Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems

Joanna Bryson, University of Bath

Alan Winfield, University of the West of England

AI is here now, available to anyone with access to digital technology and the Internet. But its consequences for our social order aren't well understood. How can we guide the way technology impacts society?

For decades—even prior to its inception—AI has aroused both fear and excitement as humanity has contemplated creating machines like ourselves. Unfortunately, the misconception that "intelligent" artifacts should necessarily be human-like has largely blinded society to the fact that we have been achieving AI for some time. Although AI that surpasses human ability grabs headlines (think of Watson, Deep Mind, or alphaGo), AI has been a standard part of the industrial repertoire since at least the 1980s, with expert systems checking circuit boards and credit card transactions.

Machine learning (ML) strategies for generating AI have also long been used, such as genetic algorithms for finding solutions to intractable computational problems like scheduling, and 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 ML and opened the door to one of the most pervasive AI abilities now available: searching through massive troves of data. Innovations in AI and ML algorithms have extended our capacity to find information in texts, allowing us to search photographs as well as 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

0018-9162/17/$33.00 © 2017 IEEE

outcomes of the recent US presidential election and UK referendum on EU membership were both influenced by the use of AI to detect and target "swing voters" via public social media. To address these and other concerns, the IEEE Computer Society Standards Activities Board is creating standards for responsible designers who will shape our brave new world and ensure AI's benefit to humanity.

DEFINING AI

Although the following definitions are not universally used, they're well-established.[1] *Intelligence* is the capacity to do the right thing at the right time, in a context where doing nothing (or making no change in behavior) 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 to the direction of light, for example. The more conventional understanding of "intelligent" includes being cognitive, that is, being able to learn new contexts and actions, and the associations between them.

AI, by convention, describes (typically digital) artifacts that demonstrate any of these capacities. So, for example, machine vision, speech recognition, pattern recognition, and static production systems are all examples of AI, with algorithms that can be found in standard AI textbooks.

Robots are artifacts that sense and act in the physical world in real time. By this definition, a smartphone is a (domestic) robot. It has not only microphones but also a variety of proprioceptive sensors that let it know when its orientation is changing or when it is falling.

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 family, workplace, government, and other organizations might regularly have some impact on our actions. Similarly, a technical system that can sense the world and select an action specific to its present context is called "autonomous" even though its actions are ultimately determined by the designers that constructed its intelligence and its operators.

CONCERNS ABOUT DOMESTIC AND COMMERCIAL AI

AI is core to some of the most successful companies in history in terms of market capitalization and, along with information and communications technology (ICT) more generally, has revolutionized the ease with which people from all over the world can create, access, and share knowledge. However, possible pitfalls of AI could have quite serious consequences. Here we briefly review some common concerns to see which are both realistic and specific to AI.

Will AI outcompete us?

Some of the most sensational fears are that, as AI increases to the point that it surpasses human abilities, it might take control over our resources and outcompete our species, leading to human extinction. AI is already superhuman in many domains. With machines, 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 than we are unaided. However, these capacities have in no sense led to machine ambition. Human memory has been outstripped by books for centuries—mere intelligence is no more of a direct threat than mere strength.

Will AI undermine societal stability?

For centuries, people have had significant concerns about the displacement of workers by technology. There is no question that new technologies disrupt communities, families, and lives, but 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 seeing a disruption that seems to be undermining political stability. This disturbance is termed *political polarization*, which seems to co-occur with inequality, although causality between these is unclear.[2] Polarization has happened before, for example, in the early 20th century, reaching its climax in World War I. New technologies could play a role in increasing inequality—and therefore polarization—by eliminating costs such as distance that formerly supported economic diversity. This time, AI and ICT might be the technologies changing the economic landscape.

Will AI harm privacy, personal liberty, and autonomy?

What really makes AI special is its relationship to information, especially personal information. Previous periods of domestic spying have been associated with everything from prejudice in opportunities to pogroms. However, AI and ICT can greatly facilitate such knowledge gathering. We are now able 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 ML, this data lets us make predictions concerning individuals' behavior and preferences, which in turn opens the possibilities of control or persecution.

CAN STANDARDS PROMOTE ETHICS IN AI?

Standards are 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.

> IEEE's Initiative for Ethical Considerations in Artificial Intelligence Systems has as its mission to "ensure every technologist is educated, trained, and empowered to prioritize ethical considerations."

Few standards explicitly address ethics in robotics and AI. One that does is British Standard (BS) 8611:2016, *Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems*.[3] Published in April 2016, it provides designers with a tool to assess ethical risk. At the heart of BS 8611:2016 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.

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."[4] The first output from the initiative is a discussion document called *Ethically Aligned Design* (EAD), version 1, published in December 2016.[4] The work of eight committees, it covers

- general 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,
- how to reframe autonomous weapons systems,
- economics and humanitarian issues, and
- law.

EAD articulates a set of about 60 draft issues and recommendations. Each committee was asked to identify issues that could be addressed through a new standard. Presently, four standards working groups are drafting candidate standards to address an ethical concern articulated by one or more of the eight committees outlined in the EAD document. The candidate standards are

- *P7000—Model Process for Addressing Ethical Concerns during System Design* (standards.ieee.org/develop/project/7000.html), which aims to establish a value-based system design methodology;
- *P7001—Transparency of Autonomous Systems* (standards.ieee.org/develop/project/7001.html), which we discuss below;
- *P7002—Data Privacy Process* (standards.ieee.org/develop/project/7002.html), which aims to create one overall methodological approach that specifies practices to manage privacy issues; and
- *P7003—Algorithmic Bias Considerations* (standards.ieee.org/develop/project/7003.html), which aims to specify methodologies to ensure that negative bias in algorithms is addressed and eliminated.

CASE STUDY: A STANDARD FOR TRANSPARENCY

P7001 is an effort in which both authors are involved. It is based on the radical proposition that it should always be possible to find out why an AS made a particular decision.

Transparency is not one thing. Clearly, elderly persons don't require the same level of understanding of their care robot as the engineer who repairs it. Nor would patients expect the same appreciation of the reasons a medical-diagnosis AI recommends a particular course of treatment as their doctor. The P7001 working group has identified five categories of stakeholder—users, safety certification agencies, accident investigators, lawyers or expert witnesses, and wider society—and proposes that ASs must be transparent to each in different ways and for different reasons.

- For users, transparency is important because it builds trust in the system by providing a simple way for users to understand what the system is doing and why.
- For AS safety certification, transparency is important because it exposes the system's processes for independent certification against safety standards.
- If accidents occur, an AS needs to be transparent to investigators; the internal process that

led to the accident must be traceable.
- Following an accident, lawyers or other expert witnesses who might be called on to give evidence require transparency to inform their evidence.
- Disruptive technologies, such as driverless cars, require a certain level of transparency for wider society to gain the public's confidence in the technology and to ensure that trust is deserved.

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 users can understand what the robot might do in different circumstances. If the robot does anything unexpected, they should be able to ask it "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. 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 we can assess an AS objectively and determine compliance. It is our aim that P7001 will also articulate transparency levels in a range that defines minimum levels up to the highest achievable standards of acceptance. The standard will provide AS designers with a toolkit for self-assessing transparency as well as recommendations for how to address shortcomings or transparency hazards.

The changes artificial intelligence and autonomous systems are bringing to the world are real, and already in progress. Although we cannot say with certainty that the situation is in hand, we as members of the global initiative are optimistic that the right steps are being taken and that IEEE will be key to ensuring that AI and ASs benefit all of humanity.

ACKNOWLEDGMENTS

We thank John C. Havens for introducing us to the IEEE initiative and forwarding Dr. Walrad's request for an article to us.

REFERENCES

1. P.H. Winston, *Artificial Intelligence*, Addison-Wesley, 1984.
2. N.M. McCarty, K.T. Poole, and H. Rosenthal, *Polarized America: The Dance of Ideology and Unequal Riches*, MIT Press, 2006.
3. *Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems*, BS 8611:2016, British Standards Inst., 2016; shop.bsigroup.com/ProductDetail?pid=000000000030320089.
4. *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems*, version 1, IEEE Standards Assoc., 2016; standards.ieee.org/develop/indconn/ec/ead_v1.pdf.

JOANNA BRYSON is an associate professor in the Department of Computer Science at the University of Bath and an affiliate of the Princeton Center of Technology Policy. Contact her at jjb@alum.mit.edu.

ALAN WINFIELD is a professor of robot ethics at the Bristol Robotics Laboratory in the University of the West of England and a visiting professor in the Department of Electronic Engineering at the University of York. Contact him at alan.winfield@uwe.ac.uk.