AI RFI Responses, October 26, 2018

Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan RFI Responses

DISCLAIMER: The RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government, National Science and Technology Council (NSTC) Select Committee on Artificial Intelligence (AI), NSTC Subcommittee on Machine Learning and AI, NSTC Subcommittee on Networking and Information Technology Research and Development (NITRD), NITRD National Coordination Office, and/or any other Federal agencies and/or government entities.
We bear no responsibility for the accuracy, legality or content of all external links included in this document.

PRINCETON UNIVERSITY

Devon Loehr, *Doctoral Student*

NCO/NITRD
490 L'Enfant Plaza SW, Suite 8001
Washington, DC 20024, USA

October 18, 2018

Dear members of the NITRD NCO,

An important issue in modern artificial intelligence research is the problem of *opacity*, also known as *model interpretability*. Modern AIs can filter spam, trade stocks, and even drive cars. However, our ability to understand *why* an AI makes its decisions has not kept up. Why was *this* message classified as spam, and not that one? Why *sell* stocks, and not buy? Why accelerate, and not brake? Understanding the answers to these questions becomes extremely important as we entrust more and more of our tasks to AIs.

This has relations to three of the existing strategies in the plan: Strategy 2, which calls for better methods of human-AI collaboration; Strategy 3, which aims for better understanding of the consequences of AI; and Strategy 4, which (among other things) seeks to ensure that AI systems operate in a “controlled, well-defined and well-understood manner." However, none of these address the issue of opacity directly, so I call for the addition of an eighth strategy:

Strategy 8: Ensure that AI are capable of explaining their decisions.

An Illustrative Story

To see why opacity may be a problem, consider the “tank story” which has been widely circulated in the machine learning field.[1] In the story, the Pentagon commissioned an artificial intelligence to automatically identify whether or not a given surveillance photo contains a tank. The AI was trained on two sets of pictures, one which contained tanks and one which did not. Eventually it was able to correctly distinguish these two sets. However, when it was given other images, its classification ability was no better than random guessing. After some time, the developers realized that all the pictures in one set were taken on a sunny day, and the others on a cloudy day. The AI had learned not to recognize tanks, but to recognize the color of the sky.

It is not entirely clear whether this story is apocryphal;[2] however, its prevalence in the community occurs because it is so eminently *plausible*. If the AI had been able to explain its decision making (for instance, via occlusion mapping)[3], then the developers would have quickly realized it was looking at the color of the sky, and not at anything related to the existence of a tank. However, the AI could not do this, and learned the wrong thing -- an outcome commonly referred to as *overfitting*. Overfitting occurs when an AI learns something what works for the situation in which is was trained, but doesn’t work in general.

Further examples

A more sinister instance of opacity is the existence of *adversarial examples*. These are slight modifications to the inputs of an AI -- changes so tiny that they would never confuse a human. However, as a result of the change, the input is interpreted completely differently. Adversarial examples can exist in image classification, natural language processing,[4] and can even be created in the real world.[5] They indicate a flaw in the AI, for if it truly understood its task (as a human does) it would not be fooled by such minor changes. We can guard against adversarial examples by ensuring that AIs have the ability to explain their decisions -- if they can do so in a way that satisfies a human, then they are unlikely to make mistakes a human would not.

Finally, we might care about the *reasons* for a decision as much as the decision itself. For example, California recently abolished its cash bail system. One component of its new system is an algorithm which is used to determine the likelihood of someone committing another crime if they were released pre-trial.[6] This is then used to determine whether to grant that person bail or not. It is reasonable to require that this algorithm *not* take certain pieces of information into account -- for example, race or religion. A good step in this direction is to simply not give the AI information about their race -- however, this does not prevent it from using information that may be correlated with race, such as where they live. Hence it is important that the AI be able to explain its reasoning -- only then can we verify it was basing its output on the “right” reasons.

A Counterargument

There do exist arguments in favor of opacity; for example, software companies may wish to keep their decision algorithms proprietary.[7] Furthermore, sometimes an AI can (correctly) deduce something so wildly unintuitive that no human could possibly come up with it. A good example here is the game of chess. A sufficiently powerful AI might make a move which it could not possibly explain to human players, which might even seem foolish at first. Yet somehow a few moves later it comes out in a much stronger position than before.[8] This occurs because the AI is not thinking like a human; it is not thinking of strategy. It is thinking of the state of the board, and of possible moves. If there were a human overseeing the AI's moves, it certainly would not have been allowed to play the same way, and would have been worse off for it.

So perhaps it is unreasonable to expect AIs to be able to explain themselves, especially as they get more and more powerful. This is a valid argument, and indicates that some discretion might be necessary in determining when exactly we expect an artificial intelligence to explain itself. However, consider an AI designed to oversee mission-critical infrastructure. If it issues an unlikely order, it seems entirely reasonable to expect it to explain itself -- otherwise humans will not be able to distinguish a moment of "artificial genius" from a flaw in its reasoning.

Conclusion

In summary, opacity is a significant problem in artificial intelligence. Although artificial intelligences are capable of giving seemingly correct answers, they might fail when presented with new data (*overfitting*), or slight modifications to the original data (*adversarial examples*). Opacity makes it very difficult to evaluate whether an AI actually understands the problem, or merely *seems* to understand. Therefore, the NITRD should encourage developers of artificial intelligence to ensure that their AIs are capable of not only giving the correct answer, but explaining how they got it.

Citations

1. Fraser, Neil. “Neural Network Follies.” Neil Fraser: Writing: Neural Network Follies, 2018, neil.fraser.name/writing/tank/.
2. Branwen, Gwern. “The Neural Net Tank Urban Legend.” Sitewide Global ATOMRSS, 24 Dec. 2015, www.gwern.net/Tanks.
3. Campbell, A. (2018). Model Interpretability with Occlusion Mapping - “An AI Tells us What it Knows When We Poke it in the Eye” - Silverpond. Retrieved October 18, 2018, from https://silverpond.com.au/2018/04/17/an-ai-tells-us-what-it-knows-when-we-poke-it-in-the-eye/
4. Pavlick, Ellie. Should we care about linguistics? [pdf]. Retrieved from http://helper.ipam.ucla.edu/publications/dlt2018/dlt2018_14546.pdf
5. Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. Retrieved from http://arxiv.org/abs/1607.02533
6. AI in the court: Algorithms help rule on jail time - SFGate. (n.d.). Retrieved October 16, 2018, from https://www.sfgate.com/business/article/AI-in-the-court-Algorithms-help-rule-on-jail-time-12547592.php
7. Burrel, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. Journals.Sagepub.Com. https://doi.org/10.1177/2053951715622512
8. How Google’s AI Viewed the Move No Human Could Understand | WIRED. (n.d.). Retrieved October 16, 2018, from https://www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand