Horizon Scanning Series

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

Fake News

This input paper was prepared by Neil Levy

Suggested Citation

Levy, N (2018). Fake News. 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.

4.1.3

How can we limit or combat the propagation of 'fake news' about AI? Might such measures themselves reduce trust?

Fostering community trust in and acceptance of AI is a multi-pronged challenge. Steps must be taken to ensure that personal data is secure, that algorithms are fair to all stakeholders, that the uses of AI are transparent and the goals themselves have broad public acceptance. These challenges are dealt with elsewhere in this report. Here we focus on another challenge, arising from false reports about AI that can be expected to circulate online (ironically, AI will facilitate the broader reach and more effective targeting of such false reports). We will discuss this challenge under the heading "fake news".

Fake news propagates freely online, and influences people's opinions. Its influence is ill-understood, but appears to be attributable to a number of independent and interacting factors. Prominent among them are the following:

(a) echo chamber effects;
(b) biased assimilation of information;
(c) confirmation bias.

While there has not yet been any systematic study of the extent to which attitudes to AI are shaped by fake news, or any indication that there is a large problem in this area yet, the examples of vaccines and GMOs provide object lessons in how such a problem may emerge, as well as indications of how we might be respond to the challenge.

Fake news refers here to the dissemination of false information in a way that is designed to sway opinion and in a format that mimics or borrows the authority of legitimate news sites. Fake news reaches a wide audience: during the 2016 US presidential election, the most popular fake stories were shared more widely than the most popular genuine new stories (Vosoughi, Roy & Aral 2018). The effects of fake news are exacerbated by the declining reach of and declining trust in mainstream media (Allcott and Gentzkow 2017). Accordingly, fact checks are of limited utility, since they will not reach a broad audience and may be distrusted.

The effects of fake news are often exaggerated. Most information consumers are exposed to a wide range of sources, and are less credulous than often believed (Guess, Nyhan & Reifler 2018). However, the example of vaccines provides a case study of how difficult it can be to correct false information.

In 1998, Andrew Wakefield and his co-authors published a paper alleging a link between the MMR vaccine and autism. After other researchers failed to replicate his findings, Wakefield was found to have undisclosed conflicts of interest. The British General Medical Council then investigated further and found a litany of other problems, from performing unnecessary and invasive procedures on children with autism to suppressing data. The paper was retracted, and Wakefield was struck off the medical register. Further work has debunked any link between vaccines and autism (e.g. Taylor, Swerdfeger & Eslick 2014). But the claims Wakefield made continue to circulate online, and they are often believed. While vaccination rates remain high overall, in some areas parents refuse to have their children vaccinated in significant numbers, and this refusal has played a role in several recent outbreaks of measles. For example, in Italy there were more than 3,300 cases of measles in the first half of 2017 alone. The vast majority of those infected – 88% – were not vaccinated, and 7% had received just one dose of the vaccine (ECDC 2017). Parents who refuse vaccination often

cite the purported link between vaccines and autism, among other reasons, for their refusal (Largent 2012).

Why have the many attempts to debunk the purported link between autism and vaccines been unsuccessful with some people? One reason is that some people live in *echo chambers*: they receive information exclusively, or very largely, from sources that support particular viewpoints. While the extent to which we live in such bubbles is often exaggerated, over 10% of US information consumers receive information only or very largely from sources that promote fake news (Guess, Nyhan & Reifler 2018). Debunking attempts fail to reach many of these people. In fact, people seem to seek out fact checks only for stories they find uncongenial. In addition, however, there is evidence that our psychological dispositions may limit the efficacy of such corrections, even when we are exposed to them.

Corrections of misinformation rarely entirely eliminates reliance on the misinformation unless the person has available what they regard as a satisfactory alternative explanation of the event (Fein, McCloskey, & Tomlinson 1997; Ecker, Lewandowsky, Swire, & Chang 2011). Hence, people may continue to rely on misinformation when trying to explain why a child was diagnosed with autism. In fact, corrections can even backfire: leaving people *more* committed to the false information than they were previously (Nyhan & Reifler 2010; Peter & Koch 2016. Note that Wood & Porter 2016 report contrary evidence). Even when information is accepted, there may be a behavioral backfire: Nyhan, Reifler, Richey & Freed (2014) found that correcting the myth that vaccines cause autism was effective at the level of belief, but actually decreased intention to have one's children vaccinated among parents who were initially least favourable to vaccines. Nyhan and Reifler (2015) documented the same phenomenon with regard to influenza vaccines.

One reason why corrections may fail is due to our biased assimilation of information: the disposition to become more strongly attached to antecedent views given genuinely mixed evidence (Corner, Whitmarsh & Xenias 2012). The confirmation bias (Nickerson 1998) may partially explain biased assimilation. The confirmation bias is a bias toward information that supports our antecedent beliefs, and away from information that undermines them. We deploy our reasoning capacities to criticise claims we dislike, but are credulous towards those we are well disposed toward.

Experimental work on how ordinary people respond to the testimony of others – their reports, whether conveyed verbally, in writing, or through the electronic media – provides clues as to how we can make people's beliefs more responsive to good evidence. We are very dependent on testimony for our knowledge of the world: without it, we would know nothing beyond what we can see and verify for ourselves, and would be ignorant about the rest of the world. We would not be able to make decisions about who to vote for, how to take care of our health, what to buy and where to go. Given this dependence, we are unsurprisingly willing to accept testimony. But experimental work shows that we filter testimony in various ways. As well as attending to the plausibility of the testimony, we also pay attention to the source of the testimony, looking to cues for putting more or less weight on what they say.

We rely, for example, on evidence that the testimony conforms to the majority opinion (Harris 2012). We also prefer testimony from prestigious individuals over testimony from less prestigious (Chudek et al. 2012). The use of these cues is unsurprising, given that the majority is more likely to be right than a minority about factual matters, and that prestigious individuals presumably owe their success to their beliefs (since their beliefs play a central role in explaining their behavior). In addition, we use cues for filtering testimony that are less obvious, like whether the testifier has previously engaged in what we regard as prosocial behavior and from out-group members (Mascaro & Sperber 2009; Sperber et al. 2010; Harris 2012).

On the basis of our growing knowledge concerning the features of testifiers which affect how people weigh testimony, we can design 'nudges' to make people more receptive to testimony. Nudges are ways of designing the context in which people form beliefs and act in ways that make them more rational (Thaler & Sunstein 2009). For example, we can ensure that testimony is tailored to the cues that particular groups of individuals are receptive to. For instance, we are more receptive to the testimony of those we perceive as sharing our values (Levy in press). Nudges can take this into account, for example ensuring that messages are promulgated by people across the political spectrum. There is evidence such nudges are effective. Corrections of false claims are effective when they come from sources that share the ideology of the hearer (Nyhan and Reifler 2013) and also when they come from sources that can be expected to find the claim they affirm contrary to their own ideological interests (Berinsky 2017). Similarly, because we are more receptive to testimony that appears to reflect the majority opinion, we might take steps to burst epistemic bubbles. Insofar as people live in such bubbles, receiving information only or very largely from individuals that share their own views, they may falsely take such views to reflect majority opinion. Bursting the bubble, by exposing them to a wider array of opinions, may make them more receptive.

Implementing these kinds of interventions without imposing censorship or limiting people's freedoms is obviously a difficult and important challenge. Some interventions may utilise the kinds of tools governments and other agencies already use without controversy. For example, messaging that uses prestigious and widely liked individuals, or individuals from across the political spectrum, should not arouse any opposition. Other interventions would require the cooperation of private corporations, such as Facebook and Twitter. These companies might ensure a more balanced diet of information is available to their users.

Nudging is controversial. Thaler and Sunstein (2008) call their program 'libertarian paternalism': it is paternalistic because it intervenes in the context of cognition and choice to promote our interests, but it is libertarian because it does not remove options or impose burdens (e.g. taxes) on any of the options. Critics have rejected the claim that nudges respect autonomy, pointing out that they bypass our reasoning capacities (see Levy 2017 for an overview of these concerns and a response to them). Regardless of whether nudges *do* respect our autonomy, they may be *perceived* to disrespect it or be otherwise unacceptably manipulative. To that extent, any attempt to increase public trust in or acceptance of AI must take into account a possible perverse effect: there is a risk that people will perceive the measures designed to increase trust as themselves untrustworthy. To avoid a possible backfire, any such measures must be designed transparently, in ways that are sensitive to public attitudes.

References.

Allcott, H. & Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31: 211-36.

Berinsky, A. J. 2017. Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science* 47: 241–262.

Chudek, M., Heller, S., Birch, S. & Henrich, J. 2012. Prestige-biased cultural learning: bystander's differential attention to potential models influences children's learning. *Evolution and Human Behavior* 33: 46–56.

Corner, A., Whitmarsh, L., & Xenias, D. 2012. Uncertainty, scepticism and attitudes towards climate change: Biased assimilation and attitude polarisation. *Climatic Change* 114: 463–478.

ECDC 2017. Epidemiological update: Measles - monitoring European outbreaks, 7 July 2017, available at https://ecdc.europa.eu/en/news-events/epidemiological-update-measles-monitoring-european-outbreaks-7-july-2017.

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. 2011. Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review* 18: 570–578.

Fein, S., McCloskey, A. L., & Tomlinson, T. M. 1997. Can the jury disregard that information? The use of suspicion to reduce the prejudicial effects of pretrial publicity and inadmissible testimony. *Personality and Social Psychology Bulletin* 23: 1215–1226.

Guess, A., Nyhan, B., & Reifler, J. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 presidential campaign. Working paper. < http://www.dartmouth.edu/~nyhan/fake-news-2016.pdf >

Harris, P. 2012. *Trusting what you're told: How children learn from others*. Cambridge: Harvard University Press

Largent, Mark 2012. *Vaccine. The Debate in Modern America*. Baltimore; Johns Hopkins University Press.

Levy, N. 2017. Nudges in a Post-Truth World. *Journal of Medical Ethics* 43: 495-500.

Levy, N. In press. Due Deference to Denialism: Explaining Ordinary People's Rejection of Established Scientific Findings. *Synthese*. https://doi.org/10.1007/s11229-017-1477-x

Mascaro, O. & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition* 112: 367–80.

Nickerson, R. S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2: 175–220.

Nyhan, B. and Reifler J. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32: 303-330.

Nyhan, B., & Reifler, J. 2013. *Which corrections work? Research results and practice recommendations*. Washington, D.C.: New America Foundation, Media Policy Initiative.

Nyhan, B., Reifler, J., Richey, S. & Freed, G.L. 2014. Effective messages in vaccine promotion: a randomized trial. *Pediatrics* 133: e835-e842.

Nyhan, B. and Reifler J. 2015. Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine* 33: 459-464.

Peter, C. & Koch, T. 2016. When Debunking Scientific Myths Fails (and When It Does Not): The Backfire Effect in the Context of Journalistic Coverage and Immediate Judgments as Prevention Strategy. *Science Communication* 38: 3-25.

Sperber, D. Clément, F, et al. (2010). Epistemic Vigilance. *Mind & Language* 25: 359-393.

Taylor, L.E., Swerdfeger, A.L., Eslick. G.D. 2014. Vaccines are not associated with autism: an evidence-based meta-analysis of case-control and cohort studies. Vaccine 32: 3623-3629

Thaler, R. & Sunstein, C. 2009. *Nudge. Improving Decisions about Health, Wealth, and Happiness*. Penguin.

Vosoughi, S., Roy, D. & Aral. S. 2018. The Spread of True and False News Online. *Science* 359.6380: 1146-1151.

Wood, T. & Porter, E. In press. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior.* https://doi.org/10.1007/s11109-018-9443-y