device would be an integral and essential physical component supporting the ethical governance of intelligent autonomous systems. Like its aviation counterpart the EBB would continuously record sensor and relevant internal status data so as to greatly facilitate (although not guarantee) the discovery of why a robot or AI made a particular decision or series of decisions – especially those leading up to an accident. EBBs would need to be designed – and certified – according to standard industry-wide specifications, although it is most likely that each application domain would have a different standard; one for driverless vehicles, another for drones and so on.

3.5 Towards Moral Machines

This paper is primarily concerned with robot and AI ethics, rather than ethical robots. But inevitably near future autonomous systems, most notably driverless cars are *by default* moral agents. It is clear that both driverless cars and assistive (i.e. care) robots make decisions with ethical consequences, even if those robots have not been designed to explicitly embed ethical values and moderate their choices according to those values. Arguably all autonomous systems implicitly reflect the values of their designers or, even more worryingly, training data sets (as dramatically shown in AI systems that demonstrate human biases [17]).

Moor [37] makes the useful distinction between *implicit* ethical agents, that is machines designed to avoid unethical outcomes, and *explicit* ethical agents, that

[5]Note that the use of the colloquialism 'black box' in aviation should not be confused with the 'black box problem' in AI, referred to in section 3(c).

is machines which either directly encode or learn ethics and determine actions based on those ethics. There is a growing consensus that near future robots will, as a minimum, need to be designed to reflect the ethical and cultural norms of their users and societies [14, 30], and an important consequence of ethical governance is that *all* robots and AIs should be designed as implicit ethical agents.

Beyond reflecting values in their design a logical (but technically very challenging) next step is to provide intelligent systems with an *ethical governor*. That is, a process which allows a robot or AI to evaluate the consequences of its (or others') actions and modify its own actions according to a set of ethical rules[6]. Developing practical ethical governors remains the subject of basic research and presents two high level challenges: (1) the philosophical problem of the formalisation of ethics in a format that lends itself to machine implementation and (2) the engineering problem of the implementation of moral reasoning in autonomous systems [11, 25, 36].

There are two approaches to addressing the second of these challenges [2]:

1. a constraint-based approach – explicitly constraining the actions of an AI system in accordance with moral norms; and

2. a training approach – training the AI system to recognise and correctly respond to morally challenging situations.

The training approach is developed for an assistive robot in [4], while examples of constraint-based approaches are explored in [5, 58]. One advantage of the constraint-based approach is that it lends itself to verification [18].

Note that equipping future IAS with an ethical governor will not abrogate or diminish the human responsibility for careful ethical governance in the design and application of those systems; rather the opposite: robots and AIs that are explicit moral agents are likely to require a greater level of operational oversight given the consequences of such systems making the wrong ethical choice [50]. Explicitly ethical machines remain, at present, the subject of basic research; if and when they become a practical reality there is no doubt that radical new approaches to regulating such systems will be needed.

4 Concluding discussion

In this paper we have argued that, while there is no shortage of sound ethical principles in robotics and AI, there is little evidence that those principles have yet translated into practice, i.e. effective and transparent ethical governance. Ethical practice starts, of course, with the individual and emerging professional codes of ethical conduct such as the recently published ACM code [1] are very encouraging. But individuals need to be supported and empowered by strong institutional frameworks and principled leadership. What would we expect of

[6]Ethical agents should be based on ethical systems other than consequentialism, however computationally modeling such systems remains a difficult research problem.

robotics and AI companies or organisations who claim to practice ethical governance? As a starting point for discussion we propose five pillars of good ethical governance, as follows.

- Publish an **ethical code of conduct**, so that everyone in the organisation understands what is expected of them. This should sit alongside a 'whistleblower' mechanism which allows employees to be able to raise ethical concerns (or 'responsible disclosure'), if necessary in confidence via an ombudsperson, without fear of displeasing a manager.

- Provide **ethics and RI training** for everyone, without exception. Ethics and Responsible Innovation, like quality, is not something that can be implemented as an add-on; simply appointing an ethics manager for instance, while not a bad idea, is not enough.

- Practice **Responsible Innovation**, including the engagement of wider stakeholders within a framework of anticipatory governance (using for instance the AREA framework [24, 46, 34]). Within that framework undertake **ethical risk assessments** of all new products, and act upon the findings of those assessments. A toolkit, or method, for ethical risk assessment of robots and robotic systems exists in British Standard BS 8611 [14], and new process standards such as IEEE P7000: *Model Process for Addressing Ethical Concerns During System Design*, are in draft.

- Be **transparent** about ethical governance. Of course robots and AIs must be transparent too, but here we mean transparency of process, not product. It's not enough for an organisation to claim to be ethical, it must also show *how* is is ethical. This could mean an organisation publishing its ethical code of conduct, membership of its ethics board if it has one (and its terms of reference), and ideally case studies showing how it has conducted ethical risk assessments alongside wider processes of anticipatory governance – these might be part of an annual *transparency report*.

- Really **value** ethical governance. Even if an organisation has the four processes above in place, it – and especially its senior managers – also needs to be sincere about ethical governance; that ethical governance is one of its core values and just not a smokescreen for what it really values (like maximising shareholder returns).

Our final point about really valuing ethical governance is of course hard to evidence. But, like trust, confidence in a company's claim to be ethical has to be earned and – as we've seen – can easily be damaged. Ethical governance needs to become part of a company's DNA, not just in product development but across the whole organisation from management to marketing. In setting

out these pillars of good ethical governance in robotics and AI we are well aware that implementing these practices will be challenging. As Boddington points out "There are indeed very hard questions about how to translate institutional ethical policies into practice" [8][p34]. We are however encouraged to see a very recent example of pressure from within a leading AI company to institute a framework not unlike the one we propose here, as set out in a letter from employees "asking leadership to work with employees to implement concrete transparency and oversight process" [40].

This paper has explored the question of ethical governance in robotics and AI. The paper has argued that ethical governance, while not a singular solution, will be critical to building public trust in robotics and artificial intelligence. It is hard to see how disruptive new intelligent autonomous systems technologies such as driverless cars, assistive robots or medical diagnosis AIs will be widely accepted and trusted without transparent, inclusive and agile ethical governance by the organisations that develop and operate them.

Acknowledgements

This work has, in part, been supported by EPSRC grant ref EP/L024861/1. Development of the ethics roadmap has greatly benefited from discussion with many individuals, especially Christophe Leroux, Vincent Muller, Noel Sharkey and Aimee Van Wynsberghe. We are also very grateful to the anonymous reviewers for their insightful comments and questions.

References

[1] Association for Computing Machinery (2018) ACM Code of Ethics and Professional Conduct, ACM. `https://www.acm.org/code-of-ethics`

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

[3] Ananny, M and Crawford, K (2016) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability, New Media and Society, 1-17.

[4] Anderson, M. and Anderson, S.L. (2014) GenEth: A General Ethical Dilemma Analyzer, in Proc. AAAI, pp.253-261.

[5] Arkin, R.C., Ulam, P., and Wagner, A.R. (2012) Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. Proc. IEEE 100 (3): 571-589.

[6] Billings, D.R., et al (2012). Human-robot interaction: developing trust in robots. Proc. seventh annual ACM/IEEE international conference on Human-Robot Interaction. ACM.

[7] Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B. and Winfield, A.F. (2017), Principles of Robotics, Connection Science, 29 (2), 124-129.

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

[9] Bonnefon, J-F., Shariff, A., and Rahwam, I. (2016) The social dilemma of autonomous vehicles, Science 352 (6293), 1573-1576.

[10] Broekaert, K. and Espinel, V.A. (2018) How can policy keep pace with the Fourth Industrial Revolution? https://www.weforum.org/agenda/2018/02/can-policy-keep-pace-with-fourth-industrial-revolution/

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

[12] Brynjolfsson, E. and Mcafee, A. (2014) The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies, MIT Press.

[13] Bryson, J. and Winfield, A. (2017), Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems, IEEE Computer 50 (5), 116-119.

[14] British Standards Institute (2016), BS8611:2016 Robots and robotic devices: guide to the ethical design and application of robots and robotic systems, ISBN 9780580895302, BSI. London.

[15] Campolo, A., Sanfilippo, M., Whittaker, M. and Crawford, K. (2017) AI Now 2017 Report. https://ainowinstitute.org/AI_Now_2017_Report.pdf

[16] Coeckelbergh, M. (2012) Can we trust robots?. Ethics and information technology 14.1: 53-60.

[17] Caliskan-Islam, A., Bryson, J. and Narayanan, A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334), 183-186.

[18] Dennis, L.A., Fisher, M., Slavkovik, M., and Webster, M. (2016): Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems 77: 1-14, Elsevier.

[19] Dillmann, R. (2004) Benchmarks for Robotics Research, EURON, April 2004. http://www.cas.kth.se/euron/euron-deliverables/ka1-10-benchmarking.pdf

[20] European Commission (2012), Special Eurobarometer 382, Public Attitudes towards Robots, Sept 2012. http://ec.europa.eu/COMMFrontOffice/PublicOpinion/index.cfm/ResultDoc/download/DocumentKy/56814

[21] European Commission (2015), Special Eurobarometer 427, Autonomous Systems, June 2015. http://ec.europa.eu/public_opinion/archives/ebs/ebs_427_en.pdf

[22] Eden, G., Jirotka, J. and Stahl, B. (2013) Responsible research and innovation: Critical reflection into the potential social consequences of ICT, in Proc. IEEE 7th Int. Conf.on Research Challenges in Information Science (RCIS 2013), 1-12, IEEE.

[23] Edwards, L. and Veale, M. (2017) Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. 16 Duke Law and Technology Review 18 (2017). Available at SSRN: https://ssrn.com/abstract=2972855 or http://dx.doi.org/10.2139/ssrn.2972855

[24] EPSRC (2016), The EPSRC AREA framework for responsible Innovation, https://www.epsrc.ac.uk/research/framework/.

[25] Fisher, M., List, C., Slavkovik, M., Winfield, A.F. (2016), Engineering Moral Machines, Informatik-Spektrum, Springer.

[26] Grand, A., Wilkinson, C., Bultitude, K. and Winfield, A.F. (2012) Open Science: A new 'trust technology'? Science Communication, 34 (5), pp. 679-689, Sage.

[27] Grimpe, B., Hartswood, M., and Jirotka, M. (2014) Towards a closer dialogue between policy and practice: responsible design in HCI, Proc. of the Conf. on Human Factors in Computing Systems, pp 2965-2974.

[28] Harper, R.H.R, (ed.) (2014) Trust, computing, and society. Cambridge University Press.

[29] House of Commons (2016), Report of the Science and Technology Committee on Robotics and artificial intelligence, Science and Technology Committee (Commons), http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf

[30] IEEE Standards Association (2016), Global initiative on Ethical Considerations in the Design of Artificial Intelligence and Autonomous Systems, http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[31] IEEE Standards Association (2017a), P7001 Transparency in Autonomous Systems, https://standards.ieee.org/develop/project/7001.html.

[32] IEEE (2017b), Ethically Aligned Design: Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), version 2, IEEE Standards Assoc., https://ethicsinaction.ieee.org/

[33] International Standards Organisation (2013), 13482:2013 Robots and robotic devices: safety requirements for personal care robots, http://www.iso.org/iso/catalogue_detail.htm?csnumber=53820

[34] Jirotka, M., Grimpe, B., Stahl, B., Eden, G., and Hartswood, M. (2017) Responsible Research and Innovation in the Digital Age, Communications of the ACM, Vol 60(5), pp 62-68.

[35] Koene, A., et al. (2015) Research Ethics and Public Trust, Preconditions for Continued Growth of Internet Mediated Research, IEEE Int. Conf. on Information Systems Security and Privacy (ICISSP 2015).

[36] Malle, B.F. (2016) Integrating robot ethics and machine morality: the study and design of moral competence in robots, Ethics and Information Technology, 18(4), 243-256.

[37] Moor, J.H. (2006) The Nature, Importance, and Difficulty of Machine Ethics, IEEE Intelligent Systems, 21(4), 18-21.

[38] Mulgan, G., (2016) A machine intelligence commission for the UK: how to grow informed public trust and maximise the positive impact of smart machines, Nesta, February 2016.

[39] Murphy, Robin and Woods, David D. (2009) Beyond Asimov: The Three Laws of Responsible Robotics. IEEE Intelligent systems. 24 (4): 14-20.

[40] O'Donovan, C. (2018) Here Is The Letter Google Employees Are Signing To Protest A Censored Search Engine For China, Buzzfeed, 16 August 2018. https://www.buzzfeednews.com/article/carolineodonovan/google-dragonfly-maven-employee-protest-demands

[41] Palmerini, E., Azzarri, F., Battaglia, A., Bertolini, A., Carnevale, A., Carpaneto, J., Cavallo, F., Di Carlo, A., Cempini, M., Controzzi, M., Koops, B.J., Lucivero, F., Mukerji, N., Nocco, L., Pirni, A., Shah, H., Salvini, P., Schellekens, M. and Warwick, K. D6.2 – Guidelines on regulating robotics. Pisa, Italy: Robolaw project, 2014, http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf.

[42] Rainey, S., and Goujon, P. (2011). Toward a Normative Ethical Governance of Technology. Contextual Pragmatism and Ethical Governance. In René von Schomberg (ed.) Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields, Report of the European Commission-DG Research and Innovation, http://dx.doi.org/10.2139/ssrn.2436399.

[43] The Rome Declaration on Responsible Research and Innovation (2014). http://www.science-and-you.com/en/sis-rri-conference-recommendations-rome-declaration-responsible-research-and-innovation

[44] Robinette, P., Wagner, A. R., and Howard, A. M. (2013). Building and Maintaining Trust Between Humans and Guidance Robots in an Emergency. In 2013 AAAI Spring Symposium Series.

[45] Schwab, K. (2017) The Fourth Industrial Revolution, Portfolio Penguin.

[46] Stilgoe, J., Owen,R., and Macnaghten, P. (2013) Developing a framework for responsible innovation. Research and Policy 42(9), pp 1568-1580.

[47] Stilgoe, J. (2018) Machine learning, social learning and the governance of self-driving cars. Social studies of science 48 (1), 25-56.

[48] Stilgoe, J. and Winfield, A. F. (2018) Self-driving car companies should not be allowed to investigate their own crashes, The Guardian, 13 April 2018.

[49] Theodorou, A., Wortham, R.H., and Bryson, J.J. (2017), Designing and implementing transparency for real time inspection of autonomous robots, Connection Science Vol 29: Issue 3.

[50] Vanderelst, D. and Winfield, A.F. (2018) The dark side of ethical robots, in Proc. AAAI/ACM conference on Artificial Intelligence, Ethics and Society, New Orleans, Feb 2018. http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_98.pdf

[51] Veruggio G (2006), EURON Roboethics Roadmap: http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf

[52] Wachter, S. and Mittelstadt, B. and Russell, C. (2017), Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR (October 6, 2017). Harvard Journal of Law and Technology, Forthcoming. Available at SSRN: https://ssrn.com/abstract=3063289 or http://dx.doi.org/10.2139/ssrn.3063289

[53] Webster, M., Dixon, C., Fisher, M., Salem, M., Saunders, J., Koay, K-L., Dautenhahn, K., and Saez-Pons, J. (2016) Toward Reliable Autonomous Robotic Assistants Through Formal Verification: A Case Study IEEE Transactions on Human-Machine Systems 46(2):186-196

[54] World Economic Forum (2018) White paper on Agile Governance: Reimagining Policy-making in the Fourth Industrial Revolution. http://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf

[55] Wilkinson, C. and Weitkamp, E. (2016) Creative research communication: Theory and practice. Manchester: Manchester University Press.

[56] Wilsdon, J. and Willis, R. (2004) See-through science: Why public engagement needs to move upstream, Demos.

[57] Winfield, A. F. (2012) Robotics: A very short introduction, Oxford University Press.

[58] Winfield, A. F., Blum, C. and Liu, W. (2014) Towards an ethical robot: Internal models, consequences and ethical action selection. In: Mistry, M., Leonardis, A., Witkowski, M. and Melhuish, C., eds. Advances in Autonomous Robotics Systems, LNCS 8717, 85-96, Springer.

[59] Winfield, A. F. and Jirotka M. (2017) The case for an Ethical Black Box, In: Gao, Y., et al eds. Towards Autonomous Robot Systems, LNAI 10454, 262-273, Springer.

[60] Winfield, A. F. (2017) A Round Up of Robotics and AI ethics: part 1 Principles, http://alanwinfield.blogspot.com/2017/12/a-round-up-of-robotics-and-ai-ethics.html

[61] Wortham, R.H. and Theodorou, A. (2017) Robot transparency, trust and utility, Connection Science, 29(3), pp 242-248.

[62] Van Wynsberghe, A., and Sharkey, N. (founders, 2016), Foundation for Responsible Robotics (2016), website http://responsiblerobotics.org/