Pre-print of article published 15 February 2019 in Nature Electronics: https://doi.org/10.1038/s41928-019-0213-6

Ethical standards in Robotics and AI

A new generation of ethical standards in robotics and artificial intelligence is emerging as a direct response to a growing awareness of the ethical, legal and societal impact of the fields. But what exactly are these ethical standards and how do they differ from conventional standards?

Alan Winfield

Standards are a vital part of the infrastructure of the modern world: invisible, but no less important than roads, airports and telephone networks. It is hard to think of any aspect of everyday life untouched by standards. The International Organization for Standardisation (ISO) – just one of several standards bodies – lists a total of 22,482 published standards. Take the simple act of brushing your teeth in the morning: there are standards for your toothbrush (both manual ISO 20126 and electric ISO 20127), your toothpaste and its packaging (ISO 11609), and the quality of your tap water (ISO 5667-5). Although it might seem odd to wax lyrical on standards, they do represent a truly remarkable body of work – drafted by countless expert volunteers – with an extraordinary impact on individual and societal health and safety.

All standards embody a principle and often it is an ethical principle or value. Safety standards are founded on the general principle that products and systems should do no harm – that they should be safe; ISO 13482, for instance, sets out safety requirements for personal care robots. Quality management standards, such as ISO 9001, describe how things should be done, and can be thought of as expressing the principle that shared best practice leads to improved quality. And technical standards, like IEEE 802.11 (better known as WiFi), can be thought of as embodying the benefits of interoperability. Even the basic idea of standards as codifying shared ways of doing things can be thought of as expressing the values of cooperation and harmonisation. All standards can therefore be thought of as *implicit* ethical standards.

We can define an *explicit* ethical standard as one that addresses clearly articulated *ethical* concerns, and seeks – through its application – to, at best remove, hopefully reduce, or at the very least highlight the potential for unethical impacts or their consequences.

What are the ethical principles which underpin these new ethical standards? An informal survey[1] in December 2017 listed a total of ten different sets of ethical principles for robotics and AI. The earliest (1950) are Asimov's laws of robotics: important because they established the principle that robots should be governed by principles. Very recently we have seen a proliferation of principles; of the ten sets surveyed seven were published in 2017.

Perhaps not surprisingly these ethical principles have much in common. In summary: robots and artificial intelligences (AIs) should do no harm, while being free of bias and deception; respect human rights and freedoms, including dignity and privacy, while promoting well-being; and be transparent and dependable while ensuring that the locus of responsibility and accountability remains with their human designers or operators. Just as interesting is the increasing frequency of their publication: clear evidence for a growing awareness of the urgent need for ethical principles for robotics and AI. But, while an important and necessary foundation, principles are not practice. Ethical standards are the next important step toward ethical governance in robotics and AI[2].

Ethical risk assessment

Almost certainly the world's first explicit ethical standard in robotics is BS 8611 *Guide to the Ethical Design and Application of Robots and Robotic Systems*[3], which was published in April 2016. Incorporating the EPSRC principles of robotics[4], BS8611 is not a code of practice, but instead guidance on how designers can undertake an *ethical risk assessment* of their robot or system, and mitigate any ethical risks so identified. At its heart is a set of 20 distinct ethical hazards and risks, grouped under four categories: societal, application, commercial & financial, and environmental. Advice on measures to mitigate the impact of each risk is given, along with suggestions on how such measures might be verified or validated. The societal hazards include, for example, loss of trust, deception, infringements of privacy and confidentiality, addiction, and loss of employment. The idea of ethical risk assessment is of course not new – it is essentially what research ethics committees do – but a method for assessing robots for ethical risks is a powerful new addition to the ethical roboticist's toolkit.

In April 2016, the IEEE Standards Association launched a global initiative on the Ethics of Autonomous and Intelligent Systems[5]. The significance of this initiative cannot be overstated; coming from a professional body with the standing and reach of the IEEE Standards Association it marks a watershed in the emergence of ethical standards. And it is a radical step. As I've argued above all standards are – even if not explicitly – based on ethical principles. But for a respected standards body to launch an initiative which explicitly aims to address the deep ethical challenges that face the whole of autonomous and intelligent systems – from driverless car autopilots to medical diagnosis AIs, drones to deep learning, and care robots to chat bots – is both ambitious and unprecedented.

Humanity first

The IEEE initiative positions human well-being as its central tenet[6]. This is a bold and political stance since it explicitly seeks to reposition robotics and AI as technologies for improving the human condition rather than simply vehicles for economic growth. The initiative's mission is "to ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity".

The first major output from the IEEE Standards Association's global ethics initiative is a discussion document called Ethically Aligned Design (EAD)[7], developed through an iterative process which invited public feedback. The published second edition of EAD sets out more than 100 ethical issues and recommendations, and a third edition will be launched early in 2019. The work of more than 1000 volunteers across thirteen committees, EAD covers: general (ethical) principles; how to embed values into autonomous intelligent systems; methods to guide ethical 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.

Each EAD committee was additionally tasked with identifying, recommending and promoting new candidate standards, and – to date – a total of 14 new IEEE standards working groups have started work on drafting so called *human* standards (Box 1).

// start Box 1

Box 1: IEEE P7000 series human standards in development

P7000 – Model Process for Addressing Ethical Concerns During System Design
P7001 – Transparency of Autonomous Systems
P7002 – Data Privacy Process
P7003 – Algorithmic Bias Considerations
P7004 – Standard for Child and Student Data Governance
P7005 – Standard for Transparent Employer Data Governance
P7006 – Standard for Personal Data Artificial Intelligence (AI) Agent
P7007 – Ontological Standard for Ethically Driven Robotics and Automation Systems
P7008 – Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
P7009 – Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
P7010 – Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems
P7011 – Standard for the Process of Identifying and Rating the Trustworthiness of News Sources
P7012 – Standard for Machine Readable Personal Privacy Terms
P7013 – Inclusion and Application Standards for Automated Facial Analysis Technology

// end Box 1

The importance of transparency and explainability

Consider P7001 as a case study. One of the general principles[8] of EAD asks "how can we ensure that autonomous and intelligent systems are transparent?", and recommends a new standard for transparency. P7001 *Transparency in Autonomous Systems* was initiated as a direct response. IEEE P7001 directly addresses the straightforward ethical principle that it should always be possible to find out why an autonomous system made a particular decision.

A robot or AI is transparent if it is possible to find out why it behaves in a certain way. We might for instance want to discover why it made a particular decision, especially if that decision caused an accident – or for the less serious reason that the robot or

AI's behaviour is puzzling. Transparency is not intrinsic to robots and AIs, but must be designed for, and it is a property which autonomous systems might have more or less of. And full transparency might be very challenging to provide, for instance in systems based on artificial neural networks (deep learning systems), or systems that are continually learning.

There are two reasons transparency is so important.

First, because modern robots and AIs are designed to work with or alongside humans, who need to be able to understand what they are doing and why. If we take an assisted living robot as an example transparency (or to be precise, *explainability*) means the user can understand what the robot might do in different circumstances. An elderly person might be very unsure about robots, so it is important that her robot is helpful, predictable – never does anything that frightens her – and above all safe. It should be easy for her to learn what the robot does and why, in different circumstances. An explainer system that allows her to ask the robot "why did you just do that?" and receive a simple natural language explanation would be very helpful in providing this kind of transparency. A higher level of transparency would be the ability to ask questions like "what would you do if I fell down?" or "what would you do if I forget to take my medicine?" This allows her to build a mental model of how the robot will behave in different situations.

And second, because robots and AIs can and do go wrong. If physical robots go wrong they can cause physical harm or injury. Real world trials of driverless cars have already resulted in several fatalities[9]**.** Even a software AI can cause harm. A medical diagnosis AI might, for instance, give the wrong diagnosis, or a biased credit scoring AI might cause someone's loan application to be wrongly rejected. Without transparency, discovering what went wrong is extremely difficult and may – in some cases – be impossible. The ability to find out what went wrong and why is not only important to accident investigators, it might also be important to establish who is responsible, for insurance purposes, or in a court of law. And following high profile accidents wider society needs the reassurance of knowing that problems have been found and fixed.

Transparency and explainability measured

But transparency is not one thing. Clearly an elderly relative does not require the same level of understanding of a care robot as the engineer who repairs it. The P7001 working group has defined five distinct groups of stakeholders (the beneficiaries of the standard): users, safety certifiers or agencies, accident investigators, lawyers or expert witness, and the wider public. For each of these stakeholder groups, P7001 is setting out measurable, testable levels of transparency so that autonomous systems can be objectively assessed and levels of compliance determined, in a range that defines minimum levels up to the highest achievable standards of transparency.

Of course, the way in which transparency is provided is very different for each group. Safety certification agencies need access to technical details of how the system works, together with verified test results. Accident investigators will need access to data logs of exactly what happened prior to and during an accident, most likely provided by something akin to an aircraft flight data recorder[10]. Lawyers and expert witnesses will need access to the reports of safety certifiers and accident investigators, along with evidence of the developer or manufacturer's quality management processes. And wider society needs accessible documentary-type science communication to explain autonomous systems and how they work. P7001 will provide system designers with a toolkit for self-assessing transparency, and recommendations for how to achieve greater transparency and explainability.

Outlook

How might these new ethical standards be applied when, like most standards, they are voluntary? First, standards which relate to safety (and especially safety-critical systems), can be mandated by licensing authorities, so that compliance with those standards becomes a *de facto* requirement of obtaining a licence to operate that system; for the P7000 series candidates might include P7001 and P7009. Second, in a competitive market, compliance with ethical standards can be used to gain market advantage – especially among ethically aware consumers. Third, there is growing pressure from professional bodies for their members to behave ethically. Emerging professional codes of ethical conduct such as the recently published ACM[11] and IEEE[12] codes of ethics and professional conduct are very encouraging; in turn, those professionals are increasingly likely to exert internal pressure on their employers to adopt ethical standards. And fourth, soft governance plays an important role in the adoption of new 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. For data- or privacy-critical applications, a number of the P7000 standards (P7002/3/4/5/12 and 13, for instance) could find application this way.

While some argue over the pace and level of impact of robotics and AI (on jobs, for instance), most agree that increasingly capable intelligent systems create significant ethical challenges, as well as great promise. This new generation of ethical standards takes a powerful first step toward addressing those challenges. Standards, like open science[13], are a trust technology. Without ethical standards, it is hard to see how robots and AIs will be trusted and widely accepted, and without that acceptance their great promise will not be realised.

Alan Winfield is Professor of Robot Ethics at the Bristol Robotics Laboratory, UWE Bristol, and visiting professor at the University of York. He chairs IEEE Standards Working Group P7001.
e-mail: Alan.Winfield@brl.ac.uk

The views expressed in this article are those of the author only, and do not represent the opinions of any organisation mentioned, or with which I am affiliated.

References

1. http://alanwinfield.blogspot.com/2017/12/a-round-up-of-robotics-and-ai-ethics.html
2. Winfield, A. F. & Jirotka, M. *Phil. Trans. R. Soc. A* **376**, 20180085 (2018); http://dx.doi.org/10.1098/rsta.2018.0085
3. British Standards Institute, BS 8611:2016 *Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems* (2016) https://shop.bsigroup.com/ProductDetail/?pid=000000000030320089
4. EPSRC, Principles of Robotics (2011) https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/
5. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html
6. http://standards.ieee.org/develop/indconn/ec/ec_about_us.pdf
7. IEEE Standards Association, Ethically Aligned Design (2017) https://ethicsinaction.ieee.org/
8. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_general_principles_v2.pdf
9. Stilgoe J, Winfield A, The Guardian, 13 April 2018 https://www.theguardian.com/science/political-science/2018/apr/13/self-driving-car-companies-should-not-be-allowed-to-investigate-their-own-crashes
10. Winfield A.F., Jirotka M. (2017) The Case for an Ethical Black Box. In: Gao Y., Fallah S., Jin Y., Lekakou C. (eds) Lecture Notes in Computer Science, vol 10454. Springer, Cham.
11. https://www.acm.org/code-of-ethics
12. https://www.ieee.org/about/corporate/governance/p7-8.html
13. Grand, A., Wilkinson, C., Bultitude, K. & Winfield, A.F. Open Science: A new ‘trust technology’? *Science Communication* **34**, 679- 689 (2012).