Standardizing Ethical Design Considerations for Artificial Intelligence and Autonomous Systems

Joanna Bryson and Alan Winfield

***Abstract*— AI is here now, available to and extending the powers of anyone with access to digital technology and the Internet. But its consequences for our social order are not only not understood, but barely even yet the subject of study. How can we guide the way technology is changing society? Since 2015, the IEEE has been developing principles for ethical *design* for intelligent and autonomous systems.**

I. Introduction

FOR decades—even prior to its inception—artificial intelligence (AI) has aroused both fear and excitement as humanity has contemplated creating machines like ourselves. Unfortunately, the misconception that 'intelligent' artefacts should necessarily be *human-like* has largely blinded society to the fact that we have been achieving AI for some time. While breakthroughs in surpassing human ability grab the headlines (think of Watson or AlphaGo), AI has been a standard part of the industrial repertoire since at least the 1980s, when expert systems began to be used to check circuit design.

Machine learning (ML) strategies for generating AI have also long been used, such as genetic algorithms for finding solutions to intractable computational problems such as scheduling, or neural networks not only to model and understand human learning, but also for basic industrial control, monitoring, and classification. In the 1990s probabilistic and Bayesian methods revolutionized machine learning and opened the door to one of the most pervasive AI technologies now available: searching through massive troves of data. The capacity to use AI to find information in texts has been extended by innovations in AI and ML algorithms that allow us to search photographs, and both recorded and live video and audio. We can translate, transcribe, read lips, read emotions (including lying), forge signatures and other handwriting, and forge video. Yet, the downside of these benefits is ever present. As we write this, allegations are circulating that the outcomes of the recent US presidential election and UK referendum on EU membership were both influenced through the use of AI for detection and targeting of 'swing voters' via their public social media use.

The IEEE Computer Society is creating standards for responsible designers to design our brave new world, and ensure its benefit to humanity.

II. Defining AI

While the following definitions are not universally used, they are well established [4]. *Intelligence* is the capacity to do the right thing at the right time, in a context where doing nothing (or making no change in behaviour) would be worse. Intelligence then requires:

- the capacity to perceive *contexts* for action,
- the capacity to *act*, and
- the capacity to associate contexts to actions.

By this definition, plants are intelligent. They can perceive and respond towards the direction of light, for example. However, the conventional understanding of "intelligent" includes being *cognitive, i.e.,* able to learn new contexts, actions, and/or associations between these.

Artificial intelligence, by convention, is a term used to describe (typically digital) artefacts that demonstrate any of these capacities. So for example, machine vision, speech recognition, pattern recognition, and static production rule systems are all examples of AI, with algorithms that can be found in standard textbooks.

Robots are artefacts that sense and act in the physical world in real time. Again, by this definition a smart phone is a (domestic) robot. It has not only microphones but also a variety of proprioceptive sensors that allow it to know when its orientation is changing or it is falling, and it can act by contacting its user.

Autonomy is technically the capacity to act as an individual. For social animals like humans, autonomy is normally situated somewhere along a scale. For example, it is fully expected that a family, place of work, government, and other organisations may regularly have some impact on our actions. Similarly, a technical system able to sense the world and select an action specific to its present context is called 'autonomous' even though its actions will ultimately be determined by the designers that constructed its intelligence, and its operators.

III. Concerns about Domestic and Commercial AI

AI is core to some of the most successful companies in history, in terms of market capitalisation, and along with ICT more generally has revolutionised the ease with which people from all over the world can create, access and share knowledge.. Yet possible pitfalls of AI could have quite serious consequences.

A. Will AI outcompete us?

Some of the most sensational claims fear that, as artificial intelligence increases to the point that it surpasses human abilities, it may come to take control over our resources and outcompete our species, leading to human extinction. AI is already superhuman in many domains. We can already do arithmetic better, play chess and go better, transcribe speech better, read lips better, remember more things for longer, and indeed be faster and stronger with machines than unaided. However, these capacities have in no sense led to machine ambition.

B. Will AI undermine societal stability?

For centuries there have been significant concerns about the displacement of workers by technology. There is no question that new technologies do disrupt communities, families, and lives, but also that historically the majority of this disruption has been for the better. In general, lifespans are longer and infant mortality is lower than ever before, and these indicators are well associated with political stability. Nevertheless, we are currently experiencing a disruption that seems to be undermining political stability, termed political polarisation,> Polarisation has happened before (e.g. in the early 20thC) and is known to co-occur with wealth inequality, although the causality between these is unclear [3]. It is possible that new technologies play a role in increasing inequality and therefore polarisation, possibly by reducing costs that formerly supported economic diversity.

C. Will AI harm privacy, personal liberty and autonomy?

When we consider the impact of AI on individual behaviour, information technology itself clearly has a specific impact. There have long been periods of domestic spying which can be associated with everything from prejudice in opportunities to pogroms. However, information technology can facilitate knowledge gathering, since it now allows us to keep (and access) long-term records on anyone who produces storable data — for example, anyone with bills, contracts, or a credit history, not to mention public writing and social media use. With machine learning, this allows us to make predictions concerning individuals' behaviour and preferences, which in turn opens the possibilities of control or persecution.

IV. Can standards promote ethics in AI? ,

Standards are consensus-based agreed 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 the user, such as safety, security and reliability.

Few standards explicitly address ethics in robotics and AI. One that does is British Standard BS 8611:2016 Guide to the ethical design of robots and robotic systems [1]. Published in April 2016, it provides designers with a tool for undertaking an ethical risk assessment. At the heart of BS 8611 is a set of twenty 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 alongside suggestions on how such measures might be verified or validated.

The IEEE's Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, a program designed to bring together "multiple voices in the AI and Autonomous Systems (AS) communities" has as its mission

> To ensure every technologist is educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems. [2]

The first output from the initiative is version 1 of discussion document Ethically Aligned Design (EAD), published in December 2016 [2]. The work of 8 committees, it covers:

1. General Principles,
2. Embedding Values into Autonomous Intelligent Systems,
3. Methods to Guide Ethical Design and Design,
4. Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI),
5. Personal Data and Individual Access Control,
6. Reframing Autonomous Weapons Systems,
7. Economics/Humanitarian Issues and
8. Law.

EAD articulates a set of about 60 draft issues and recommendations. Each committee was asked to identify issues which might be addressed through a new standard.

To date four Standards Working Groups are drafting candidate standards to address an ethical concern articulated by one or more of the 8 committees outlined in the EAD document. The candidate standards are:

- P7000 – Model Process for Addressing Ethical Concerns During System Design[1]. P7000 aims to set out a Value-Based System Design methodology.
- P7001 – Transparency of Autonomous Systems (AS)[2]. (See detail below).
- P7002 – Data Privacy Process[3]. P7002 aims to set out one overall methodological approach that specifies practices to manage privacy issues.
- P7003 – Algorithmic Bias Considerations[4]. P7003 aims to specify methodologies for assurance of how negative bias in algorithms has been addressed and eliminated.

A. P7001 – Transparency of Autonomous Systems (AS)

We take as an example one standard effort, on which both the authors are participating. P7001 is based on the radical proposition that it should always be possible to find out why an AS made a particular decision.

[1]https://standards.ieee.org/develop/project/7000.html
[2]https://standards.ieee.org/develop/project/7001.html
[3]https://standards.ieee.org/develop/project/7002.html
[4]https://standards.ieee.org/develop/project/7003.html

Transparency is not one thing [cite the Theodorou et al paper, attached?]. Clearly an elderly person doesn't require the same level of understanding of her care robot as the engineer who repairs it. Nor would a patient expect the same appreciation of the reasons a medical diagnosis AI recommends a particular course of treatment as their doctor. We identify five categories of stakeholder, and argues that AS must be transparent to each in different ways and for different reasons. These stakeholders are: (1) users, (2) safety certification agencies, (3) accident investigators, (4) lawyers or expert witnesses and (5) wider society.

1) For users, transparency is important because it builds trust in the system, by providing a simple way for the user to understand what the system is doing and why.
2) For safety certification of an AS, transparency is important because it exposes the system's processes for independent certification against safety standards.
3) If accidents occur, AS will need to be transparent to an accident investigator; the internal process that led to the accident must be traceable.
4) Following an accident, lawyers or other expert witnesses who may be required to give evidence require transparency to inform their evidence. And
5) for disruptive technologies, such as driverless cars, a certain level of transparency to wider society is needed in order to build public confidence in the technology.

Of course the way in which transparency is provided is likely to be very different for each group. If we take a care robot as an example, transparency means the user can understand what the robot might do in different circumstances; if the robot should do anything unexpected she should be able to ask the robot 'why did you just do that?' and receive an intelligible reply.

Safety certification agencies will need access to technical details of how the AS 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 (and it should be illegal to operate an AS without such a system). And wider society would need accessible documentary-type science communication to explain an AS (such as a driverless car autopilot) and how it works.

In P7001, we aim to develop a standard that sets out measurable, testable levels of transparency in each of these categories (and perhaps new categories yet to be determined), so that Autonomous Systems can be objectively assessed and levels of compliance determined. It is our aim that P7001 will also articulate levels of transparency in a range that defines minimum levels up to the highest achievable standards of acceptance. The standard will provide designers of AS with a toolkit for self-assessing transparency, and recommendations for how to address shortcomings or transparency hazards.

V. Conclusion

In conclusion, the changes artificial intelligence and autonomous systems are bringing to the world are real, and already in progress. While we cannot say with certainty that the situation is in hand, the authors, as members of the global initiative, are optimistic that the right steps are being taken, and that the IEEE may be key to ensuring that Artificial Intelligence and Autonomous Systems benefit all of humanity.

Acknowledgments

The authors would like to thank John C. Havens for introducing us to the IEEE initiative, and this writing opportunity.

Author information:

Joanna Bryson , Reader (Associate Professor), Intelligent Systems Group, Department of Computer Science, University of Bath, Bath, BA2 7AY United Kingdom and affiliate of the Princeton Center of Technology Policy.

E-mail: jjb@alum.mit.edu

Alan Winfield, Professor of Robot Ethics, Bristol Robotics Laboratory, University of the West of England, Bristol BS16 1QY, UK; Visiting Professor, Department of Electronic Engineering, University of York, York YO10 5DD, UK.

Email: Alan.Winfield@uwe.ac.uk

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

[2] IEEE Standards Association (2016), Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.

[3] McCarty, N.M., Poole, K.T. and Rosenthal, H. (2006), Polarized America: the dance of ideology and unequal riches, MIT Press.

[4] Winston, P.H. (1984), Artificial Intelligence, Addison-Wesley.