THE ROYAL SOCIETY

UK Science & Innovation Network

Machine learning and AI for social good

Note of a discussion convened by the Royal Society and UK Foreign and Commonwealth Office Science and Innovation Network on 6 December 2017.

Background

The Royal Society's machine learning project set out to investigate the potential of machine learning over the next 5 – 10 years and the barriers to realising that potential. Its report *Machine learning: the power and promise of computers that learn by example* called for action in key areas to support the development of machine learning and to help share the benefits of this technology across society[1].

In December 2017, the Royal Society and the UK Foreign and Commonwealth Office Science and Innovation Network convened a discussion on the margins of the Neural Information Processing Systems (NIPS) conference to explore current developments in machine learning, emerging applications and how machine learning and AI can be developed in ways that support broad societal benefits.

This note summarises the main points raised during the meeting, which took place under the Chatham House Rule. This summary is non-attributable and does not reflect a consensus amongst those present or the views of the sponsoring organisations[2].

Applying machine learning and AI to societal or public policy issues: opportunities and challenges

Opportunities for new applications

A growing community of researchers is interested in questions about how to develop machine learning for societal benefit. This community includes those that approach such questions from a philosophical perspective (for example investigating the type of utility functions with which systems should be designed) and those interested in questions around the application of machine learning and AI. These questions are not about whether AI or machine learning are inherently good or bad, but how, where and why society should make use of AI technologies.

The development of AI for social good could refer to: applications; research into areas of societal impact or concern; sustainable development goals; freely available technologies from which people derive value or enjoyment; or systems that increase productivity and support economic growth.

1. These areas are: creating an amenable data environment, building skills at all levels, supporting businesses, facilitating public engagement and dialogue and advancing research in key areas of interest.
2. The Chatham House Rule reads as follows: When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed. See https://www.chathamhouse.org/about/chatham-house-rule

Research into applications of AI for societal benefit is already being carried out by a range of projects and initiatives. For example:

- Workshops at the NIPS 2017 conference were exploring applications of AI in healthcare[3], transport[4] and socially-assistive robotics.
- Teams in the AI Xprize competition are developing applications across these areas and more, including education, drug-discovery and scientific research[5].

Using data for social good

Applying AI to public policy challenges often requires access to complex, multi-modal data about people and public services. While many national or local government administrations, or non-governmental actors, hold significant amounts of data that could be of value in applications of AI for social good, this data can be difficult to put to use. Institutional, cultural, administrative, or financial barriers can make accessing the data difficult in the first instance. If accessible in principle, this type of data is also often difficult to use in practice: it might be held in outdated systems, be organised to different standards, suffer from compatibility issues with other datasets, or be subject to differing levels of protection. Use of such data is also often linked to questions about governance and the ways in which regulatory systems allow data to be used.

One approach to facilitating further activity in applying machine learning to public policy challenges could be for government administrations to identify forthcoming areas of interest and to create frameworks that enable access to data that is relevant to these.

In addition to these questions about accessibility, the complexity associated with these datasets – or the need to combine multiple datasets to understand the different facets of a policy challenge – means that they often require significant effort to make them usable. Support for data engineering is therefore important and needs to be budgeted for in terms of both funding and project time[6].

Incentive structures for research into applications of interest

In addition to requiring access to data, successful research in areas of social good often requires interdisciplinary teams that combine machine learning expertise with domain expertise. Creating these teams can be challenging, particularly in an environment where funding structures or pressure to publish certain types of research may contribute to an incentives structure that favours problems with 'clean' solutions.

In this context, challenge-based funding can play a role in spurring research in an area of interest and in fostering collaborations between public, private and third sector actors. Such collaborations can support research in areas not traditionally supported by funding from commercial interests and where existing funding structures might not be able to support longer-term or open-ended research areas.

Areas of research interest in machine learning for social good

In addition to those questions that arise in using machine learning for specific applications or policy challenges, there exists a number of technical challenges that may shape how machine learning is able to respond to questions of societal interest[7].

Fairness and inequality

Bias in machine learning systems can arise in different ways: systems can inherit subjective biases from the data on which they are trained, or they can use personal characteristics as predictive of outcomes in ways that society may deem inappropriate.

There are already examples of areas in which machine learning applications have raised questions about fairness and inequality. For example, a natural language system developed by Google was found to assign 'positive' or 'negative' labels to personal characteristics in its sentiment analysis in a biased way[8]. If not managed and addressed, such biases could have implications for access to finance, effectiveness of healthcare and the ability to use machine learning systems for functions such as image recognition, amongst other applications.

3. https://nips.cc/Conferences/2017/Schedule?showEvent=8728 and https://nips.cc/Conferences/2017/Schedule?showEvent=9561
4. https://nips.cc/Conferences/2017/Schedule?showEvent=8755
5. https://ai.xprize.org/news/blog/ai-xprize-top-10-teams
6. For further discussion of this, see: *Machine learning: the power and promise of computers that learn by example*, The Royal Society, April 2017, available at royalsociety.org/machine-learning
7. The Royal Society has commented on a number of these areas in further detail in its machine learning report, calling for a new wave of machine learning research to address them.
8. As discussed in Kate Crawford's NIPS Keynote, available at https://www.youtube.com/watch?v=fMym_BKWQzk

Advances in machine learning research could help address questions about bias and inequality, by creating systems that can use data in accordance with anti-discrimination provisions, for example by restricting how different inputs are used. This offers the hope that algorithmic bias may be addressed and removed from machine learning systems, potentially making them 'fairer' than their human counterparts.

Interpretability

Some machine learning methods are highly accurate, but difficult to interpret. This can restrict the extent to which they are used in some applications, particularly those in safety-critical environments. Interpretability can also be important in helping researchers understand how or where there may be biases in a machine learning system.

There are a number of ways in which technical advances can improve the interpretability of machine learning systems and these are being actively pursued in areas of the research community.

Questions around AI safety

Progress in reinforcement learning[9] has underpinned a number of recent advances in machine learning, such as the success in playing Go[10] and further advances could support progress in robotics, helping robots to understand how to move through their environment[11].

As agents move through their environment using reinforcement learning, it is important that they do so in a way that does not harm themselves or others. Questions about safe exploration in reinforcement learning are likely therefore to play an important role in developing algorithms that can minimise such risks and develop goals and priorities that are desirable for society. As part of these calculations, an agent may need to weigh the value of 'curiosity' against the need to invest resources into exploring potential actions and their consequences.

Creating an environment that supports the application of machine learning for social good

Indications from conferences such as NIPS are that the research community is beginning to engage with questions about both the applications of machine learning and its future research directions, though some question the extent of this engagement.

As the field progresses, supporting the application of AI for social good will require a policy environment that enables access to appropriate data and that supports research to address areas of societal interest. It will also require support for skills development, both in the AI community and for the wider population, as the implications of AI-enabled automation for the labour market become clearer.

The research community has an active role to play in creating an environment of careful stewardship as machine learning develops, whether through understanding the impact of bias (or similar ethical challenges) in their research, engaging in debates about the place of machine learning in society, or by developing codes of conduct and professional practices around their field. Initiatives such as the AI Index are also helping communicate progress in the field, thereby helping to broaden understanding of its societal implications[12].

The Royal Society has set out how a new wave of research in key areas – including privacy, fairness, interpretability and human-machine interaction – could support the development of machine learning in a way that addresses areas of societal interest. As research and policy discussions around machine learning and AI progress, the Society will be continuing to play an active role in catalysing discussions about these challenges.

9. Reinforcement learning seeks to understand which actions will most effectively achieve a desired endpoint as an agent or computer program interacts with its environment, making sequential decisions in a way that aims to maximise its future rewards.
10. See: *Machine learning: the power and promise of computers that learn by example*, The Royal Society, April 2017, page 27, available at royalsociety.org/machine-learning
11. See, for example, Pieter Abbeel's NIPS Keynote, available at https://www.youtube.com/watch?v=TyOooJC_bLY
12. https://aiindex.org