Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Public Communication

This input paper was prepared by Mark Alfano

Suggested Citation

Alfano, M (2018). Public Communication. Input paper for the Horizon Scanning Project “The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing” on behalf of the Australian Council of Learned Academies, www.acola.org.

The views and opinions expressed in this report are those of the author and do not necessarily reflect the opinions of ACOLA.

Report for Australian Council of Learned Academies

Mark Alfano, Delft University of Technology & Australian Catholic University with input from: Marc Cheong, Monash University, Adam Carter, Glasgow University, and Emily Sullivan, Delft University of Technology.

1 Background and international context

Artificial intelligence (AI) of various levels of sophistication and specialization is encroaching on every aspect of contemporary life. It can expand the decision-making capacities of over-stretched laborers, bureaucrats, managers, and professionals. It can furnish guidance at any time, on any day, in any place, at very low cost. It can often make decisions that are at least as good as those typically made by experts. And unlike experts, AI doesn't retire or die after a few decades. Also unlike experts, AI is incapable of holding personal grudges, nor does it get tired, bored, or angry. It helps people find information, make friends, navigate cities, determine whom to hire and fire, predict epidemics, diagnose medical conditions, and identify and track criminals.

Until recently, decision-making in these domains was the exclusive purview of human adults. Our epistemic, ethical, and political capacities enable us to engage in such activities and — in the ideal case — explain our decisions to the people they affect, to the general public, and to ourselves. This explanatory ability and the normative expectations surrounding it are a reason why the European Union articulated the right to explanation in the General Data Protection Regulation.[1]

The human capacity not only to make good-enough decisions but also to explain our decisions to affected and interested parties is an essential ingredient in liberal democracies and republics. One of the key differences between the status accorded a mere political subject and that accorded a citizen is that citizens are presumed to be autonomous in the sense that they make choices based on reasons. This is only possible in a social and political environment in which people have adequate epistemic access to the reasons that bear on their choices. In addition, one of the epistemic presuppositions of democratic deliberation is that citizens have access to enough of the same information and truths that they share common ground on which to debate policies, institutions, and other arrangements.

The growing use of online media has brought the problems of filter bubbles (Pariser 2011), echo chambers (Nguyen forthcoming), and group polarization (Sunstein 2017) into focus. Recent journalism has labeled the current era a time of "epistemic crisis" (Roberts 2017a) in which "tribal epistemology" dominates (Roberts 2017b). Levy (2017) goes so far as to suggest that the best response to the fake news crisis is to cut oneself off entirely from many sources of information.

2 The problem of explainability in AI

The algorithms underlying AI are sophisticated, but their workings are often difficult to decipher. For example, one of the most important algorithms underlying the search functionality of Google and various other engines is PageRank (Brin & Page 1998; Brin et al. 1998). Under certain conditions, this algorithm is capable of harnessing the wisdom of crowds (Masterson et al. 2016). However, it can be difficult to explain PageRank to people without a strong mathematical background. More recent developments, such as Google's TensorFlow, are opaque even to their designers. In a story about the development of TensorFlow, Lewis-Kraus (2016) quotes some of the main engineers involved in the project admitting that "They didn't know themselves why [their code] worked." AI of this sort does not rely on branching decision trees where the meaning of each branch is explicable or on the tallying of interpretable points. Instead, it is built on artificial neural networks that respond holistically to a very large number of variables based on very large training datasets.

[1] URL = < https://gdpr-info.eu/art-22-gdpr/ >.

In cases of supervised learning, we have some idea of what an AI is sensitized to (and not sensitized to) because the training data are coded by humans with human-interpretable categories. However, supervised learning is guaranteed to embed human biases and systematic errors in the algorithms trained with it (Caliskan-Islam et al. 2016). And when training data is not made publicly accessible, it can be difficult to understand or explain how (and whether) errors arise. For an example of the bizarre outputs that TensorFlow sometimes delivers, see Jon Christian's (2018) journalistic exposé on Google Translate, which has recently translated repeated use of the word 'dog' (allegedly in Maori) into English as: "Doomsday Clock is three minutes at twelve We are experiencing characters and a dramatic developments in the world, which indicate that we are increasingly approaching the end times and Jesus' return."

AI based in supervised learning is troubling enough, but in cases of unsupervised learning it is in principle impossible to assess outputs for accuracy or reliability (Hastie et al. 2008).

3 Social epistemology, democracy, and AI

The problem of explainability relates directly to the epistemic and political acceptability of AI. AI increasingly determines how citizens acquire information. Many people get their news from Twitter (which relies on PageRank) and Facebook (which relies on EdgeRank, a variant of PageRank). At the same time, they search for information and translate texts from other languages using Google's tools, which rely on PageRank, TensorFlow, and other AI infrastructures and algorithms.

It's not hard to envision a future in which, through some combination of negligence and malicious interference, the technologies described above produce devasting consequences. A significant proportion of the population could end up deeply misinformed or disinformed, and it would be very difficult to trace, track, and address the causes. For example, PageRank could be hijacked by creating websites, social media accounts, and links that systematically violate the constraints that enable it to harness the wisdom of crowds. This could be done by an organized group of trolls or a hostile foreign power. EdgeRank and TensorFlow can be hijacked in the same way. Indeed, there is evidence that this has already happened in connection with the Brexit referendum (Booth et al. 2017; Sabbagh 2018), the 2016 US Presidential election (Smith 2018), and other high-stakes decisions.

When conspiracy theories, extremist content, outright propaganda, and other epistemically suspect sources of information are promoted by on Facebook, Twitter, YouTube, and other platforms that increasingly use AI, it's hard to know why. This is in part because of the explainability problem mentioned above. It's exacerbated when the training data and code these companies use is not released for inspection, criticism, and correction.

To make matters worse, even if training data and code are released, the personalization of people's newsfeeds and search results makes it difficult or even impossible to reproduce the processes that led to troubling outcomes (Alfano et al. 2018). This in turn means that it is difficult or even impossible to diagnose and correct these processes.

For example, Google creates suggestions either by aggregating other users' data or by personalizing for each user based on their location, search history, or other data. It takes a user's own record of engagement as the basis for delivering search results and video recommendations. Engagement, in this context, refers to all recorded aspects of a user's individual online behavior. This includes their browsing history (which sites/links they visit, frequency of such visits, etc.), their search history, their record of sharing and "liking" posts on social media, their email record (if, for example, they use Google for both search and email), their physical location (if, for example, they use Google's location services for navigation, or merely having the 'location history'[2] feature active), and so on. While it is possible to disguise these aspects of one's online signature in various ways, most users

[2] URL = <https://support.google.com/accounts/answer/3118687?hl=en>.

neglect to do so. In addition to the individual's own record of engagement, others' records of engagement can be used to profile that individual. To the extent that your record of engagement — even in depersonalized aggregated form — is more similar to that of one set of users than that of another set of users, you're liable to be profiled among the former.

Profiling enables both predictive and prescriptive analytics to tell a user what to think and what to do. This is especially worrisome when the process bypasses the user's capacity for reasoning. Following Koralus & Mascarenhas (2013) on reasoning in general and Koralus & Alfano (2017) on moral reasoning in particular, we can construe reasoning as the iterative, path-dependent process of asking and answering questions. Profiling enables online interfaces such as Google to tailor both search suggestions (using predictive analytics) and answers to search queries (using prescriptive analytics) to an individual user.

Consider a simple example: predictive analytics will suggest, based on a user's profile and the initial text string they enter, which query they might want to run. For instance, if you type 'why are women' into Google's search bar, you are likely to see suggested queries such as 'why are women colder than men', 'why are women protesting', and 'why are women so mean'. And if you type 'why are men' into Google's search bar, you are likely to see suggested queries such as 'why are men jerks', 'why are men taller than women', and 'why are men attracted to breasts'. These are instances of predictive analytics. The same predictive searches conducted in another geographic location, at another time, by an account with a different history and social graph will yield different results.[3]

Prescriptive analytics in turn suggests answers based on both the query someone runs and their profile. In its most naïve form, a search for the query 'cafe' returns results for cafés nearest to the user; the top results will differ for someone in Amsterdam as opposed to Abuja. To continue with our prior examples: in response to 'why are women colder than men', one of Google's top suggestions is a post titled "Why are Women Always Cold and Men Always Hot," which claims that differences between the sexes in the phenomenology of temperature are due to the fact that men have scrotums.[4] In response to 'why are men jerks', one of Google's top suggestions is a post titled "The Truth Behind Why Men are Assholes," which contends that men need to act like assholes to establish their dominance and ensure a balance of power between the sexes.[5] And in response to 'why are women so mean', Google suggests posts answering questions about why beautiful women in particular are so mean, why women are so mean to each other, and why women are so mean to men. Most of these posts have a strongly misogynistic flavor.

In cases like this, Google suggests questions and then answers to those very questions, thereby closing the loop on the first stage of an iterative, path-dependent process of reasoning. If reasoning is the process of asking and answering questions, then the interaction between predictive and prescriptive analytics can largely bypass the individual's contribution to reasoning, supplying both a question *and* its answer.

Consider next the path-dependency mentioned above. Which question you ask depends in part on both the questions you asked previously and the answers you accepted to those questions. If both the initial question and its answer are shaped by predictive and prescriptive analytics, then the first question-answer pair in the process of reasoning largely bypasses the human's contribution. But that in turn means that subsequent questions and answers depend on this bypassing, potentially sending the user deeper into an epistemic and ethical morass.

[3] Depending on a user's profile, the content of search results can be subject to change, as in the case of Google's Personalised Search, which can "customize search results for you based upon 180 days of search activity linked to an anonymous cookie in your browser". See URL = < https://googleblog.blogspot.com.au/2009/12/personalized-search-for-everyone.html >.

[4] URL = < https://www.qualityhealth.com/womens-health-articles/why-women-always-cold-men-always-hot >.

[5] URL = < http://elitedaily.com/dating/sex/men-assholes/ >.

To illustrate, suppose you were interested in anything beginning with the text string 'alt', such as 'alternative energy'. You type these first three letters into Google's search bar, and it suggests 'alt right'. Though you weren't initially interested in this query, the suggestion piques your curiosity. You run the 'alt right' query, and several of the top results are videos on YouTube (a subsidiary of Google's parent company, Alphabet). The top result is a video by *The Atlantic* titled, "Rebranding White Nationalism: Richard Spencer's Alt-Right."[6] After watching this eleven-minute video, you allow the top suggested video (as determined by the video you clicked on and your own profile) to auto-play. It's a clip titled "White nationalist Richard Spencer talks to Al Jazeera."[7] When the video ends, you allow the next top suggested video to auto-play: "BEST OF Richard Spencer vs Hostile Audience at Texas A&M."[8] This is a post by the white supremacist account Demography is Destiny. It celebrates Spencer's political positions and those like them. The first three letters of an innocent online search have been hijacked: in just a few steps, you went from the start of a query about alternative energy to Demography is Destiny.[9]

YouTube is especially adept at this kind of epistemic seduction because it uses AI to find patterns in individuals' and groups' preferences, then recommend clips that they're most likely to engage with (Newton 2017):

> "We knew people were coming to YouTube when they knew what they were coming to look for," says Jim McFadden, the technical lead for YouTube recommendations, who joined the company in 2011. "We also wanted to serve the needs of people when they didn't necessarily know what they wanted to look for."'

McFadden's team succeeded. The vast majority of the time people spend watching videos on YouTube is now driven by algorithmic recommendations rather than search or linking. Whistleblower Guillaume Chaslot, who was fired by YouTube in 2013 for raising this criticism, has shown that YouTube recommendations are systematically biased in favour of bizarre, violent, and extremist content (Lewis 2018).

Even if only a portion of the population is influenced in the ways envisioned, our democratic institutions will suffer. People will find themselves in disagreement about what should be basic common knowledge. Each side will be able to point to their own sources of information as an epistemic backstop. Determining which sources are problematic will be difficult or impossible both because the AI that recommends the sources is difficult or impossible to explain and because the training data and code are treated as proprietary intellectual property.

4 Recommendations

There is no single solution to this problem. However, several remedies are likely to be helpful in combination. First, corporations such as Facebook, Google, and Twitter should be required by law to open up both their datasets and the AI algorithms and infrastructures they use. Google is ahead of the game on this, with a large and important initiative spearheaded by Margaret Mitchell. Second, research on the explainability gap in AI should be funded by the government and industry. Third, the Australian government should seriously consider following the EU in upholding a legal right to explanation — and go further than the EU in enforcing this right.

References:

Alfano, M., Carter, A., & Cheong, M. (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association.*

[6] URL = < https://www.youtube.com/watch?v=kVeZ0_Lhazw >.

[7] URL = < https://www.youtube.com/watch?v=nhrJg4FqzTA >.

[8] URL = < https://www.youtube.com/watch?v=FxfDOOY2H28>.

[9] See also Tufekci (2018) for a recent discussion of this phenomenon.

Booth, R., Weaver, M., Hern, A., Smith, S., & Walker, S. (2017, November 15). Russia used hundreds of fake accounts to tweet about Brexit, data shows. *The Guardian*. URL = < https://www.theguardian.com/technology/2018/jan/17/facebook-inquiry-russia-influence-brexit >.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine, WWW 1998. In *Seventh International World-wide Web Conference*. Brisbane, Australia.

Brin, S., Page, L., Motwami, R., & Winograd, T. (1998). *The PageRank citation ranking: Bringing order to the web*. Stanford University Technical Report.

Caliskan, Bryson, & Narayanan (2016). Semantics derived automatically from language corpora contain human-like biases. Doi = < 10.1126/science.aal4230 >.

Christian, J. (2018, 21 July). Why is Google Translate spitting out sinister religious prophecies? *Motherboard*. URL = < https://motherboard.vice.com/en_us/article/j5npeg/why-is-google-translate-spitting-out-sinister-religious-prophecies >.

Hastie, Tibshirani, & Friedman (2008). *The elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Koralus, P. & Alfano, M. (2017). Reasons-based moral judgment and the erotetic theory. In J.-F. Bonnefon & B. Trémolière (eds.), *Moral Inferences*. 77-106. Routledge.

Koralus, P. and Mascarenhas, S. (2013). “The erotetic theory of reasoning: Bridges between formal semantics and the psychology of propositional deductive inference.” *Philosophical Perspectives*, 27: 312-365.

Levy, N. (2017). The bad news about fake news. *Social Epistemology Review and Reply Collective*, 6(8): 20-36.

Lewis, P. (2018, February 2). ‘Fiction is outperforming reality’: How YouTube’s algorithm distorts truth. *The Guardian*. URL = < https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth >.

Lewis-Kraus, G. (2016, December 14). The great AI awakening. *The New York Times*. URL = < https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html >.

Masterson, G., Olsson, E., & Angerre, S. (2016). Linking as voting: How the Condorcet jury theory in political science is relevant to webometrics. *Scientometrics*, 106: 945-66.

Newton, C. 2017. How YouTube perfected the feed. *The Verge*. URL = < https://www.theverge.com/2017/8/30/16222850/youtube-google-brain-algorithm-video-recommendation-personalized-feed >.

Nguyen, C. T. (forthcoming). Cognitive islands and runaway echo chambers: Problems for epistemic dependence on experts. *Synthese.*

Pariser, E. (2011). *The Filter Bubble: What the Internet is Hiding from You*. New York: Penguin Press.

Roberts, D. (2017a, November 2). America is facing an epistemic crisis. *Vox Media*. Url = <www.vox.com/policy-and-politics/2017/11/2/16588964/america-epistemic-crisis>.

Roberts, D. (2017b, May 19). Donald Trump and the rise of tribal epistemology. *Vox Media*. Url = <www.vox.com/policy-and-politics/2017/3/22/14762030/donald-trump-tribal-epistemology>.

Sabbagh, D. (2018, January 18). Facebook to expand inquiry into Russian influence of Brexit. *The Guardian*. URL = < https://www.theguardian.com/technology/2018/jan/17/facebook-inquiry-russia-influence-brexit >.

Smith, D. (2018, February 18). Putin’s chef, a troll farm and Russia’s plot to hijack US democracy. *The Guardian*. URL = < https://www.theguardian.com/us-news/2018/feb/17/putins-chef-a-troll-farm-and-russias-plot-to-hijack-us-democracy >.

Sunstein, C. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Tufekci, Z. (2018, March 10). YouTube, the Great Radicalizer. *The New York Times*. URL = < https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html >.