A conversation with Stuart Russell on February 28, 2014

Participants

- Stuart Russell –Professor of Computer Science and Smith-Zadeh Professor in Engineering, University of California, Berkeley; Advisory Board Member, Centre for the Study of Existential Risk (CSER)
- 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 Stuart Russell.

Summary

GiveWell spoke with Stuart Russell as part of its investigation of potential social implications of artificial intelligence (AI) research. Conversation topics included: assessing risks from super-intelligent systems, attitudes in the AI research community, and logical next steps in research.

Incentives for AI development

Significant advances in AI, once a distant theoretical possibility, are increasingly likely. The economic incentives for improvements to AI are immense. Individual companies or researchers can gain billions of dollars for marginal improvements in AI algorithms. As more consumers and industries use AI techniques, the market pressure for advancements in AI will likely grow.

The development of AI may mirror the development of nuclear weapons in that, once the technology exists, there will be a race to improve and deploy it. In such situations, regulation may struggle to keep up.

Ethical issues related to AI

Some ethical issues related to AI are already being discussed. For example, there are difficult issues around modern autonomous weapon systems. While official US policy is that drone strikes are only to be permitted with the authorization of a human operator, the current design and deployment of drones could be modified easily to allow the operator to send the drone to acquire and eliminate human targets based on its programmed criteria.

The lack of a single, unified international organization for addressing responsible AI policy has made it harder to address the ethical questions posed by autonomous weapons earlier.

International support for banning autonomous weapons is significant; the United Nations, Human Rights Watch and the International Committee for Robot Arms Control (ICRAC) all support a ban. There is rising interest in more benign applications of drones, such as search & rescue systems, ambulances, or transport vehicles, but these would be harder to develop than fully autonomous weapons, which only require the combination of a few existing technologies. Because military thinkers are divided on the benefits of autonomous weapons, a meaningful international treaty limiting them may still be possible.

Assessing potential AI risks and benefits

The study of how powerful AI systems may impact society is still in an early stage. While the expected impacts are large and likely positive on balance, the likely variance in impacts is also very high. (That is, AI is likely to have major impacts, and the direction and magnitude of those impacts in terms of human welfare is quite uncertain.) While there has been some speculation, there has been little organized study of the risks from powerful AI systems:

- The Association for the Advancement of Artificial Intelligence (AAAI) convened a panel that addressed AI risks a few years ago, but its primary conclusion was that such technology was too far in the future to be of present concern.
- The US Air Force (USAF) has started a program to examine Testing & Evaluation (T&E) and Verification & Validation (V&V) in autonomous systems. USAF is likely to have a more narrow focus on the next generation of military craft and weapons systems.
- The Artificial General Intelligence Society (AGI) sometimes addresses the topic of risk, but only in broad brushstrokes.

While powerful intelligent systems could have global ramifications, there has been limited serious discussion of potential risks in the AI research community. Part of the reason for that is that there is no obvious framework in which to study these questions in a rigorous fashion.

Reasons to address AI risks now

The rationale for delaying consideration of risks is not as strong as it was five to ten years ago. Progress in many subareas of AI has accelerated rapidly and there have been advances in autonomous system designs. Due to recent technological development and the strong economic incentives to develop AI technology, AI risks require present-day attention.

Current research on AI risks

There is little research on AI risks now. The Centre for the Study of Existential Risk (CSER), affiliated with Cambridge, is trying to advance the discussion by moving questions about risk from the realm of abstract theorizing to an area of genuine technical inquiry, but their efforts are still at an early stage. They hope to engage more people and agencies in exploring and funding AI risk assessment.

Previous workshops on "human-level AI" have focused on feasibility of AI development rather than the variance of the potential impact. Professor Russell ran one workshop in the late 90s called "The Big Picture," which left him with the impression that most people in the field were not interested in focusing on the question of what long-term impacts major advances in AI would produce.

Barriers and challenges to understanding AI risks

Despite the sizable community of academics and professionals working on AI, there is little attention to long-term risks, for several reasons.

Perceived seriousness

For years, the idea of human society being harmed by highly advanced machines has been the focus of science fiction and a fringe part of the AI community (usually lacking strong technical credentials). While concern among other groups has risen during the last few years as the speed of new developments has increased, the problem of long-term AI risks is still viewed by some as a less than serious area of study.

Defensiveness

Experts in AI are deeply invested in the potential benefits that improved AI systems could provide. Claims that highly advanced systems would pose a risk are sometimes seen as an attack on the field motivated by technophobia rather than rational concern. This reduces the amount of energy that experts prefer to spend on exploring long-term risks.

Professor Russell believes that other researchers are mistaken about the tradeoffs involved. He believes that the potential benefits of advanced AI are enormous, but that if they are deployed recklessly, then none of that upside will be realized. Just as with nuclear power, such advanced technology will only be tolerated by society if it is seen as safe.

Opportunity costs

Most experts are working on research and problems that are most relevant to the near future. Giving attention to abstract long-term AI risks would divert time and attention away from ongoing projects.

Lack of cohesion

The international AI community is large but fragmented. There is no single international organization offering definitive guidelines for ethics and best practices in designing highly intelligent systems.

Challenges and goals in super-intelligent systems

A number of conceptual questions about powerful intelligent systems make it difficult to imagine their potential impact on society. Some of these questions could be addressed as part of a research agenda on AI risks.

Defining value

It is difficult to imagine how AI creators might be able to instill an AI with the ability to weigh options in an acceptable way. It is unclear how to encode a utility function that AI could use to determine the best outcome without the risk that it might make unacceptable decisions. Attempting to bypass the need to actually write the function by 'teaching' the machine through reinforcement learning may also be problematic. In order to teach through reinforcement, one must define a condition under which the reward signal (the indication that the system has done something good) is generated. The so-called "paperclip argument" described in Nick Bostrom's work applies to both approaches.

The only readily available model for preparing an intelligent system to behave somewhat rationally in the human context is child-rearing, a process that evolved over time and is far from perfect. Whether any of the same principles of child-rearing will be applicable to artificial systems is a very open question.

Intelligent agents vs. intelligent theorem-provers

There are different risks and benefits associated with intelligent agents (AI that can make decisions) and intelligent "theorem-provers" (AI that can only evaluate the validity of propositions). Intelligent agents carry greater potential risks and greater potential benefits. A super-intelligent theorem-prover would be a powerful tool but would still present some dangers. Most of those dangers would be bounded by human willingness to act on the

information the system provided. The key question would be what persons or organizations would be allowed to possess or use such systems.

Advanced AI systems would likely be much easier to police than engineered biological threats. Once the design principles for super-intelligent AI are understood, it will still require significant infrastructure and expertise to create one. Conversely, nearly anyone can splice genes into bacteria in a rudimentary lab. Modern-day restrictions on nuclear technology could act as a model for controlling intelligent theorem-provers.

Verification

A critical concept in addressing risk in designs for AI is verification. Theoretically, a sufficiently powerful theorem-prover should be able to examine a proposed design for a super-intelligent agent and determine the safety of the design. However, there is debate among experts about whether it would be possible to formally specify the desirable design parameters of such an agent.

Creating a framework to understand AI risk

The key obstacle to progress on AI risk research in the status quo is the lack of a technical framework in which to work. The best way to create a technical framework for research to understand and mitigate AI risk may be to arrange a one or two week retreat of about a dozen researchers. Following the retreat, it would likely take a few years to publish the framework and obtain some useful technical results. Then, it may be possible to move from research proposals to a fully funded research program by getting funding from government science funders (e.g. DARPA or the National Science Foundation). Funding from either DARPA or the NSF prior to the development of such a framework is conceivable but unlikely.

Others to talk to about these issues

- Eric Horvitz, Microsoft Research
- Bart Selman, Professor of Computer Science, Cornell University
- Murray Shanahan, Professor of Cognitive Robotics, Imperial College London
- Nick Bostrom, Professor of Philosophy, Oxford University