A conversation with Tom Mitchell on February 19, 2014

Participants

- Tom Mitchell — Professor at the School of Computer Science and Chair of the Machine Learning Department, Carnegie Mellon University
- Alexander Berger — Senior Research Analyst, GiveWell
- Jacob Steinhardt — Graduate student, Computer Science, Stanford University

Note: This set of notes was compiled by GiveWell and gives an overview of the major points made by Dr. Tom Mitchell.

Summary

GiveWell spoke with Dr. Tom Mitchell as part of its shallow investigation of potential social implications of artificial intelligence (AI) research. The topics of discussion included the importance of resolving issues surrounding privacy and data sharing, the potential impact of viruses, and the risks from very advanced artificial intelligence.

Privacy and data sharing

There is a huge amount of data available today, from records of our online behavior to readings from smartphone sensors, that could be used to help resolve important social problems. There are many potential uses for this data that are technologically feasible but that are not pursued due to a combination of privacy concerns and issues of data ownership; improving understanding of these issues and making better tradeoffs could create significant social value.

Data from mobile devices could be used to determine the most effective treatments for diseases, to understand the health impacts of lifestyle variation, to manage traffic, etc. For example, smartphones can record and store measurements of one's location over the last week. This data could be useful in helping prevent a pandemic if, when a patient is diagnosed with a particularly contagious disease, data from smartphones could be used to identify everyone who had been in extended contact with them over the past few days. Then, those individuals could be contacted and given instructions to help ensure that they were treated promptly if infected.

Privacy concerns

There are privacy concerns that prevent us from aggregating the data necessary for many socially beneficial projects. There has not been very much thoughtful or careful discourse about these issues amongst the general public, policymakers, or researchers. Improving these groups' understanding of the relevant issues might make beneficial applications of data more feasible.

For example, most people seem to have a poor understanding of the graded nature of privacy reductions, and instead think of possible tradeoffs in terms of a dichotomy between

“private” and “not private;” most people have a poor understanding of the benefits that could be achieved by modest reductions in privacy; and most people (including policymakers) don't realize the extent to which it is possible for technologies like privacy-preserving data mining to open up more favorable tradeoffs between privacy and social benefit. Overall, there is very little serious discussion of beneficial applications that might be made possible by effectively sharing more data.

There is a need for more informed decisions about how to navigate tradeoffs between privacy and social benefits. These decisions are the responsibility of policymakers and the public at large.

There is also an opportunity for work that makes it possible to achieve greater social benefits with the same levels of privacy or to achieve greater privacy without sacrificing the usefulness of data, such as further research in privacy-preserving data mining.

Data sharing

Ownership of data also inhibits many socially beneficial applications of data. Private corporations are typically motivated to share data only if they can expect to profit by doing so, and this is often not the case for socially valuable applications of data. In other cases, there are regulations that prevent data owners from sharing data at all (such as the Health Insurance Portability and Accountability Act (HIPAA), which prevents medical care providers from sharing data).

It is possible that these problems could be ameliorated by guidelines, regulations, or incentives provided by the government, and that a better understanding of the potential applications of data sharing would facilitate improvements.

Viruses

The potential impact of viruses is growing and is already quite large. Most people use mobile devices with a wide range of sensors and credentials, so a sophisticated virus that gains access to those devices could conceptually do a lot of damage: it could record, block, or make phone calls or emails, it could make financial transactions, it could record audio or video, it could track the user’s location, and so on. These capabilities might be used in combination; for example, a virus might impersonate an individual over email and then prevent them from being contacted so that they could not fix the problem. Progress in machine learning may lead to more intelligent behavior by viruses over time, which could increase their potential impact.

Moreover, there are clear incentives for unscrupulous individuals to design and improve viruses, so as more damaging viruses become possible it seems likely that some individuals will make them.

Risks from advanced AI

GiveWell asked about the threat posed by extremely intelligent AI systems to society at large, a concern which is raised (for example) in the "Ethics and Risks" section of Stuart Russell and Peter Norvig's textbook, Artificial Intelligence: A Modern Approach*.*

Risks from very advanced AI do not seem to be as serious as other challenges we currently face, such as limiting the impact of computer viruses and navigating tradeoffs concerning privacy and data sharing. Risks from advanced AI are probably not a significant concern over the coming decades (e.g. the next 20-40 years). Unlike the situation with respect to viruses, there are few incentives for researchers to develop an AI that might pose a threat to society more broadly, which decreases the likelihood of a serious incident.

In the very long run it's possible to imagine such a threat arising accidentally. For example, today some researchers study AI systems that interact with the world passively in order to learn language, such as Professor Mitchell's "Never-Ending Language Learner." There is interest in enabling such systems to be more active, perhaps by hosting webpages or sending emails. Today we can mostly understand what these systems can learn and why they make the decisions that they do, but it is possible to imagine them becoming more sophisticated in the future such that it is harder to understand exactly what they are doing or why. If a sophisticated system were for some reason motivated to actively make its behavior obscure, then it might have some risk of behaving in an unanticipated and harmful fashion.

It would make sense to think at least a little bit about these questions, and for AI researchers to have a better understanding of how they could or should respond to possible problematic situations. At the moment there is little focused discussion of these issues, though Eric Horvitz organized a panel that discussed some of them in his role as President of the Association for the Advancement of Artificial Intelligence (AAAI).

People for GiveWell to talk to

- **Eric Horvitz,** past President of AAAI, convened an AAAI Presidential Panel on Long-Term AI Futures.
- **Manuela Veloso**, current President of AAAI.
- **Thomas Dietterich**, Professor at Oregon State University.