Explicitly ethical standards for robotics

Cian O'Donovan[a]

Working paper for discussion at the international symposium on post-automation,
SPRU – Science Policy Research Unit, University of Sussex[b]
11-13th September, 2019

Abstract

This paper explores how explicitly ethical standards for robotics are peer-produced. It describes the motivations, organisation and practices of standardization contributed by a globally distributed community of experts. The research question asks what kind of rules for robots are being created through standardization and what are the motivational and organizational features of this knowledge production? In addressing this question, I reflect on how ethical principles are applied in practice within the field of autonomous and intelligent systems and what implications this may have for the governance of robotics innovation. The paper directly responds to the aims of the workshop by speculating on the potential for post-automation robotics innovation pathways that are not automatically determined, but arrived at by means of broad participation in governance decisions and innovation processes.

Keywords

Robotics and autonomous systems, artificial intelligence, peer-production, standards, science and technology studies, ethics

Table of contents

[a] Department of Science and Technology Studies, University College London. Email c.o'donovan@ucl.ac.uk

> *"The point about principles is that principles are not a method [bangs table]. Principles are simply a set of values, if you like, an expression of ideals. What we need is a method, something that someone can actually do. A standard is something you can do... a standard is a way of doing things well [banging]."*

- Senior roboticist and standards contributor at a UK robotics research laboratory.

1. Introduction

Standards are means by which social and technological worlds are brought together to make things work. They are the rules, guidelines and procedures which ensure electricity gets to our homes; trains traverse our rail networks; data flows over ethernet, wi-fi and 4G; and safety is assured through testing and good practice (Hughes, 1983; Kaijser and Vleuten, 2006; Vinsel, 2019). In this paper I discuss recent initiatives designed to create *explicitly ethical standards* for robotics and autonomous and intelligent systems[3]. The central research is what kind of rules for robots[4] are being created through standardization and what are the motivational and organizational features of this knowledge production?

This paper addresses the research question through interviews with key protagonists involved in one particular standardization project, the IEEE's *Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems*. In Section 2 I set the scene and give an overview of the state of the art in the development of ethical standards in and for A/IS. In Section 3 I describe the motivations and organization of the knowledge production processes of a set of standards currently under development, the IEEE P7000 Series. In Section 4 I discuss implications for the governance of emerging technologies and for ideas about post-automation – that is alternative arrangements for how people, robotics and autonomous systems can exist in more appropriate, caring and sustainable configurations. Interview data is supported by observation at a set of academic, industry and policy conferences and workshops as well as through reviews. Notes on methodology and other descriptive details are appended in an annex.

2. Concerns about innovation in robotics, autonomous and intelligent systems

A strategic focus on innovation in robotics is a constituent part of Industrie 4.0 rhetorics in Germany and elsewhere, the UK's Industrial Strategy and similar framings of industry-technology change around the world. Supporters of innovation under these framings encourage speed of research, development and diffusion over other considerations of emerging technologies. The distribution of benefits, allocation of costs, mapping of often ambiguous impacts and uncertain outcomes are secondary concerns, if they feature at all. Yet societal concerns, voiced through social movement organizations, collective action, media commentary and academic research suggest that existing governance of robotics and autonomous systems technologies is insufficient (Torresen, 2018; Winfield *et al.*, 2019).

[3] Space constraints preclude an in-depth discussion of an emerging issue in the field, that of ethics-washing (Wagner, 2018).

[4] By which I refer throughout to a broad set of technologies such as vision systems, machine learning and other software and hardware that constitute what I call robotics and autonomous and intelligent systems, or simply A/IS or robotics.

While once situated predominantly in industrial and military settings (Noble, 2011), A/IS are now increasingly prevalent in a far wider range of industries and life domains, such as health and social care, education, service sectors and non-industrial design and manufacturing (Prescott and Caleb-Solly, 2017; Winfield *et al.*, 2019; O'Donovan and Smith, 2020). Quality assurance and safety issues that were at the fore in manufacturing and heavy industry, are now augmented by a range of other concerns such as data protection, privacy, transparency and issues of democracy (Mahieu *et al.*, 2018; Whittaker, M., Crawford, K., Dobbe, 2018).

Issues including uncertainty, complexity, multiple and sometimes opaque interests and the uneven distribution of costs and harms mean that the governance of emerging technologies (Rotolo, Hicks and Martin, 2015) is inherently difficult. The field of robotics is no exception. Mitigation strategies include demand side market solutions, such as the withdrawal by advertisers of ads from platforms; self-assessment and regulation by firms, such as corporate social responsibility efforts; state intervention policies; regulatory oversight by novel or existing regulators (Koene, 2019).

One set of responses to these concerns has been the production of ethical frameworks with which to govern A/IS (Winfield, 2017). But ethics achieve force in the world only through interlinked and often complex changes in the cultures and practices of innovation and through implementation and imbedding in regulatory and political structures such as soft and hard law. Effective frameworks require operationalization. Standardization offers one such means for integrating values and ethics into the practices of the development and use of emerging technologies. By closely examining such standards, how they are created and by whom, I seek to illuminate one avenue by which A/IS may be governed.

Like all forms of knowledge production, standards are the outcome of social and often political and politicised processes. They are often peer-produced by teams of globally distributed inter-disciplinary and inter-sectoral expert volunteers. The production of standards are usually sponsored by transnational standard organisations such as the IEEE and ISO, or national organisations such as the British Standards Institution. Production is financed through various models. Through sponsorship by large technology firms for example, through voluntary labour, or through membership fees and other revenues of professional bodies. Standards may be free at the point of use, or be controlled via strict IP law where end use is charged.

Standards are sites of value judgements, and represent decisions and ways of knowing the world, and ways of knowing how the world should be. Standards are, according to Bryson and Winfield (2017):

> "consensus-based agreed-upon ways of doing things, setting out how things should be done. If a system or process can be shown to do things as prescribed, it is said to be compliant with the standard. Such compliance provides confidence in a system's efficacy in areas important to users, such as safety, security, and reliability".

Standards have had a significant impact on individual and societal health, well-being and safety. Without standards, large scale socio-material systems such as energy infrastructures and medical technologies would not function in the way they do today.

For example, for autonomous guided vehicles, and for industrial robots and robot systems. These standards include, for AGVs, according to Franklin (2019):

> *"the U.S. standard B56.5 and the International Organization for Standardization (ISO) 3691-4; and for industrial robots and robot systems, R15.06 in the U.S., the national adoption of ISO 10218-1,2. However, neither standard fully addresses the current state-of-the art of robot mobility. R15.06 was developed at a time when industrial robots were bolted in place, not mobile. B56.5 was developed around the capabilities of devices that did not possess sufficient autonomy to operate safely away from their predetermined paths"*

So, can standards help re-configure a social and technological configurations that incorporate a plurality of ethical standpoints, promotes a diversity of values, is commensurate with sustainability, and broadens participation in who gets a say in governing emerging autonomous and intelligent systems. In short, can standards be incorporated as a means of promoting post-automation innovation?

3. From principles to practice: IEEE P7000 Series

The IEEE is a professional membership organisation of over 400,000 people. The IEEE Standards Association, its standards division, has a portfolio of over 1,100 active standards and over 500 standards under development (as of 2017). According to its own promotional material, its work on ethics can be broken into three interdependent area. First, establishing codes of ethics and professional guidelines that might help define intended behaviours for its members, other professionals in the field, and influence third party ethics frameworks such as AI Now and national policy reports. Second, work on impacting human behaviour in the context of these guidelines, through for example, ethics education. Third, work on ethical and societal impact of the technologies themselves.

In April 2016, as part of this third component, the IEEE-SA established the Global Initiative on Ethics of Autonomous and Intelligent Systems to bring together global experts in robotics, autonomous systems, engineering, data sciences, social sciences, humanities and ethics. By May of 2018 there were more than 850 A/IS ethics professionals involved from the US, the EU, Australia, India, China, Korea, Japan and other nations. It had 13 committees creating content in addition to out-reach support and dissemination workers. The *Global Initiative* aimed in their words to provide "an incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies" (*The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.*, 2019). The Global Initiative has two operational pillars. The 294 page Ethically Aligned Design public discussion document (IEEE, 2019), detailed in Annex 2. and the P7000 Series of ethical standards

3.1. The IEEE P7000 Series of ethics standards

The IEEE P7000 Series has been established as a core outcome of the EAD. According to Winfield 2019, in addition to their main objectives, each EAD writing committee was tasked with "identifying, recommending and promoting new candidate standards, and a total of 14 new IEEE standards working groups have started work on drafting so-called human standards", listed in Table 1. While more traditional standards focus on technology interoperability, functionality, safety, and trade facilitation, "the IEEE P7000 series addresses specific issues at the intersection of technological and ethical considerations" (Havens2019:70). And although there are existing standards that relate to A/IS and aspects of wellbeing such as safety, the P7000 series is the first to deal with explicitly ethical concerns.

Table 1. IEEE P7000 Series standards in development

Standard	Description
7000	Engineering Methodologies for Ethical Life-Cycle Concerns Working Group
7001	Transparency of Autonomous Systems
7002	Personal Data Privacy Working Group
7003	Algorithmic Bias Working Group
7004	Standard for Child and Student Data Governance
7005	Employer Data Governance Working Group
7006	Personal Data AI Agent Working Group
7007	Ontological Standard for Ethically Driven Robotics and Automation Systems
7008	Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems Working Group
7009	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
7010	Well-being Metric for Autonomous and Intelligent Systems (A/IS)
7011	News Site Trustworthiness Working Group
7012	Machine Readable Privacy Terms Working Group
7013	Benchmarking Accuracy, Increasing Transparency, and Governing Use of Automated Facial Analysis Technology Working Group

Each of the individual standards projects were initiated due to what participants have called "the necessities and requirements arising from the IEEE Global Initiative.] (Havens and Hessami, 2019). So, what is it that these standards do? According to Havens and Hessami (2019, p. 70)

> *“like its technical standards counterparts, the IEEE P7000 series empowers innovation across borders and yields societal benefits”.*

This is a substantial claim. And it is remarkable that even in presenting solutions aimed at mitigating the harms of A/IS innovation, this language reinforces mainstream science and technology policy's most dominant trope: innovation as an unquestionable force for progress. Decades of research in the critical social sciences, and centuries of social movement building and collective action tell us this is not the case. However, it would be both unfair, and premature to judge these claims based on rhetoric alone. Unfair, because in the EAD and supporting documentation there is strong evidence for the inclusion of a plurality of perspectives and ways of knowing and understanding the world. This evidence suggests the report to date is far from a siloed effort authored by a narrow interest group. Premature because the standards are not yet published. As such, it is appropriate that this paper assess the possibilities for post-automation that the P7000 series presents. To explore these possibilities, I examine in closer detail the development process behind one particular standard, P7003.

3.2. P7003 Algorithmic Bias Considerations: addressing the 'messy' problem of fair algorithms.

Any A/IS that will produce different results for some people than for others is open to challenges of being biased. Examples could include: security camera applications that detect theft or suspicious behaviour, marketing automation applications that calibrate offers, prices, or content to an individual's preferences and behaviour. These and other issues of bias in algorithms have been increasingly problematised in recent years (O'Neil, 2016; Noble, 2018; Eubanks, 2019).

Motivations behind the P7000 Series

The creators of P7003 aim to address these problems (Koene, Dowthwaite and Seth, 2018):

> *"IEEE P7003 is aimed to be used by people/organizations who are developing and/or deploying automated decision (support) systems (which may or may not involve AI/machine learning) that are part of products/services that affect people. Typical examples would include anything related to personalization or individual assessment, including any system that performs a filtering function by selecting to prioritize the ease with which people will find some items over others (e.g. search engines or recommendation systems)."*

In short, the standard is designed to provide the creator of algorithms certification oriented methodologies to provide accountability and clarity around how algorithms are targeting, assessing and influencing users and stakeholders. It is intended that follow-up certification under the standard will allow algorithm creators to communicate to their users, as well as regulators, that up-to-date best practices were used in the design, testing and evaluation of the algorithm to avoid unjustified differential impact on users.

The requirements specification provided by the IEEE P7003 promise that the standard will allow algorithm creators to communicate to users, and regulatory authorities, that up-to-date best practices were used in the design, testing and evaluation of the algorithm to attempt to avoid unintended, unjustified and inappropriate differential impact on users (IEEE, 2016):

> *"The standard describes to help users certify how they worked to address and eliminate issues of negative bias in the creation of their algorithms, where "negative bias" infers the usage of overly subjective or uniformed data sets or information known to be inconsistent with legislation concerning certain protected characteristics (such as race, gender, sexuality, etc); or with instances of bias against groups not necessarily protected explicitly by legislation, but otherwise diminishing stakeholder or user well-being and for which there are good reasons to be considered inappropriate.*
>
> *Possible elements include (but are not limited to): benchmarking procedures and criteria for the selection of validation data sets for bias quality control; guidelines on establishing and communicating the application boundaries for which the algorithm has been designed and validated to guard against unintended consequences arising from out-of-bound application of algorithms; suggestions for user expectation*

> *management to mitigate bias due to incorrect interpretation of systems outputs by users (e.g. correlation vs. causation).*

The underlying the standard design process in the P7003 project has problematised the issue of bias in some interesting ways. Notably, the lead authors suggest the issue of bias in algorithms as inherently socio-technical. Fairness according to (Koene, Dowthwaite and Seth, 2018) is fundamentally a "societally defined construct". These authors bring to bear on the problem observations about cultural differences between nations and jurisdictions as well as cultural changes in time. And so for them, not only must algorithms be transparent, but they must be adaptable to change through balanced processes. According to (Koene, 2019), their problematisation of bias has drawn attention to a number of key questions should be addressed when developing of deploying an algorithmic system. Such as who will be affected? What are the decision or optimisation criteria? How are these criteria justified? Are these justifications acceptable in the context where the system is used? Such a problematisation would seem amenable to issues post-automation, to which I will return in Section 4.

<u>Organisation of the P7000 Series</u>

So how do these standards get made and by whom? Standards, according to (Winfield, 2019), are:

> *"community-driven initiatives, formed by dedicated, expert volunteers who propose, debate, draft and re-draft until a consensus is met. They are an important step in establishing trust in a new technology, and therefore realizing its full potential."*

One interviewee describes an early meeting of one of the P7000 standards

> *"The people who turned up were an extraordinary, eclectic mix. This is one of the exciting things that you're, sitting around the table with philosophers, with of course, all the usual suspects. Engineers, computer scientists, technologists, but also diplomats. People from the United Nations, people from, quangos, and NGOs. The UN and, and all kinds of, of [banging] amazing unions, as well, of course, as, industry people. People from companies and so on. So a really extraordinary mix of, backgrounds. People from the WEF, people from, Oxfam, you know.*
>
> *"A really extraordinary, range of interests and backgrounds. Now, I think, when it comes to volunteering to work on working groups, I suspect it narrows a little, but not, not completely. So, many of those kinds of backgrounds are also represented in the working groups. Of course, working groups are a different story because, you know. They, they need a bit more commitment."*

While the interviewee above stressed the organisation contributors came from, the P7003 working group consisted of individuals rather than affiliations. At the time of writing there are between 80 and 100 people on the working groups listserv. Between 20 and 30 members participate in working calls each month with a core of eight to ten people responsible for driving the work forward.

The core working group team members and expertise came from the UK's Engineering and Physical Science Research Council project, *Unbias: Emancipating Users Against Algorithmic Biases for a*

Trusted Digital Economy[5]. It seems that the Ethically Aligned Design initiative created the space of possibility and opportunity for the development of the standard, but the motivation of the particular stance on bias in algorithms as well as core personnel came 'pre-packaged.

The P7003 working group have created what they call four multi-disciplinary foundational sections. A taxonomy of algorithmic bias, person categorisation and identifying affected population groups, legal frameworks related to bias and a psychology of bias. These have been followed by six development sections that work on algorithmic system design stages' assurance of representativeness of testing/training/validation data' evaluation of algorithmic processing; assessment of resilience against external manipulation to bias; and documentation of criteria, scope and justification of choices. The up-front and transaction costs of joining the working group is low. One interviewee suggested people turned up on calls and they were unsure how they got there or what they had to contribute, but they were welcome anyway.

4. Discussion and conclusions: Opportunities and potentials for post-automation

What role then standards and post-automation? Here I understand post-automation in relation to the workshop call[6] as Post– in the sense of recognising and reappraising the human agency in 'automation' technologies. Post– in the sense that users appropriate automation technologies into non-industrial and new-industrial spaces beyond conventional manufacturing circuits and logics. Post– in the sense that groups are following alternative work-life visions, e.g. pursuing creative livelihoods and environmental sustainability with these technologies, and as such seek sociotechnical relations that contrast starkly with the labour productivities pursued in conventional automation. Post– because we are witnessing reappraisal and tailoring of a subversions of technologies whose genealogies go back to an earlier wave of struggle over automation in manufacturing, and which is instructive for collective action today. And post– in the sense that social theory is currently inadequate to understanding the causes, consequences and social significance of this phenomenon.

From the evidence gathered to date, IEEE P7000 Series standards are predominantly methodologically individualistic: they seek to avoid harms done to individual humans but are less concerned with social and collective issues. Yet, notably in the case of P7003, the project working group's conceptualisation of the world illustrates an awareness of the complexity and situated-ness of algorithms, and the importance of context in mitigating harms. P7003 seems particularly reflexive in this regard. In its foundational principles it acknowledges that 'algorithmic systems do not exist in a vacuum', that they are built, deployed and used by 'people within organisations, within a social, political, legal and cultural context.' If post-automation pathways are to be developed, then explicitly recognising the situated and contingent nature of automation is a must.

The work-to-date also points to roles that the peer-production of standards might play in instigating a collective intelligence (Benkler, Shaw and Hill, 2015)) that may, undergird the governance of

[5] See https://unbias.wp.horizon.ac.uk/ (last accessed July 20th, 2019)

[6] Acknowledging the call by Smith and Fressoli

autonomous and intelligent systems. This is notable for at least two reasons. First, the production of standards (potentially) lays outside of state and firm control. Second, it draws attention to the creation and strengthening relations between collective (human) intelligence and A/IS, an issue which has risen up the agenda of innovation agencies such as NESTA in recent years[7].

Also intriguing at this stage of the research are questions about who is creating these standards and how. We must acknowledge homogeneity and exclusions here along some dimensions. A US and western Europe focus in working group composition. Gender composition that based on current evidence is representative of robotics and engineering fields rather than technology users. And I remain uncertain about class and other demographic features. Yet, evidence suggests a broader base of participants in terms of knowledge discipline and diversity of occupations. Furthermore, the production and use of standards introduces further locations and strategies of contention and unruliness. These might augment strikes and other workplace demonstrations in the repertoire of tech-workers seeking to reclaim agency in pursuit of steering innovation in societally appropriate directions. This broadening out of participation and opening up the outputs of knowledge production to new locations is to be, at least conditionally, welcomed.

There is also some evidence for spillover effects - in other words, unintended but substantive consequences of this knowledge production. We've seen ideas from the Ethically Aligned Design project emerge through other venues such as parliamentary reports. Speculatively, it seems the standardisation process lends legitimacy, credibility and importantly an enrolment mechanism to ongoing activities and outputs connected to the wider Global Initiative, and vice versa - an opening-up of the locations where these.

Summarising the together the above discussion I outline some thoughts in relation to these dimensions of post-automation in Table 2.

Table 2. Post-automation potentials

Reappraising human agency	Standards offer a location at which to re-orientate human agency.
New appropriation locations	One of the drivers of these standards is the role that A/IS are now playing in locations beyond industrial setting. Actors at these locations include children, elderly and sick people. However, it is unclear if and how these standards might enhance agency for these people to deal with A/IS technologies on their own terms.
New forms of sociotechnical relations	It is not so clear in what sense groups pursuing creative livelihoods and sustainability with these technologies are impacted. This is a clear potential location of peer-produced governance arrangements that are outside the control of state or firm actors
Subverting industrial technology	Explicitly ethical standards offer another location, in addition to strikes and other demonstrations where tech-workers themselves can reclaim agency in pursuit of steering innovation in societally appropriate directions
Beyond current social theory	New forms of collective intelligence

[7] See https://www.nesta.org.uk/project/centre-collective-intelligence-design/ last accessed July 20th, 2019

Broadening out and opening up ethical inputs and opportunities for post-automation	People doing ethics in AI related fields tend to be ethicists and engineers according to (Mahieu *et al.*, 2018) who write that that to get a thorough understanding of, and grip on, all the hard ethical questions of a digital society, ethicists, policy makers and legal scholars will need to familiarize themselves with the concrete and practical work that is being done across a range of different scientific fields to deal with these questions. Initial findings are encouraging in this regard.

In organising the discussion of the production of the IEEE P7003 standard, I loosely followed (Benkler et al 2015) framework which assesses motivation and organisation, for the moment at least leaving out an explicit focus on quality. Future work might specifically address issues of quality within the P7000 series projects. Indeed, a major question left unaddressed in this paper concerns the quality and efficacy of the standards themselves. What impact, negative or positive, do these standards have on the world. In terms of the standards themselves, we cannot know. The IEEE will publish standards as they are published, most likely starting some time in 2020. Clearly further research is needed to track the efficacy. Most optimistically, these standards will provide institutional steering, directing innovation pathways towards more ethically appropriate ends, though we must be suspicious of any ethical framework that neglects to make clear whose ethics matter.

Word count: 4,039

Acknowledgements

The research was funded through the European Union’s Horizon 2020 research and innovation program under grant agreement No. 788359 (“SCALINGS: Scaling up co-creation? Avenues and limits for including society in science and innovation”). I am grateful for the participation of interviewees and other expert respondents.

Biography

Cian O’Donovan is a research associate at the Department of Science and Technology Studies, University College London and associate faculty at SPRU - Science Policy Research Unit, University of Sussex. He is currently researching the practices, policies and post-automation potentials of robotics innovation in Europe (www.scalings.eu).

ORCID https://orcid.org/0000-0003-4467-9687

References

Benkler, Y., Shaw, A. and Hill, B. (2015) 'Peer Production: A Modality of Collective Intelligence', in Malone, T. and Bernstein, M. (eds) *Handbook of Collective Intelligence*. Cambridge, Massachusetts: MIT Press, pp. 1–27.

Bryson, J. and Winfield, A. (2017) 'Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems', *Computer*, 50(5), pp. 116–119. doi: 10.1109/MC.2017.154.

Eubanks, V. (2019) *Automating inequality: how high-tech tools profile, police, and punish the poor.* New York, NY: St. Martin's Press.

Franklin, C. (2019) *New U . S . Standard Sets the Bar for Industrial Mobile Robot Safety*, *International Federation of Robotics*. Available at: https://ifr.org/post/new-u.s.-standard-sets-the-bar-for-industrial-mobile-robot-safety ! (Accessed: 11 June 2019).

Havens, J. C. and Hessami, A. (2019) 'From Principles and Standards to Certification', *Computer*. IEEE, 52(4), pp. 69–72. doi: 10.1109/MC.2019.2902002.

Hughes, T. P. (1983) *Networks of Power: Electrification in Western Society, 1880-1930*. Baltimore: The Johns Hopkins University Press.

IEEE (2016) 'P7003 Algorithmic Bias Considerations', pp. 5–6.

IEEE (2019) *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. First edit. Institute of Electrical and Electronics Engineers (IEEE).

International Panel on Social Progress (2018) *Rethinking Society for the 21st Century : Report of the International Panel on Social Progress. Volume 1, Volume 1,*. Cambridge, UK.

Kaijser, A. and Vleuten, E. van der. (2006) *Networking Europe: transnational infrastructures and the shaping of Europe, 1850-2000*. Sagamore Beach: Science History Publications / USA.

Koene, A. (2019) 'Industry Standards as vehicle to address socio-technical AI challenges', in *JRC HUMAINT project Winter School on AI: ethical, social, legal and economic impact*. Available at: https://www.slideshare.net/AnsgarKoene/industry-standards-as-vehicle-to-address-sociotechnical-ai-challenges.

Koene, A., Dowthwaite, L. and Seth, S. (2018) 'IEEE P7003^{TM} standard for algorithmic bias considerations', *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE/ACM, pp. 38–41. doi: 10.1145/3194770.3194773.

Mahieu, R. *et al.* (2018) 'From dignity to security protocols: a scientometric analysis of digital ethics', *Ethics and Information Technology*. Springer Netherlands, 20(3), pp. 175–187. doi: 10.1007/s10676-018-9457-5.

Noble, D. F. (2011) *Forces of production : a social history of industrial automation*. New Brunswick, N.J.: Transaction Publishers.

Noble, S. U. (2018) *Algorithms of oppression: data discrimination in the age of Google*. New York, NY: New York University Press.

O'Donovan, C. and Smith, A. (2020) 'Technology and Human Capabilities in UK Makerspaces', *Journal of Human Development and Capabilities*. Taylor & Francis, 21(1), pp. 63–83. doi:

10.1080/19452829.2019.1704706.

O’Neil, C. (2016) ‘Weapons of math destruction: how big data increases inequality and threatens democracy’. New York: Random House. Available at: http://link.overdrive.com/?websiteID=100434&titleID=2525880.

Prescott, T. and Caleb-Solly, P. (2017) ‘Robotics in Social Care: A Connected Care EcoSystem for Independent Living’, *EPSRC UK-RAS Whitepaper*, pp. 2–25.

Rotolo, D., Hicks, D. and Martin, B. R. (2015) ‘What is an emerging technology?’, *Research Policy*, 44(10), pp. 1827–1843. doi: 10.1016/j.respol.2015.06.006.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019) *IEEE*. Available at: https://ethicsinaction.ieee.org/ (Accessed: 18 July 2019).

Torresen, J. (2018) ‘A Review of Future and Ethical Perspectives of Robotics and AI’, *Frontiers in Robotics and AI*, 4(January). doi: 10.3389/frobt.2017.00075.

Vinsel, L. (2019) *Moving violations. Automobiles, experts, and regulations in the United States*. Johns Hopkins University Press.

Wagner, B. (2018) ‘Ethics as an escape from regulation: From ethics-washing to ethics-shopping.’, in Hilderbrandt, M. (ed.) *Being Profiling. Cogitas ergo sum*. Amsterdam: Amsterdam University Press.

Whittaker, M., Crawford, K., Dobbe, R. et al. (2018) *AI Now Report 2018*, *AI Now*. Available at: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

Winfield, A. (2017) *A Round Up of Robotics and AI ethics*, *Alan Winfield’s Web Log*. Available at: http://alanwinfield.blogspot.com/2017/12/a-round-up-of-robotics-and-ai-ethics.html (Accessed: 21 February 2019).

Winfield, A. *et al.* (2019) ‘Ethical Issues for Robotics and Autonomous Systems’. UK-RAS Network. Available at: https://www.ukras.org/wp-content/uploads/2019/07/UK_RAS_AI_ethics_web_72.pdf.

Winfield, A. (2019) ‘Ethical standards in robotics and AI’, *Nature Electronics*. Springer US, 2(2), pp. 46–48. doi: 10.1038/s41928-019-0213-6.

Annex 1. Methods and materials

By closely examining standards, how they are created and by whom, this paper seeks to illuminate one avenue by which autonomous and intelligent systems are governed. Like all forms of knowledge production, standards are the outcome of social and often political and politicised processes. They are often peer-produced by teams of globally distributed inter-disciplinary and inter-sectoral expert volunteers. They are sites of value judgements, and represent decisions and ways of knowing the world, and ways of knowing how the world should be. As such, a qualitative approach is appropriate, and I follow a situated research approach that uses case study and a strategic literature review of academic and grey literature.

The methodology consists of two parts. First a close reading of the Ethically Aligned Design, the substantial and ambitious written output from The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2019). Particular attention was paid to explicit and implicit visions of contemporary and future roles for robotics in social settings, ethical frameworks and actor positions.

In the second part of the methodology I sought to develop a narrative of how the EAD came to be and how principles evident in the document were operationalised in the IEEE P7000 series of standards. This work is ongoing, and in this draft I report on one of 14 projects within the series, the P7003 Standard for Algorithmic Bias Considerations. I report on this project for pragmatic reasons. First constraints of space in this discussion document. And second according to participants it has made good progress and as such makes for a compelling case study. Of course, reasons why other projects may not be as advanced are relevant to this study and may be discussed in future drafts of this paper.

For the interviews, I chose a semi-structured, narrative-generating approach (Flick, 2010; Lamnek, 1989) to gain insights into the motivations, the accompanying visions and expectations, and the assessments of the actual developments from most of the actors involved.

Insights gained from literature reviews and readings were augmented by a series of visits to Bristol Robotics Lab, the largest such facility in the UK, in 2018 and 2019 as well as major robotics conferences, workshops and networking events.

Annex 2. Background on the Global Initiative and Ethically Aligned Design, Vol. I.

The first draft of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* was released as a Request for Input in December of 2016 and received over two hundred pages of feedback about the draft. Version 2 was released in December 2017 and the final version, the first edition, was launched in February of 2019. During the course of these drafts, the EAD established eight general principles that aim to "further human values and 'ensure trustworthiness' of A/IS". These are concerned with human rights, wellbeing, what they term data agency, effectiveness of technologies, transparency, accountability and responsibility, and finally awareness of misuse and competence.

The construction of these principles and the document itself followed generally the IEEE-SA model of collaborative project management. Authority for chapters, sections and discrete writing tasks is delegated and distributed to smaller teams. IEEE-SA provides some project resources such as governance arrangements for group formation and management including collaboration and consensus guidelines, communications infrastructure such as email list-servers and bespoke project management tools, established document review procedures, editing and meeting planning guidance. The majority of meetings take place through monthly video conferences. Some financial assistance is typically provided for occasional face to face meetings however participation in the project is, apart from a small coordination team, voluntary. These operating principles have special importance for the IEEE-SA because the U.S. Department of Justice has held that standards organizations are responsible for the actions of their standards developers. While this is not directly applicable to work on the EAD, the principle has informed the tools and culture of collaborative working within the IEEE.

Using the eight principles as a foundation, the EAD is being used as a foundational document to underpin a number of ongoing initiatives and interventions. These include the creation of the IEEE P7000 Series of ethical standards, an online course for business professionals, the creation of a fascinating glossary of A/IS Ethics terms that recognises and maps concepts across diverse knowledge communities such as legal scholars, economists, engineers and social scientists. The document's prescriptions and recommendations are also used as arguments with which to influence government reports; it is cited for example in the House of Lords report, "AI in the UK, ready, willing and able". In this regard it is loosely analogous in both production and dissemination to the IPCC climate change reports and the recent International Panel on Social Progress project and report (International Panel on Social Progress, 2018).