



Editor: Daniel Zeng,
University of Arizona and
Chinese Academy of Sciences,
zengdaniel@gmail.com

AI Ethics: Science Fiction Meets Technological Reality

Daniel Zeng, *University of Arizona and Chinese Academy of Sciences*

There's no doubt that AI is becoming a permanent component of general public consciousness: AI-enabled technologies are now commonplace in mainstream applications.

On mobile devices, intelligent personal assistants such as Apple's Siri, Microsoft's Cortana, and Google's Google Now are touted as a successful integration of a variety of AI technologies such as natural language processing, knowledge discovery, intelligent interface, and recommender systems. A new generation of cognitive systems, best represented by IBM's famed Watson system, is opening up new venues for commonsense reasoning, knowledge engineering, and task automation. Autonomous driving, based on a number of AI technologies such as robotics, planning, intelligent control, and multi-agent systems, promises huge benefits in road safety. The expected reduction in traffic congestion and the likely major changes in human driving behavior and mobility patterns, along with a potentially revolutionary impact on the auto industry, have fascinated technologists and the general public alike. In military contexts, autonomous logistical systems based on robotic vehicles of varying kinds, and robotic combat capabilities, are already a reality or not far from it.

The past years have also seen exciting developments in AI methodology itself. Take the example of deep learning. In just a few years, deep learning and the related end-to-end computing paradigm have blossomed into an active field of study with major success in a variety of applications, rekindling the interest in one of the old AI traditions concerning the development of human-inspired general problem-solving mechanisms.

It's safe to say that a new era of AI is coming!

Growing Interest in AI Ethics

As the interactions between AI-enabled technologies and users (individuals, organizations, governments, and society at large) intensify and deepen, the interest in the ethical aspects of AI and AI applications is rapidly growing. Major actions are being taken by AI technology providers and governments to deal with such issues. For instance, in early 2014, Google established an internal AI ethics board to ensure the proper application of AI technologies. We suspect that it's very likely that other major players in the AI industry will soon set up similar mechanisms. Multiple countries are deliberating legal guidelines governing the use of driverless cars—for instance, in the US, legislation has been passed in several states and

We'd Like to Hear from You

Letters to the Editor:

Send letters, including a reference to the article in question, to bkirk@computer.org.

Letters will be edited for clarity and length.

Articles:

If you're interested in submitting an article for publication, see our author guidelines at www.computer.org/intelligent/author.htm.

**IEEE Computer Society
Publications Office**

10662 Los Vaqueros Circle, PO Box 3014
Los Alamitos, CA 90720-1314

Associate Manager, Editorial Services
Product Development

Brian Kirk
bkirk@computer.org

Editorial Management
Jenny Stout

Publications Coordinator
isystems@computer.org

Director, Products & Services
Evan Butterfield

Senior Manager, Editorial Services
Robin Baldwin

Digital Library Marketing Manager
Georgann Carter

Senior Business Development Manager
Sandra Brown

Senior Advertising Coordinator
Marian Anderson
manderson@computer.org

Submissions: For detailed instructions and formatting, see the author guidelines at www.computer.org/intelligent/author.htm or log onto *IEEE Intelligent Systems'* author center at Manuscript Central (www.computer.org/mc/intelligent/author.htm). Visit www.computer.org/intelligent for editorial guidelines.

Editorial: Unless otherwise stated, bylined articles, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *IEEE Intelligent Systems* does not necessarily constitute endorsement by the IEEE or the IEEE Computer Society. All submissions are subject to editing for style, clarity, and length.

Washington, DC, to allow their use, and in Germany, a committee consisting of researchers, practitioners from the auto industry, and the government has recently formed to draft a legal framework to regulate such vehicles.

Research interest in AI ethics (broadly construed) is growing, but ethical studies of AI aren't actually new. Before the establishment of AI as an academic discipline, philosophers, science fiction writers, and futurists were already discussing the ethical ramifications of machine intelligence. Such discussions are the precursors of AI ethics studies. In today's technological backdrop, old and new AI ethical discussion threads alike, to a large extent, are no longer fantasy pastimes. The legal, policy, and societal relevance and significance of AI ethics are now almost palpable.

Major Discussion Threads

Here, I summarize the major discussion threads concerning AI ethics. Needless to say, this summary isn't meant to be comprehensive. Rather, I'm aiming at the ongoing hotspots of meaningful discussions that might trigger further research, legislative actions, or broader societal repercussions.

Discussion Thread 1: "Technological Singularity and AI Doomsday Scenarios"

From AI's earliest days, philosophers, technologists, futurists, and the general public have been fascinated and, in some cases, frightened by the potential of AI and its possibly drastic impact on society and human civilization itself. A stream of science fiction movies and TV series dramatize various viewpoints, ranging from AI doomsday scenarios to AI/robots helping humanity prevail in difficult situations or saving civilization from the dark side of humanity itself. Many important figures have jumped into this discussion as well. Stephen

Hawking is a leading voice cautioning about the potential threats posed by AI, famously saying, "The development of full artificial intelligence could spell the end of the human race." Technologist Elon Musk posits that the development of AI is like "summoning the demon." Bill Gates has also joined them, suggesting that AI could be an existential threat to humans.

At the other end of the spectrum, several influential thinkers propose the idea that AI could culminate in human immortality and become a phenomenon at the cosmological scale redefining consciousness and matter. *Technological singularity* has been used to refer to what might happen as artificial intelligence exceeds human intellectual capacity and control.

The following statement, from Eric Horvitz, director of Microsoft Research Lab (and a member of this publication's advisory board), and his colleagues, represents a well-balanced and pragmatic viewpoint: "AI doomsday scenarios belong more in the realm of science fiction than science fact. However, we still have a great deal of work to do to address the concerns and risks afoot with our growing reliance on AI systems."

Discussion Thread 2: "Impact of Automation on Economy and Employment"

There's no doubt that the application of AI leads to increased productivity in specific areas of the economy. Yet, unintended adverse effects could occur. Take, for example, high-frequency or algorithmic trading, which typically involves AI elements. Put in the limelight by the US equity market flash crash in May 2010, automated trading practices have triggered a lively debate. One school of thought is that algorithmic trading has improved the overall quality of markets through lowered trading costs, more liquid markets, and reduced discrepancies

in prices across related markets. The counterargument concentrates on a mirage of negative impact: creating something resembling a single point of failure, benefiting the very few at the expense of the very many, and trading ahead of market orders to the detriment of long-term investors.

At the macro level, the total return on AI investment needs to be carefully weighed, especially when considering potential downsides such as increased inequality and unemployment: increased automation could push income distribution further toward the winners-take-all scenario. However, although traditional economic agents—the providers of cheap labor and “ordinary” capitalists—might be increasingly squeezed by automation, a new group of people who can innovate, design, and develop new products, services, and business models might emerge as winners in the new knowledge economy.

Discussion Thread 3: “Legal Ramifications and Accountability”

The current legal system, as well as social norms and ethical standards governing accountability and credit/blame assignment, are made for humans and relatively low-tech technological underpinnings. AI-enabled hardware and software systems, as they’re embedded in the modern-day societal fabric, are starting to challenge today’s legal and ethical systems.

For instance, in the current legal system, intent determines whether an act is considered a crime. As autonomous, intelligent, and adaptive AI entities roam across both the physical and cyber worlds, how will we define crime and intent? In particular, the era of driverless cars will emerge in the not-so-distant future, raising new questions in liability law. How can legislative bodies, jointly with technology providers and consumers, and society

at large, design new legal frameworks that encourage the realization of the safety benefits in autonomous driving technology? How much legal responsibility and protection should the driverless car manufacturers and designers enjoy? In the medical domain, who will be responsible for accidents or errors made by robotic surgical systems or automatic diagnostic systems?

Isaac Asimov’s classical work on robots raised many interesting questions about the responsibilities of robots and how to interrogate the ones that kill their human masters. In the futuristic but entirely conceivable robot era, can a robot be put on a trial? How much responsibility should the robot’s designer assume? Can a robot be called on to testify against itself?

Discussion Thread 4: “Privacy Considerations and Human Rights”

Intelligent softbots and assistive robots are already collecting and processing a lot of information from their human users in an array of task environments. Privacy concerns abound in such cases and have attracted a lot of attention. For instance, Ryan Calo wrote a chapter in *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press, 2011), discussing three aspects of privacy: direct surveillance, increased access, and social meaning.

The major implications of robotics on human rights are becoming a reality. According to a *Guardian* report published on 20 June 2014, the Skunk Riot Control Copter, a drone armed with plastic bullets and pepper spray built by a South African company, has been sold to an international mining company interested in using it to suppress labor riots. “Shaking the Foundations: The Human Rights Implications of Killer Robots,” a report released by Human Rights Watch and Harvard Law School’s International Human Rights Clinic on 12 May 2014, finds

fully autonomous weaponry could pose major challenges to fundamental human rights. Such concerns have also been discussed in high-profile international events—for instance, the first multilateral talks on killer robots opened at the UN in May 2014.

Discussion Thread 5: “Human-AI Relations”

Although humanoid robots are still a work in progress, it’s not inconceivable that such robots or AI machines assuming other outward forms will interact with humans holistically in the future, including at the emotional level. As such interaction intensifies and grows, interesting ethical considerations concerning human-AI relations will arise. Can robots really understand our feelings? Do we have feelings for robots?

In recent years, a growing number of robots have been developed to care for or accompany children, the elderly, and the disabled. Without a doubt, these robots will provide a lot of benefits. However, it’s entirely possible that their users could develop various forms of attachments to these robots. Such attachments could lead to interesting and challenging relational and ethical issues, warranting further research.

Discussion Thread 6: “Robot Rights”

Humans have human rights. Animals have animal rights. Should robots have robot rights? Should they be treated as conscious beings? Does the First Amendment protect robots’ speech? Do we need laws to protect robots from being abused?

It’s still too early to formulate specific answers to these “robot rights” questions given the current state of the art of robotics. Nonetheless, these questions are already drawing significant attention from the general public. In February 2015, heated discussions erupted in both mainstream and social

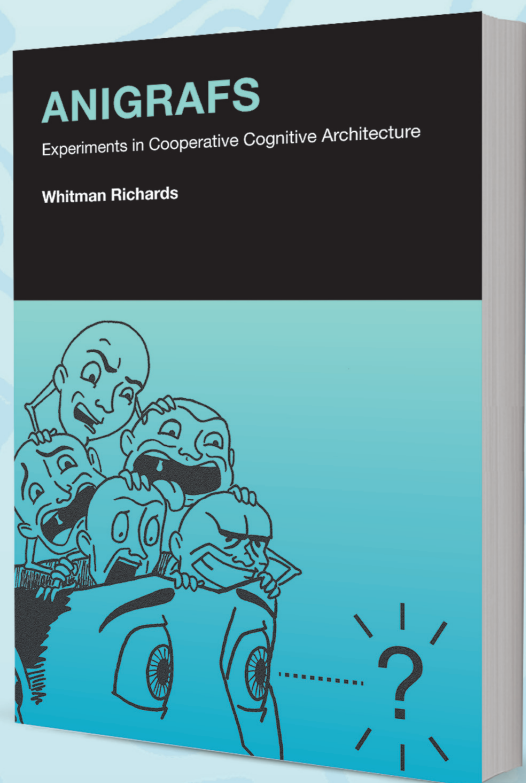
media about a video showing Boston Dynamics employees kicking a robot dog to show its robustness. A widely held opinion is that although this dog isn't a real, living thing, kicking it is inappropriate and can be construed as a violent behavior. A prevailing thought is that although we don't necessarily need to care about robot rights, abusing robots (just as abusing animals) very likely will make people more abusive toward other people. As such, protecting robot rights indirectly protects human rights. The emerging field of *robotics ethics*, short for robotics ethics, is concerned with these and other issues, such as how "humans design, construct, use and treat robots and other artificially intelligent beings."

Discussions about AI ethics are fascinating, thought-provoking, forward-looking, and in no small part entertaining, but one thing is clear: AI ethics is no longer a purely philosophical topic. Nor does it still fall under the realm of creative thinking by science fiction writers and futurists. Because of the increasingly deeper and wider adoption of AI technologies, AI ethical issues are becoming much more real, with practical relevance and significance. This emerging field calls for rigorous research and collaborative efforts from disciplines as disparate as philosophy, law, anthropology, economics, politics, computer security, and, of course, different branches of AI itself.

Most people with more than a cursory interest in AI are familiar with Asimov's famed three laws of robotics. Against the backdrop of a newly kindled interest in AI ethics, several researchers in Europe argued a few years ago that an extended set of ethical rules are needed for building robots. Five laws were put forth:

- "Robots should not be designed solely or primarily to kill or harm humans."
- "Humans, not robots, are responsible agents. Robots are tools designed to achieve human goals."
- "Robots should be designed in ways that assure their safety and security."
- "Robots are artifacts; they should not be designed to exploit vulnerable users by evoking an emotional response or dependency. It should always be possible to tell a robot from a human."
- "It should always be possible to find out who is legally responsible for a robot."

Of course, we can always argue whether these laws or certain components of them make sense or not. But as to direction, such potentially actionable guidelines and shared understandings might be what the AI community needs to further push the technological envelop while remaining mindful about our humanistic and societal bases. ■



ANIGRAFS

Explorations of a Collaborative Cognitive Architecture

Whitman A. Richards

An innovative proposal for understanding how mental organisms make decisions and control behavior.

128 pp., 7 color illus., 49 b&w illus., \$25 paper

INTELLIGENCE EMERGING

Adaptivity and Search in Evolving Neural Systems

Keith L. Downing

An investigation of intelligence as an emergent phenomenon, integrating the perspectives of evolutionary biology, neuroscience, and artificial intelligence.

508 pp., 196 illus., \$50 cloth

mitpress.mit.edu