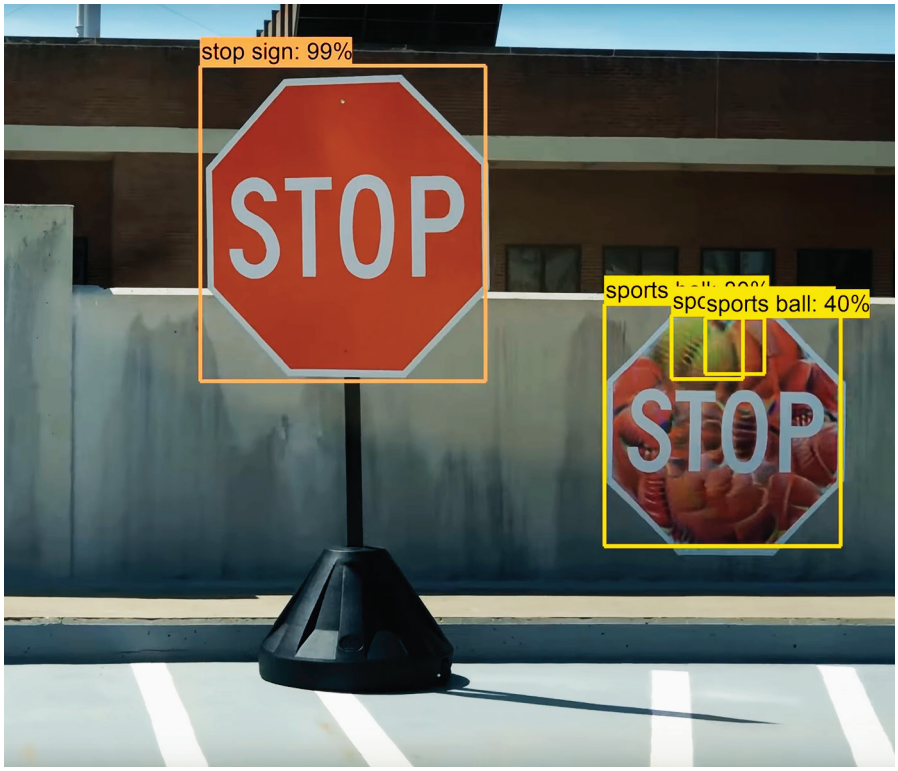
DOI:10.1145/3379920

Peter J. Denning and Dorothy E. Denning

The Profession of IT
Dilemmas of Artificial Intelligence

Artificial intelligence has confronted us with a raft of dilemmas that challenge us to decide what values are important in our designs.

Many speakers have pointed to various challenging ethical and design dilemmas raised by AI technology—we will describe 10 of the most prominent ones in this column. The first few are mostly technical; they arise from seemingly impenetrable complexity of the new technology. The final few ethical and design dilemmas include strong social dimensions; they arise from the difficulty of resolving emotional value conflicts to everyone's satisfaction.



An example of the stop-sign fragility problem: Will a driverless car's road-sign recognizer correctly see a stop sign?

SCREEN CAPTURE OF THE STOP-FRAGILITY PROBLEM COURTESY OF POLO CLUB OF DATA SCIENCE, GEORGIA TECH/YOUTUBE. BASED ON PAPER ARXIV.ORG/ABS/1804.05810

Explainability

The most common AI technology is the artificial neural network (ANN). An ANN consists of many layers of artificial neurons interconnected via weighted links. ANNs are not programmed in the conventional way by specifying the steps of an algorithm. Instead they are trained by showing them large numbers of examples of input-output pairs and adjusting their internal connection weights so that every input gives a correct output. The matrix of connection weights can amount to several gigabytes of storage. In effect, an ANN encodes the training examples of a function in its connection matrix and extrapolates them to estimate the outputs for data not in the training examples.

What happens if the human operator wants to know why the network generated an unexpected or erroneous output? In a conventional program, the operator would locate the code segment responsible for the output and if necessary repair it. In a neural network, the operator sees no algorithmic steps, just an unintelligible gigabyte size matrix of weights. How the weights relate to the unexpected output is totally opaque. It is a hot research area to find ways to augment neural networks so that their outputs can be explained.

Fragility

Neural networks can be quite sensitive to small changes in their inputs. For example, changing a few pixels of a trained input image can cause the output to change significantly even though the human operator cannot see a difference in the image. This leads to uncertainty in whether to trust a neural network when it is presented with new data on which it was

not trained. For example, when shown a new photo of a person's face, will it identify it as that person or someone else? Will a road sign recognizer in a driverless car correctly see a stop sign, and stop?

The sensitivity to small input changes is a vulnerability. A new subfield, "adversarial AI," has sprung up to find defenses against an adversary seeking to cause a neural network to malfunction. In one famous experiment, a road-sign recognizer was confused by an image of a stop sign on which small squares of masking tape were applied at strategic locations; instead of saying "stop sign" the network said "speed limit sign." In the current state of the art, it appears small changes in sensor outputs that feed a neural network can produce significantly wrong outputs. What looks to a human like a small continuous change to the input looks to the network as a discontinuous jump to a new state.

Fragility can also be seen when comparing neural networks. Suppose two neural networks are each trained from a different training set taken as a sample from a larger population. By all standard measures the two training sets are fair representatives of the population. When this is tried in practice, the two networks can respond with different outputs when shown the same input. Statistically minor changes in the training data can result in major changes of the output.

Researchers are looking for improved methods to measure the sensitivity of neural networks to small changes in their inputs, and ways to ensure a small input change results only in a small output change.

Bias

This is an issue that arises with the training data of neural networks. A bias in the training data can skew outputs. Many people are concerned about police use of neural networks trained by faces of predominately white people that give wrong identifications of faces of people of color. The bias of the training data may be invisible to the people running the training algorithms and only becomes visible in the results when the network is presented with untrained inputs.

It is as hot research area to find ways to augment neural networks so that their outputs can be captured.

The bias issue is further complicated by the fact that human beings are inherently biased. Each person has an individual way of interpreting the world that does not always agree with others. What appears as bias to one person may appear as fairness to another. What one person sees as the solution to a bias problem may appear as a new bias to another. This aspect of bias cannot be resolved within the technology by new statistical methods. It demands that humans respect each other's differences and negotiate solutions for conflicts.

Fakes

Tools for editing images, videos, and soundtracks are being combined with AI tools to produce convincing fakes.[1] They cannot be distinguished from real images, videos, or soundtracks without advanced equipment and forensic skills. These digital objects often contain biometric data of specific individuals, used for identification. How can we trust digital identifications when digitized forms of traditional identifications cannot be distinguished from fakes?

High Cost of Reliable Training Data

Neural networks require large training sets. Getting properly labeled data is time consuming and expensive. Consider the labor costs of a training scenario. Trained physicians must review colon images to identify suspicious polyps and label the images with their diagnoses. Suppose training a suspicious-polyp recognizer needs a million labeled images and a physician can diagnose and label an image in six minutes. Then 100,000 physician hours are needed to complete the labeling. If physicians were paid $50 an hour for this job, the training set would cost $50 million.

Training is also energy-intensive: a training that takes several days is as computationally intensive as bitcoin mining.

This means good quality training sets are hard to come by.

To keep the costs down there is a lot of interest in open source training sets. Users of these training sets are right to be concerned over the quality of the data because the persons contributing might be low-wage amateurs rather than well-paid professionals. There are reports of exactly this happening in open datasets that are then used to train medical diagnosis networks.

So even if developers are determined to avoid bias by getting large datasets, they will be expensive and right now it is difficult to determine their quality.

The big tech companies have a lot of reliable raw data about their users but are not sharing.

Military Uses of AI

Project Maven is a U.S. Pentagon project to use AI to give drones the power to distinguish between people and objects. Google was a partner and outsourced image differentiation to a company that used captchas to distinguish people from other objects. The gig workers looking at the captchas did not know they were teaching an AI system for a military purpose. When 3,000 Google employees formally protested, saying Google should not be developing technologies of war, Google executives decided not to renew the Maven contract.

Aversion to research for the military has been a difficult issue in universities since the days of the U.S. Vietnam war. Most universities divested themselves of laboratories that researched such technologies. Most DOD contracts are with private companies that are not involved with universities. With the large influx of new graduates into the big tech companies, the same aversion is now showing up among employees of private companies. The dilemma is in how to balance the need for national defense with the desire of many employees to avoid contributing to war.

Weapons and Control

The military's interest in AI to distinguish potential targets for drone attacks introduces another dilemma: Should a drone be allowed to deploy its weapon without an explicit command from a human operator? If AI is used in any weapons system, should a human have the final say in whether a weapon is launched?

Looking to the future, AI may also facilitate the creation of inexpensive weapons of mass destruction. Stuart Russell, a computer science professor at UC Berkeley and an AI pioneer issued a dire warning about AI controlled drones being used as WMD.[2] He produced a video, "Slaughterbots," which presented a near-future scenario where swarms of cheap drones with on-board facial recognition and a deadly payload assassinate political opponents and perform other atrocities. A swarm of 25,000 drones could be as destructive as a small nuclear bomb at a tiny fraction of the price.

Russell worries not only about the destructive potential of current AI technology, but about even more destructive potential of advanced AI. He says the creation of a super-intelligent computer would be the most significant event in human history—and might well be its last.

Issac Asimov postulated the famous Three Laws of Robotics in 1950 but no one has found a way to enforce them in the design of robots. The dilemma is: Should we continue to work on developing general AI when we do not know if we can control it?

Employment and Jobs

There is widespread fear that AI-powered machines will automate many familiar office tasks and displace many jobs. This fear is not unique to AI technology. For hundreds of years, new technologies have stirred social unrest when workers felt threatened by loss of their jobs and livelihoods. The fear is heightened in the modern age by the accelerated pace of AI automation. A century ago, a technology change was a slow process that took a generation to be fully adopted. Today, a technology change can appear as an avalanche, sweeping away jobs, identities, and professions in just a few years. Although the historical record says the new technology is likely to produce more jobs in the long run than it displaces, the new jobs require new skill sets the displaced workers do not have. The appearance of new jobs does not help the displaced.

Should we work on developing general AI when we do not know if we can control it?

One solution to this is regional training centers that help displaced workers move into the new professions. Unfortunately, the investment in such centers is currently limited.

Another proposed solution is the Universal Base Income (UBI), which would give every adult a monthly stipend to make up for income lost to automation. This proposal is very controversial.

Surveillance Capitalism

Surveillance capitalism is a term coined by Shoshana Zubhoff to describe a new phenomenon arising in the commercial space of the Internet.[3] The issue is that most online services capture voluminous data about user actions, which the service provider then sells to advertisers. The advertisers then use AI to target ads and tempt individuals into purchases they find difficult to resist. They also use AI to selectively customize information to individuals to manipulate their behavior such as their thinking about political candidates or causes.

The phenomenon is spreading to app developers as well. Their apps are Internet connected and provide data from mobile device sensors. A growing number are opting for "X as a service," meaning function X is no longer provided as installable software, but is instead a subscription service. In addition to a steady stream of monetizable personal data, this strategy provides a steady stream of income from subscribers.

Many of these services and apps are so attractive and convenient the tide to adopt them will not soon reverse. The dilemma for app developers is to find a way that provides the service without compromising individual user control over their data. The dilemma for citizens is how to effectively resist the trend to monetize their personal data and manipulate their behavior.

Decision Making

Dilemmas arise around machines that make decisions in lieu of humans. Consider the self-driving car when the sensors indicate "pedestrian ahead." How does the car decide between applying the brakes abruptly and potentially harming the occupant, or applying the brakes moderately and potentially hitting the pedestrian? Or, should the car swerve into the car alongside or drive off a cliff? Or do we hand control to the human and let that person choose an alternative? More generally, do we want machines to only make recommendations or machines that make and act on decisions autonomously? Is it even possible for machines to "act ethically"? Or is that something only humans can do?

Conclusion

None of these dilemmas is easily resolved. Many can be couched as ethical dilemmas that no professional code of ethics has been able to answer. Some of these dilemmas make obeying Asimov's First Law impossible: no matter what action is taken (or not taken), a human will get hurt. Software developers face major challenges in finding designs that resolve them.

References

1. Farid, H. *Fake Photos*. MIT Press Essential Knowledge Series, 2019.
2. Russell, S.R. *Human Compatible: AI and the Problem of Control*. Allen Lane of Penguin Books, Random House, UK, 2019.
3. Zuboff, S. *The Age of Surveillance Capitalism*. Public Affairs Books, 2019.

For more on autonomous vehicles, see Awad et al. on p. 48.

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for information innovation at he Naval Postgraduate School in Monterey, CA, USA, is Editor of ACM *Ubiquity*, and is a past president of ACM. The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

Dorothy E. Denning (dedennin@nps.edu) is Distinguished Professor Emeritus of Defense Analysis at the Naval Postgraduate School in Monterey, CA, USA.

Copyright held by authors.