A conversation with Tom Dietterich on April 28th, 2014

Participants

- Tom Dietterich – Distinguished Professor and Director of Intelligent Systems, School of Electrical Engineering and Computer Science, Oregon State University
- Alexander Berger – Senior Research Analyst, GiveWell

Note: This set of notes was compiled by GiveWell and gives an overview of the major points made by Professor Dietterich.

Summary

GiveWell spoke with Professor Dietterich about potential social implications of artificial intelligence (AI) research and potential ways to improve AI safety research.

Potential AI risks and benefits

The AI community is concerned about potential societal risks posed by advanced AI. However, just as the developers of jet propulsion may not be well-positioned to assess whether the technology would be used to break the land speed record or to build bombers, AI researchers may not be particularly well-suited to addressing questions of AI safety. Certainly contributions from other fields will also be needed.

AI-enabled crime

AI-enabled crime is one important risk in the short term. Intelligent malware that mutates over time has not yet appeared but is feasible. This is particularly worrying because those who would implement such malware probably would not care about its potentially wide-ranging negative effects. Applications of such AI include:

- Malware that could imitate users on, e.g., banking websites.
- Intelligent bots that could spread misinformation via social networks in order to cause panic or to manipulate public opinion, with the goal of affecting stock movements. Such misinformation campaigns could be very effective.
- AI-automated insider trading, which may already be occurring.

There is no AI law enforcement sub-community, though the US Department of Defense does pay some attention to such issues.

Autonomous weapons systems

It is feasible that autonomous weapons systems, such as intelligent drones, will be developed. Autonomous drones might be more consistent in following military rules of engagement than drones operated by humans. However, there is a risk of such technology falling into the hands of enemies or criminals, in which case, it would follow whatever rules they chose.

Several prominent researchers, including UC Berkeley Professor Stuart Russell and Noel Sharkey, have advocated that the AI community urge governments to ban the development and use of autonomous weapon technologies. However, detecting violations of such a ban could be difficult, and reliance on an invasive inspection regime might not be effective.

The Singularity and machine memory

Although he does not rule out the possibility entirely, Professor Dietterich does not see evidence that a Singularity-like chain reaction of exponential recursive improvement in machine intelligence is likely, though many writings on AI accept this as a premise.

Research is being done on "lifelong" machine learning wherein machines learn from their experiences at a steady rate. However, most intelligent systems cannot be run indefinitely because of their memory limits. These limits would prevent their capabilities from increasing exponentially. There has been very little research on machine "forgetting," or how to prevent a long-running intelligent agent from becoming overwhelmed by its own memories. Professor Dietterich says that very little current research is investigating this problem.

Professor Dietterich's current research focuses on machine systems that learn from their experiences, but these systems can only learn a small amount before halting.

Professor Dietterich was involved in the CALO (Cognitive Assistant that Learns and Organizes) project funded by the Defense Advanced Research Projects Agency (DARPA) during the 2000s. CALO could only run for about a week before filling its memory disk drives. Several researchers from the CALO project went on to develop Apple's Siri personal assistant. Siri clearly poses no danger of runaway recursive increase in capability.

Benefits of automated decision-making

Automated decision-making is ideal for problems that need to be solved quickly and that machines can solve better than humans. There is less benefit to giving autonomy to a machine for decisions that are made at a slower pace.

It may not be desirable for factory robots to think autonomously. It is more efficient to build non-autonomous machines that have a fixed set of goals that they work to achieve.

Safety in autonomous systems

It is important to research ways of placing appropriate limits on machine autonomy. For example, mixed autonomy systems are designed to function autonomously but consult human operators when making difficult decisions. The primary challenge with such

machines is ensuring that they can tell when a given decision is difficult in the relevant sense.

The military has funded some research on mixed autonomy machines. Existing drone weapon systems currently require several pilots and support staff per machine. The military would like to have one pilot be able to control several machines, which might require mixed autonomy systems or some similar technical innovation.

Research by Oren Etzioni and Dan Weld at the University of Washington has found that it would it be quite difficult to implement something like Isaac Asimov's three laws of robotics due to the difficulty of deciding whether a law applies to a particular action (e.g., will the action contribute directly or indirectly to injuring a human being?). The AI community has shown some interest in safety research like the work of Etzioni and Weld on Asimov's laws, and it is considered a legitimate object of study. Funding agencies tend to be less interested in safety research than in improving the basic capabilities of AI systems.

Safety in robotics

Safety is a major concern in robotics, as errors in code can cause erratic and potentially dangerous behavior.

Developing robots that can preserve their own safety when exploring new environments is a major area of research. Professor Dietterich currently researches how to give machine intelligence systems models of their own competence, which he calls "competence modeling." In dynamic situations, such as flying, it is desirable for machines to be able to anticipate and avoid situations too complex for them to successfully navigate.

It is desirable for machine intelligence systems to be able to predict the consequences of their actions and formulate successful plans even when the information available is limited. This is currently much more challenging for robots than for humans, but people also often fail to anticipate the effects of their actions.

Implications of AI for the labor market

Machines have eliminated many middle management positions. More research on the effects of improving machine intelligence on the labor market would be beneficial. Economists and sociologists, rather than computer scientists, are likely best positioned to conduct such research.

Computer scientists are not always able to anticipate what role new computer technologies will play in society. For example, although computer scientists had developed the Internet protocol suite in the 1980s, they did not anticipate the huge expansion of the Internet during the 1990s or that the primary role of computers among the general public would be as a communication technology (e.g., email, social networks, etc.).

Artificial general intelligence

Though most research in the mainstream AI community focuses on building specific tools for specific applications, most AI researchers believe that building an Artificial General Intelligence (AGI) with many of the same decision-making and perceptual learning capabilities as humans is possible. The research currently being done on AGI may also provide insight into how the human mind works; similarly, research in human cognition and neurophysiology may provide guidance for constructing AI systems.

There may not be strong economic incentives for developing AGI. There is significant economic value in automating tasks that machines can perform better than humans, but Professor Dietterich sees much less economic incentive to automate tasks that humans already perform well.

Improving AI safety research

Professor Dietterich would be happy to see a foundation support meetings of experts from the fields of computer science, law, philosophy, and economics to discuss issues of AI safety. The Association for the Advancement of Artificial Intelligence (AAAI) would be an excellent organizer for such meetings.

In the near term, there is a need for discussion of public policy questions raised by developments such as self-driving cars and the use of AI in medical devices – for example, what regulations to place on self-driving cars and how to certify the safety of systems that learn and adapt.

Offering research grants would likely increase interest among academics to study safety questions, though evaluating research proposals might be difficult.

In the long run, it would be beneficial to interest the National Science Foundation (NSF) in funding AI safety research. This would ideally involve the collaborative support of both the Directorate for Computer and Information Science and Engineering (CISE) and the Directorate for Social, Behavioral and Economic Sciences (SBE). CISE has been a key funder of much AI research in the last 25 years. DARPA has also been a key funder in the last dozen years and previously, especially during the 1970s.

Dr. Dietterich was a member of the AAAI Presidential Panel on Long Term AI Futures, headed by Eric Horvitz. That meeting was interesting but did not produce many concrete results. Much more investigation is needed of these questions and issues.