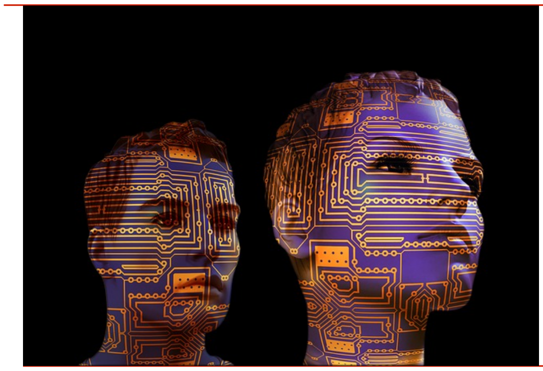
CUSPE Communications
Oct 2016



Advanced Artificial Intelligence: Policy and Strategy

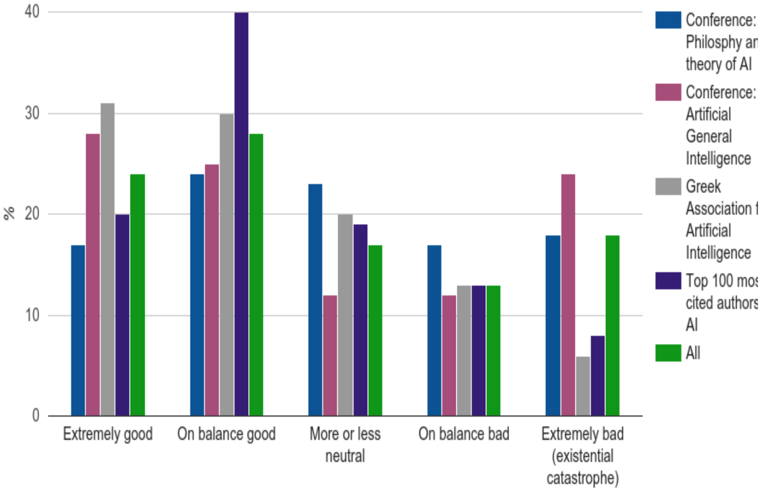
Elizabeth Barnes, Computer Laboratory, University of Cambridge

Artificial intelligence is complex to define and frequently misunderstood. Colloquially, the term usually tends to be used when computers make progress on a new type of task that was thought to require 'intelligent behaviour' and which previously only humans were capable of performing. A more general definition is a computer system that behaves as an 'agent' capable of taking in information from its environment and making decisions in order to achieve its programmed goals. In popular culture, AI is usually depicted as a humanoid and human-like robot, but AI can take many forms: from Netflix's recommendation algorithm to self-driving cars. Current AI systems are 'narrow' - only capable of carrying out specific tasks. In the future it may be possible to build artificial general intelligence or 'AGI': AI systems that can carry out most tasks as well as or better than humans.

AI technology has the potential to bring huge benefits to society. Current technology is already helping in applications such as extracting useful information from large amounts of data, allowing us to improve medical diagnosis for example. In general, increasingly intelligent AI will help us solve increasingly important and difficult problems. However, it is also possible that advanced AI could be highly destructive and even lead to human extinction (an 'existential threat') if we do not manage its development correctly, as discussed by an increasing number of experts in the field [1] [2]. Given the potential for an existential catastrophe, which entails not only the tragedy of billions of deaths but also the loss of all possible future flourishing, the stakes are extremely high. Currently, we do not know exactly when, how or even if AGI will be developed. What can be done at this stage to ensure that advanced AI is robustly beneficial?

Fig 1 - Estimated probability of different outcomes from human-level general intelligence

AI experts' expected outcomes of human-level general intelligence. These are preliminary surveys with some methodological and response rate issues and do not constitute a reliable prediction, but serve to illustrate the lack of consensus, with experts considering both extremely good and catastrophically bad outcomes quite likely.



Data from Bostrom and Müller [3]

Problem Description

It is useful to discuss a functional definition of general intelligence. Here we will define general intelligence as the ability to successfully achieve goals in a wide range of environments. It is uncertain where the upper bound on intelligence is; it could be far above humans. There are constraints on human abilities due to our biology and evolutionary history — such as limits on brain size imposed by the need to fit through the birth canal or the high calorie demands of supporting a large brain — that do not restrict non-biological systems. Although present technology is far from human-level general intelligence, in the long run we may be able to create highly competent AI systems that are far better than humans at achieving whatever goals we give them. On the one hand, this means that the AI would be able to resolve problems that are currently impossible for humans. On the other hand, it would be very difficult for a human to intervene to stop these goals being carried out once the system is running. In this scenario, the choice of the goals that the AI is programmed with is likely to strongly shape the future - for better or for worse. The question then is whether it is easy to choose goals that will be beneficial for humanity if we program a highly intelligent agent to carry them out. There are reasons to believe that this may be very hard.

One issue is that humans are highly adapted to predicting the behaviour of other humans, but we may find it much more difficult to reliably predict the behaviour of AI systems, especially given our strong tendency to anthropomorphise [4]. AI algorithms may respond unpredictably to new inputs, even if to a human the new data seems similar to the training data. As AI is put in control of more critical systems, this becomes an increasingly serious problem. There is already some evidence of surprising behaviour in Narrow AI systems, such as ‘flash crashes’ with trading algorithms, classifier algorithms that learnt to distinguish photos by the lighting rather than the objects that the humans noticed as the distinguishing feature, and a circuit design algorithm that ‘hacked’ its task by designing a circuit to amplify background signals from nearby computers rather than generating them itself as intended [5]. This means that a beneficial-seeming goal might be achieved in a surprising way, which may result in unforeseen harmful side effects.

One issue is that humans are highly adapted to predicting the behaviour of other humans, but we may find it much more difficult to reliably predict the behaviour of AI systems

Another possible problem with advanced AI is that seemingly innocuous goals might be harmful when taken to extremes by a highly competent system. This is because for any goal there exist sub-goals that are important for achieving any task, which include things like self-preservation and acquiring large amounts of resources [6]. So an AI system maximising the probability of achieving whatever goals it was programmed to carry out might start by ensuring it would not get shut down, acquiring large quantities of money, and concealing its actions from humans so they would not attempt to interfere with carrying out the goal. It is possible this could put AI systems in resource conflict with humans, even when we have carefully specified the system’s goals to be something useful. A simplistic example would be a system instructed to cure cancer as quickly as possible. Truly optimising only for this and disregarding any other considerations might lead, for example, to turning a large proportion of the planet into computers in order to model protein interactions and design a cure.

To avoid this sort of problem, it is necessary to specify all the things you do and don't want the system to do whilst carrying out its assigned tasks. However, because of the unpredictability problem described above, this is very hard to do 'case-by-case'. In the circuit design example, how likely is it that the programmers would have thought to specify 'Make this circuit produce this signal, but don't do it by turning the components into a radio receiver and amplifying background signals'? An AI that is intelligent enough to solve complex problems that humans cannot solve will come up with solutions that humans cannot anticipate.

A potential better solution is to set the system's goals to something like 'Do what we meant' or 'Do want we would have wanted'. However, specifying these sort of concepts in computer code is beyond our current capabilities. This is the heart of the AI safety research problem: how to set the goals of an advanced AI system such that it will not cause destruction in the process of achieving the goals we have specified.

It is important to note that none of the concerns discussed here presuppose that an AI will be sentient, will 'rebel' against its creators, or will have human-like emotions of anger, resentment or malice as usually depicted in fiction. Rather, the concern is that the system will do exactly what it is programmed to do.

There are other important issues around AI development, including how to manage the social and economic impact of automation, and how to safeguard increasingly weaponizable systems that are under AI control (such as cars or critical elements of infrastructure) from deliberate external manipulation. Although in this report I am focusing on longer-term implications of AI, the shorter-term issues are also very much deserving of attention.

Obviously, researchers want the technology they create to benefit society. However, there are generally stronger incentives for groups to develop increasingly competent AI systems than to do research into safe development. An AI catastrophe would impact people across the world and possibly even all future generations, making safety research a common good, but incentives for individuals or companies in the field are usually to produce AI systems that can perform some new function rather than doing speculative research into addressing some problem that may or may not arise in future. Safety research is therefore likely to be neglected due to tragedy of the commons effects.

Present technology is still a very long way from AGI, and we should not restrict current research, which will bring huge benefits in many different fields. However, although there is not a consensus that AI poses a serious danger in the long term, neither is there a consensus that it will be safe. Given the high stakes involved, we should investigate and take the potential risks seriously even at this early stage.

Policy Approaches

There are parallels with current or emerging technologies that have potential for great upsides and serious risks - for example synthetic biology, nuclear technology, nanotechnology, or geoengineering. We can learn from historical attempts to mitigate risks from emerging technologies, such as the 1975 Asilomar conference on recombinant DNA [5], or the Biological and Chemical weapons conventions [6]. Other areas also provide examples of what can go wrong. Current lack of public understanding of the risks and benefits of GM has lead to restriction of beneficial research and reduced uptake of beneficial technologies [6].

In this situation, when dealing with highly uncertain but potentially catastrophic risks, we cannot afford to proceed by trial and error but must use 'anticipatory policymaking'. This would be aided by the development of a set of technical milestones for future development with appropriate policy actions at each stage. Useful strategies for this include improving our ability to forecast technological progress and the outcome of different policy approaches, using new techniques for eliciting and aggregating knowledge such as subsidised prediction markets [6].

Ensuring sufficient dialogue between those with technical expertise in AI and policymakers is essential to ensure policies are developed that are both scientifically appropriate and politically feasible. In general, we need to develop policies that foster a culture of safety, openness and collaboration, and that favour the development of safety techniques in advance of the development of more powerful systems.

Recommendations for policymakers

I. Increase information flow between industry, academics, and policymakers through conferences, reports and intergovernmental panels - for example a Science and Technology committee enquiry, or a report modelled on the Stern Review. Policymakers should listen to research institutes working on these issues such as the Centre for the Study of Existential risk, or the Future of Humanity Institute.

II. Develop milestones against which progress in AI can be tracked, and anticipatory policy frameworks that governments, researchers and industry should abide by at each milestone. Revise and update this guidance regularly through collaboration with industry and academics as we gain new information.

III. Incentivise safety research by preferentially funding research focused on safety and value alignment over research focused purely on increasing AI capabilities, and by creating platforms to enable pre-competitive industry collaboration on safety-related issues that provide little competitive advantage.

IV. Improve forecasting and strategy - *e.g.*, through projects similar to the US's IARPA ACE forecasting tournament, which aims to identify the best strategies for soliciting and aggregating expert judgement on uncertain topics. Prediction markets or other knowledge aggregation methods allow people to place bets on the probability of particular events occurring and be rewarded for correct predictions, with multiple benefits: people with special knowledge are incentivised to share it, 'superforecasters' with much higher than average prediction ability can be identified, and aggregating the judgements of large numbers of people can reduce bias and improve accuracy if done correctly. This information can improve the quality of policymaking when dealing with uncertain future events such as AI development.

Research priorities in computer science

Many priorities for beneficial development are common to short and long timescales. These include economic, policy and strategic questions about the impacts of AI, such as its impact on employment. They also include technical questions in the field of computer science. It is uncertain exactly what research is needed, but some of the following may be helpful:

I. Transparency: developing techniques to investigate the features an AI system is using to make a decision. This allows humans to spot when the system has learnt an incorrect algorithm.

II. Avoiding negative side effects: if we have an AI system with a particular goal, how can we ensure it does not disturb or damage its environment while carrying out that goal?

III. Safe exploration: how do we ensure that a machine learning system during training only experiments in a safe way?

IV. Predictability and robustness: designing systems that will behave safely even with novel inputs, and avoid e.g. 'flash crashes'. This involves flagging when the environment has changed and proceeding with caution by e.g. requesting human oversight.

V. Value alignment: developing techniques to ensure an AI acquires the intended goal function. Inverse reinforcement learning is a potentially promising method, but other approaches are needed. Verifying whether the goal function has been learned or specified correctly may also be important. [7].

VI. Control: building safeguards into a system to enable human control. Increasing our ability to rapidly intervene and contain any damage done if a system fails.

VII. Developing safe virtual environments for testing AI systems

Recommendations for researchers

I. Consider working in one of the areas highlighted as important for beneficial AI development. These are explained in detail in the paper 'Concrete Problems in AI Safety' and Future of Life Institute's paper 'Research priorities for robust and beneficial artificial intelligence' and are summarized above. These areas are receiving a growing amount of interest and funding. [12][13]

II. If you have relevant technical or policy expertise, consider becoming an advisor to a research centre working on these issues – for example, the new Leverhulme Centre for the Future of Intelligence or the Centre for the Study of Existential Risk in Cambridge, or the Future of Humanity Institute in Oxford.

References

[1] Bostrom, Nick. Superintelligence. Oxford: OUP, 2014.;
[2] Russell, Stuart. "The Long-Term Future Of AI". Cs.berkeley.edu. N.p., 2016. Web. 7 Mar. 2016.;
[3] Müller, Vincent C. and Bostrom, Nick. "Future progress in artificial intelligence: A Survey of Expert Opinion", in Vincent C. Müller (ed.), Fundamental Issues of Artificial Intelligence; 2014
[4] Nass, Clifford and Youngme Moon. "Machines And Mindlessness: Social Responses To Computers". Journal of Social Issues 56.1 (2000): 81-103. Web.

About the Author

Elizabeth is a computer science student at Cambridge. Her interests include risks from emerging technologies such as AI, geoengineering and synthetic biology. She has been a contributing author to The Wilberforce Society on AI policy. She is a fellow of the Pareto program, a project of the Centre For Effective Altruism that supplies young people with the skills, resources and network to undertake ambitious projects that have an exceptionally large positive impact on the world. She is the president and founder of Future of Sentience Cambridge, a student society with the mission to empower future leaders to steer technological development wisely.