CHAPTER B-4

Does Debunking Work? Correcting COVID-19 Misinformation on Social Media[*]

Timothy Caulfield[**]

Citation: Timothy Caulfield, "Does Debunking Work? Correcting COVID-19 Misinformation on Social Media" in Colleen M. Flood, Vanessa MacDonnell, Jane Philpott, et al., eds., Vulnerable: The Law, Policy and Ethics of COVD-19 (University of Ottawa Press, 2020) 183-200.

Abstract

A defining characteristic of this pandemic has been the spread of misinformation. The World Health Organization (WHO) famously called the crisis not just a pandemic, but also an "infodemic." Why and how misinformation spreads and has an impact on behaviours and beliefs is a complex and multidimensional phenomenon. There is an emerging rich academic literature on misinformation, particularly in the context of social media. In this chapter, I focus on two questions: Is debunking an effective strategy? If so, what kind of counter-messaging is most effective? While the data remain complex and, at times, contradictory, there is little doubt that efforts to correct misinformation are worthwhile. In fact, fighting the spread of misinformation should be viewed as an important health and science policy priority.

[*] I would like to thank Gordon Pennycook, Kate Starbird, Briony Swire-Thompson, Robyn Hyde-Lay, Sandro Marcon, Darren Wagner, and the rest of our CIHR COVID-19 rapid response team. I am also grateful for the funding support from the Canadian Institutes of Health Research, Alberta Innovates, the Ministry of Economic Development, Trade and Tourism, the Government of Alberta, and the Government of Canada for their generous support of the following projects: *Coronavirus Outbreak: Mapping and Countering Misinformation* and *Critical Thinking in the Digital Age: Countering Coronavirus Misinformation*.

[**] Canada Research Chair in Health Law and Policy, Professor, Faculty of Law and School of Public Health, Research Director, Health Law Institute, University of Alberta.

Résumé
La démystification fonctionne-t-elle ? Rectifier la désinformation sur les médias sociaux au sujet de la COVID-19

Cette pandémie est marquée par la propagation de la désinformation. L'Organisation mondiale de la santé a qualifié cette crise non seulement de pandémie, mais aussi d'« infodémie ». Pourquoi et comment la désinformation se propage-t-elle et a-t-elle une incidence sur les comportements et les croyances ? Il s'agit d'un phénomène complexe et multidimensionnel. On assiste à l'émergence d'une riche littérature universitaire sur le sujet, en particulier dans le contexte des médias sociaux. Dans ce chapitre, je me concentre sur deux questions : la démystification est-elle une stratégie utile ? Si oui, quel type de contre-message est le plus efficace ? Si les données restent complexes et parfois contradictoires, il apparaît néanmoins que les efforts visant à corriger la désinformation en valent la peine. En fait, la lutte contre la diffusion de fausses informations devrait être considérée comme une priorité des politiques sanitaires et scientifiques.

A defining characteristic of the COVID-19 pandemic has been the spread of misinformation.[1] The WHO famously called the crisis not just a pandemic, but also an "infodemic."[2] Misinformation includes the suggestions that the coronavirus is both caused by 5G wireless technology and is a bioweapon. Cow urine and bleach have been put forward as cures. And enumerable wellness gurus have pushed immune-boosting supplements and diets—all science-free nonsense, of course. It has been suggested that this noise has already, *inter alia*, caused physical harm[3] and financial loss,[4] impacted health and science

[1] Areeb Mian & Shujhat Khan, "Coronavirus: The Spread of Misinformation" [2020] BMC Medicine, online: *BMC Medicine* <https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01556-3>.

[2] World Health Organization, "Infodemic Management" (15 April 2020), online: *World Health Organization* <https://www.who.int/teams/risk-communication/infodemic-management>.

[3] Alistair Smout & Paul Sandle, "Misinformation Ruins Lives, UK Fact-Checker Says", *National Post* (30 April 2020), online: <https://nationalpost.com/pmn/entertainment-pmn/misinformation-ruins-lives-uk-fact-checker-says>.

[4] Greg Iacurci, "Americans Have Lost $13.4 Million to Fraud Linked to Covid-19", *CNBC* (15 April 2020), online: <https://www.cnbc.com/2020/04/15/americans-have-lost-13point4-million-to-fraud-linked-to-covid-19.html>.

policy,[5] added confusion and distraction to an already chaotic information environment,[6] heightened stigma and prejudice,[7] and made it more difficult to implement needed health policy initiatives.[8]

Much of this misinformation is spreading on social media,[9] which has included the use of bots and strategic disinformation campaigns.[10]

It is worth noting that social media has also played a constructive role. It has, for instance, been used as a tool for communicating preventative strategies and mapping the spread of the virus.[11] And

[5] Michael Liu et al, "Internet Searches for Unproven COVID-19 Therapies in the United States" Research Letter (29 April 2020) JAMA Intern Medicine at E1 DOI: <10.1001/jamainternmed.2020.1764>: "Demand for chloroquine and hydroxychloroquine increased substantially following endorsements by high-profile figures and remained high even after a death attributable to chloroquine-containing products was reported".

[6] See generally Amy Mitchell, J Baxter Oliphant & Elisa Shearer, "About Seven-in-Ten U.S. Adults Say They Need to Take Breaks From COVID-19 News" (29 April 2020) at 4, online: *Pew Research Center* <https://www.journalism.org/2020/04/29/about-seven-in-ten-u-s-adults-say-they-need-to-take-breaks-from-covid-19-news/> (it was found that 86% believe that misinformation is causing either a great deal (49%) or some (37%) confusion about basic facts). See also Michael Sean Pepper & Stephanie Burton, "Sheer Volume of Misinformation Risks Diverting Focus from Fighting Coronavirus", *The Conversation* (29 April 2020), online: <https://theconversation.com/sheer-volume-of-misinformation-risks-diverting-focus-from-fighting-coronavirus-137408>.

[7] Harrison Mantas, "COVID-19 Infodemic Exacerbates Existing Religious and Racial Prejudices" (1 May 2020), online: *Poynter* <https://www.poynter.org/reporting-editing/2020/covid-19-infodemic-exacerbates-existing-religious-and-racial-prejudices/> ("COVID-19 has inflamed fears of outsiders across the globe").

[8] See Leonardo Bursztyn et al, "Misinformation During a Pandemic" (2020) Becker Friedman Institute [working paper] at abstract: "While our findings cannot yet speak to long-term effects, they indicate that provision of misinformation in the early stages of a pandemic can have important consequences for how a disease ultimately affects the population." See also Mian & Khan, "Public Confusion Leaves Citizens Unprepared for Combatting a Public Health Crisis", *supra* note 1 at 2.

[9] See Soroush Vosoughi, Deb Roy & Sinan Aral, "The Spread of True and False News Online" (2018) 359:6380 Science 1141 at 1141, DOI: <10.1126/science.aap9559>, where the authors analyzed millions of social media shares and came to the grim conclusion that "falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information."

[10] Ryan Ko, "Social Media Is Full of Bots Spreading COVID-19 Anxiety. Don't Fall for It" (2 April 2020), online: *Science Alert* <https://www.sciencealert.com/bots-are-causing-anxiety-by-spreading-coronavirus-misinformation>: "These fake accounts are common on Twitter, Facebook, and Instagram. They have one goal: to spread fear and fake news."

[11] Katherine Ellison, "Social Media Posts and Online Searches Hold Vital Clues about Pandemic Spread", *Scientific American* (30 March 2020), online: <https://

it has served as a primary source of news for many in the general public.[12] Indeed, more and more people are turning to social media to keep up-to-date on developments surrounding the pandemic.[13] It has been reported that Twitter had about "12 million more daily users in the first three months of 2020 than in the last three of 2019."[14]

Still, in the context of the "infodemic," social media platforms have been the focus of much of the concern and policy activity.[15] There is some suggestion that the spread of overt misinformation—that is, misinformation provided by known "fake news" sources—on some platforms, such as Facebook, has decreased since the implementation of platform countermeasures, including removing fake accounts and tweaking their algorithm to reduce the reach of debunked articles.[16] But on other platforms, including Twitter, the situation has

www.scientificamerican.com/article/social-media-posts-and-online-searches-hold-vital-clues-about-pandemic-spread/>.

[12] See e.g. Alaa Abd-Alrazaq et al, "Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study" (2020) 22:4 J Medicine Internet Research e19016, DOI: <10.2196/19016>, where the authors analyzed 2.8 million tweets on the pandemic and found tweets on issues such as the source, cause, economic consequences, and treatments and cures, concluding: "Social media provides an opportunity to directly communicate health information to the public."

[13] Jeffrey Gottfried & Elisa Shearer, "News Use Across Social Media Platforms 2016" (16 May 2016), online: *Pew Research Center* <https://www.journalism.org/2016/05/ 26/news-use-across-social-media-platforms-2016/>.

[14] Jon-Patrick Allem, "Social Media Fuels Wave of Coronavirus Misinformation as Users Focus on Popularity, Not Accuracy", *The Conversation* (6 April 2020), online: <https://theconversation.com/social-media-fuels-wave-of-coronavirus-misinformation-as-users-focus-on-popularity-not-accuracy-135179>. See also Vengattil Munsif & Dave Paresh, "Twitter Ad Sales Hit by Coronavirus but Active Users Soar" (23 March 2020), online: *Reuters* <https://www.reuters.com/article/us-health-coronavirus-twitter/twitter-ad-sales-hit-by-coronavirus-but-active-users-soar-idUSKBN21A3HY>.

[15] Ramez Kouzy et al, "Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter" (2020) 12:3 Cureus e7255, DOI: <10.7759/cureus.7255>.

[16] Hunt Allcott, Matthew Gentzkow & Chuan Yu, "Trends in the Diffusion of Misinformation on Social Media" (2019) 6:2 Research & Politics 1 at abstract: "Our results suggest that the relative magnitude of the misinformation problem on Facebook has declined since its peak." See also Paul Resnick, Aviv Ovadya & Garlin Gilchrist, "Iffy Quotient: A Platform Health Metric for Misinformation" (18 October 2018) at 1, online: *School of Information Center for Social Media Responsibility, University of Michigan* <https://csmr.umich.edu/wp-content/uploads/2018/10/UMSI-CSMR-Iffy-Quotient-Whitepaper-810084.pdf>: "there has been gradual improvement in Facebook's Iffy Quotient since mid-2017, with a substantial cumulative impact. […] In 2016 the Iffy sites' share of attention was about twice as high on Facebook as Twitter; now it is 50% higher on Twitter."

worsened.[17] Much of the misinformation about the coronavirus remains unchecked and continues to circulate, especially on Twitter.[18]

Why and how misinformation spreads and has an impact on behaviours and beliefs is a complex and multidimensional phenomenon.[19] There is an emerging rich academic literature on misinformation, particularly in the context of social media.[20] Here, I make no attempt to provide a comprehensive overview of that work. Rather, I focus on two relatively narrow questions: Is debunking an effective strategy; If so, what kind of counter-messaging is most effective? The goal of this article is to bring together relevant empirical research and expert commentary to serve as a resource and guide in the battle against misinformation (hence the heavy referencing) and to stand as a defence of these efforts.[21]

[17] Allcott, Gentzkow & Yu, *supra* note 16.

[18] J Scott Brennen et al, "Types, Sources, and Claims of COVID-19 Misinformation" (7 April 2020) at 1, online: *Reuters Institute for the Study of Journalism, University of Oxford* <https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation>: "On Twitter, 59% of posts rated as false in our sample by fact-checkers remain up." See also Craig Timberg, "On Twitter, Almost 60 Percent of False Claims about Coronavirus Remain Online—Without a Warning Label", *Washington Post* (7 April 2020), online: <https://www.washingtonpost.com/technology/2020/04/07/twitter-almost-60-percent-false-claims-about-coronavirus-remain-online-without-warning-label/>.

[19] Dietram A Scheufele & Nicole M Krause, "Science Audiences, Misinformation, and Fake News" (2019) 116:16 PNAS 7662 at 7662, DOI: <10.1073/pnas.1805871115>: "[W]e show how being misinformed is a function of a person's ability and motivation to spot falsehoods, but also of other group-level and societal factors that increase the chances of citizens to be exposed to correct(ive) information."

[20] See generally Yuxi Wang et al, "Systematic Literature Review on the Spread of Health-related Misinformation on Social Media" (2019) 240:112552 Social Science & Medicine 1 at 1, DOI: <10.1016/j.socscimed.2019.112552>: "Overall, we observe an increasing trend in published articles on health-related misinformation and the role of social media in its propagation." See also Denise-Marie Ordway, "Fake News and the Spread of Misinformation: A Research Roundup" (1 September 2017), online: *Journalist's Resource* <https://journalistsresource.org/studies/society/internet/fake-news-conspiracy-theories-journalism-research/>.

[21] The word "debunking" is less than ideal, as some may feel it fails to capture the need to listen to and engage the public. It can also be associated with a more aggressive, or mocking, approach (a strategy I criticize below). However, in total, with those critiques noted, I still feel it is a good catch-all word that, as defined by Amy Sippitt, can be used to refer to "factual messages which seek to rebut inaccurate factual claims." See Amy Sippitt, "The Backfire Effect: Does It Exist? And Does It Matter for Factcheckers?" (March 2019) at 7, online: *Full Fact* <https://fullfact.org/blog/2019/mar/does-backfire-effect-exist/>.

Is It Worth It?

Let's start with two of most frequently raised arguments *against* vigorously countering the spread of misinformation. One is that correcting misinformation online is simply ineffective. Dumping more science on people has little impact, it is often said, because attempting to correct a misperception can cause individuals to become *more* entrenched in their beliefs. This phenomenon—usually called the "backfire effect"—has received a lot of attention and is often noted whenever there is a call for more individuals to get actively involved in the countering of misinformation. Debunking doesn't work, it is argued.[22]

But how strong is the backfire phenomenon? There are several well-known studies associated with the birth of this concern. Probably the most influential is a study published in 2010 where the researchers explored the impact of corrected news articles that contained a misleading claim by a politician. It was found that "corrections frequently fail to reduce misperceptions among the targeted ideological group" and there were "several instances of a 'backfire effect' in which corrections actually increase misperceptions among the group in question."[23] As a result of this and several other studies, there now seems to be a widely accepted belief that the backfire effect is a dominant phenomenon that makes debunking a near futile exercise.[24]

[22] See, for example, Christian Bokhove, "Beware: Debunking Research Myths Can Backfire on You" (19 July 2019), online: *Tes* <https://www.tes.com/magazine/article/beware-debunking-research-myths-can-backfire-you>.

[23] Brendan Nyhan & Jason Reifler, "When Corrections Fail: The Persistence of Political Misperceptions" (2010) 32 Political Behaviour 303, DOI: <10.1007/s11109-010-9112-2>.

[24] See, for example, Julie Beck, "This Article Won't Change Your Mind", *The Atlantic* (11 December 2019), online: <https://www.theatlantic.com/science/archive/2017/03/this-article-wont-change-your-mind/519093/>; "The Backfire Effect: Why Facts Don't Win Arguments" (15 October 2013), online: *Big Think* <https://bigthink.com/think-tank/the-backfire-effect-why-facts-dont-win-arguments>. See also Erin Brodwin, "Facebook's Covid-19 Misinformation Campaign Is Based on Research. The Authors Worry Facebook Missed the Message" (1 May 2020), online: *StatNews* <https://www.statnews.com/2020/05/01/facebooks-covid-19-misinformation-campaign-is-based-on-research-the-authors-worry-facebook-missed-the-message/>, where it is noted that Facebook's coronavirus misinformation strategy is "designed to avoid what's known as the backfire effect." Why the "backfire effect" gained so much traction is an interesting question on its own, one which is beyond the scope of this piece. But I think that the fact it feels intuitively correct is a big part of its appeal. It is hard to change opinions.

In reality, the backfire effect seems to be a relatively rare occurrence.[25] Indeed, Brendan Nyhan, the lead author of the 2010 study, has noted that their results often have "been overstated and oversold,"[26] in part because their conclusions may be quite context specific.[27] A 2019 comprehensive analysis of the available research concluded that the existing body of evidence—much of it published after the 2010 study—found no backfire effect and that "most recent studies now suggest that generally debunks can make beliefs in specific claims more accurate."[28] For example, a study published in 2019 found that "evidence of factual backfire is far more tenuous than prior research suggests. By and large, citizens heed factual information, even when such information challenges their ideological commitments."[29] Another study from 2019 found that "debunking" works—if done using appropriate strategies (more on that below)—and "no evidence" that "rebutting science denialism in public discussions backfires, not even in vulnerable groups (for example, U.S. conservatives)."[30] To be fair, motivated reasoning (constructing rationales to fit a pre-existing position) and other cognitive biases (for example, confirmation bias) have been shown to influence what information we see online and elsewhere.[31] Still, for many areas of science, at least some research has found that differences in scientific belief are driven mostly by levels of

[25] Indeed, some have gone so far as to call its existence a myth. See, for example, Laura Hazard Owen, "The 'Backfire Effect' Is Mostly a Myth" (22 March 2019), online: *NiemanLab* <https://www.niemanlab.org/2019/03/the-backfire-effect-is-mostly-a-myth-a-broad-look-at-the-research-suggests/>.

[26] See 8 January 2018 tweet by lead author Brendan Nyhan, where he states: "[T]he research findings, including accounts of my own backfire effect paper with @jasonreifler, have often been overstated and oversold" (3 January 2020 at 8:21), online: *Twitter* <https://twitter.com/brendannyhan/status/948544775799607296?lang=en>.

[27] For example, see Sippitt, *supra* note 21 at 10, who notes that the experiment "purposefully covered a highly controversial topic in American politics [WMD in Iraq] where people would have prior beliefs" and as such "it's arguably unsurprising that individuals were unpersuaded by a single news item."

[28] See *ibid* at 5.

[29] Thomas Wood & Ethan Porter, "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence" (2019) 41 Political Behaviour 135.

[30] Philipp Schmid & Cornelia Betsch, "Effective Strategies for Rebutting Science Denialism in Public Discussions" (2019) 3 Nature Human Behaviour 931 at abstract.

[31] For example, see Dan Kahan, "The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It" in RA Scott and SM Kosslyn, eds, *Emerging Trends in the Social & Behavioral Sciences* (Wiley Library Online, 2016), DOI: <10.1002/9781118900772.etrds0417>.

scientific knowledge and not motivated reasoning.[32] So while a backfire effect may occur in some circumstances—this is an area where more research would be helpful—it certainly isn't such a robust and measurable phenomenon that it should stop us from mounting efforts to counter misinformation on social media.

The second and perhaps more challenging critique of correcting and debunking is that it may inadvertently help to spread misinformation.[33] Specifically, there might an "illusory truth" effect.[34] Studies have consistently found that merely exposing people to an idea increases the believability of that idea.[35] In many ways this is how "fake news" works.[36] A study by Gordon Pennycook et al, for

[32] Jonathon McPhetres & Gordon Pennycook, "Science Beliefs, Political Ideology, And Cognitive Sophistication" (2020) at abstract, online: *OSF Preprints* <https://osf.io/ad9v7/>: "We also found little evidence of motivated reasoning; reasoning ability was instead broadly associated with pro-science beliefs. Finally, one's level of basic science knowledge was the most consistent predictor of people's beliefs about science. Results suggest educators and policymakers should focus on increasing basic science literacy and critical thinking rather than the ideologies that purportedly divide people."

[33] This is also often called the backfire effect, though it is a different phenomenon than that described by Nyhan & Reifler in "When Corrections Fail," who coined the phrase. As such, I usually treat them as distinct and refer to this as the "spreading" concern.

[34] Melissa Healy, "Misinformation About the Coronavirus Abounds, but Correcting It Can Backfire", *Los Angeles Times* (8 February 2020), online: <https://www.latimes.com/science/story/2020-02-08/coronavirus-outbreak-false-information-psychology>: "Sometimes the effort to correct misinformation involves repeating the lie. That repetition seems to establish it in our memories more firmly than the truth."

[35] See Jonas De keersmaecker, David Dunning & Gordon Pennycook, "Investigating the Robustness of the Illusory Truth Effect Across Individual Differences in Cognitive Ability, Need for Cognitive Closure, and Cognitive Style" (2020) 46:2 Personality and Social Psychology Bulletin 204. Indeed, this effect can still have an impact even if the information runs counter to an existing knowledge base. See, for example, Lisa K Fazio et al, "Knowledge Does Not Protect Against Illusory Truth" (2015) 144 J Experimental Psychology 993 at 993: "Contrary to prior suppositions, illusory truth effects occurred even when participants knew better."

[36] See, for example, Danielle C Polage, "Making Up History: False Memories of Fake News Stories" (2012) 8:2 Europe's J Psychology 245; Christopher Paul & Miriam Matthews, "The Russian 'Firehose of Falsehood' Propaganda Model: Why It Might Work and Options to Counter It" (2016), online: *RAND* <https://www.rand.org/pubs/perspectives/PE198.html>. I have argued that this is also one reason that celebrities can have such a large impact on the spread of misinformation. See, for example, Timothy Caulfield, "Celebrities like Gwyneth Paltrow Made the 2010s the Decade of Health and Wellness Misinformation", *NBC News* (27 December 2019), online: <https://www.nbcnews.com/think/opinion/

example, found that even a single exposure to misinformation could increase subsequent perceptions of accuracy.[37]

So, does this mean that debunking misinformation and conspiracy theories on social media—which often, of necessity, will include a restatement of the problematic belief—has the potential to do more harm than good? While the speculation about the problem of spreading is rooted in evidence about the possible impact of exposure to misinformation, there does not appear to be much direct empirical evidence that debunking actually has this problematic impact. Indeed, a recent study (still in preprint at time of this writing) explored this exact concern by analyzing whether a debunking of a new piece of misinformation—a not widely known and novel myth or conspiracy theory—led to an increase in beliefs about the claim. They found that corrections that "repeated novel misinformation claims did not lead to stronger misconceptions compared to a control group never exposed to the false claims or corrections."[38] As a result of this finding—which fits with other works on this point[39]—the authors conclude, "it is safe to repeat misinformation when correcting it, even when the audience might be unfamiliar with the misinformation."[40]

The timing of a correction may also be relevant here. Claire Wardle, executive director of an institute dedicated to fighting misinformation, suggests that if you debunk a bit of misinformation too early, you may give it unintended oxygen and allow it to spread further.[41] But once the public awareness of a particular myth, conspiracy

celebrities-gwyneth-paltrow-made-2010s-decade-health-wellness-misinformation-ncna1107501>. See also Mathew Ingram, "Amplifying the Coronavirus Protests", *Columbia Journalism Review* (22 April 2020), online: <https://www.cjr.org/the_media_today/amplifying-coronavirus-protests.php>, where it is noted that less-than-ideal reporting of lockdown protests may have given them more legitimacy than the objective numbers might have suggested was appropriate.

[37] Gordon Pennycook, Tyrone D Cannon & David G Rand, "Prior Exposure Increases Perceived Accuracy of Fake News" (2018) 147:12 J Experimental Psychology: General 1865, DOI: <10.1037/xge0000465>.

[38] Ullrich KH Ecker, Stephan Lewandowsky & Matthew Chadwick, "Can Corrections Spread Misinformation to New Audiences? Testing for the Elusive Familiarity Backfire Effect" (2020) [working paper], DOI: <10.31219/osf.io/et4p3>.

[39] Ullrich KH Ecker et al, "The Effectiveness of Short-Format Refutational Fact-Checks" (2020) 111:1 British J Psychology 36 at 36: "[W]e found no evidence for a familiarity-driven backfire effect."

[40] *Ibid*.

[41] Claire Wardle, "What Role Should Newsrooms Play in Debunking COVID-19 Misinformation?", *Nieman Reports* (8 April 2020), online: <https://niemanreports.org/articles/what-role-should-newsrooms-play-in-debunking-covid-19-mis-

theory, or item of misinformation hits a tipping point—that is, the item is starting to be shared more widely—it is important to vigorously counter. If we wait too long to attempt a correction, it may become increasingly difficult to stop the momentum of the misinformation.[42] As we have seen with issues like the myths surrounding vaccination, once a conspiracy theory gets a strong foothold in the public conscious, it can be difficult to dislodge.

A better interpretation of the existing literature is that while we need to be cognizant of the spreading concern, the evidence is far from definitive and what evidence is available suggests it doesn't often happen. There are, of course, many other challenges associated with efforts to correct misinformation, such as the possibility for a range of additional unintended consequences (for example, general warning tags skewing how people perceive legitimate news).[43] But despite the need for more research, there is nothing in the existing research to suggest debunking is a futile exercise. On the contrary, as we will see, there is a growing body of evidence that tells us correcting

information/>. See also Whitney Phillips, "The Oxygen of Amplification: Better Practices for Reporting on Extremists, Antagonists, and Manipulators Online" (2012), online: *Data & Society* <https://datasociety.net/library/oxygen-of-amplification/>; Susan Benkelmam, "Getting it Right: Strategies for Truth-Telling in a Time of Misinformation and Polarization" (11 December 2019), online: *American Press Institute* <https://www.americanpressinstitute.org/publications/reports/strategy-studies/truth-telling-in-a-time-of-misinformation-and-polarization/>: "Journalists must ask themselves whether a falsehood has become so significant that it needs to be knocked down."

[42] There is some recent evidence to support this view. See e.g. Wasim Ahmed et al, "COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data" (2020) 22:5 J Medicine Internet Research e19458 at abstract: The authors found that "there was a lack of an authority figure who was actively combating such [5g] misinformation" on social media. What is needed, they conclude, is the "combination of quick and targeted interventions oriented to delegitimize the sources of fake information."

[43] John M Carey et al, "The Effects of Corrective Information about Disease Epidemics and Outbreaks: Evidence from Zika and Yellow Fever in Brazil" (2020) 6:5 Science Advances 1 at 9, DOI: <10.1126/sciadv.aaw7449>: "[A] general warning about the presence of fake news has been found to decrease belief in the accuracy of both false and legitimate news headlines." For a study that found the opposite effect, see Gordon Pennycook et al, "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings" (2020) Management Science [forthcoming], DOI: <10.2139/ssrn.3035384>. While placing "fake news" warnings on social media content can have a positive impact, this study found that "the presence of warnings caused untagged headlines to be seen as more accurate than in the control" (at abstract).

misinformation should be viewed as a vitally important science and health policy activity.

What Kind of Counter-Messaging Works?

As with the research on the challenges associated with correcting misinformation, the data surrounding effective debunking strategies is messy and context-dependent. More research on how best to deal with misinformation is clearly needed,[44] but there is little doubt that countering misinformation can have a positive impact.[45] Indeed, silence in the face of misinformation seems likely to be the worst strategy. A 2019 study, for example, found that not responding to misinformation "has a negative effect on attitudes towards behaviours favoured by science."[46] But what kind of social media counter is likely to have the biggest positive result? Below is a list of some of the general themes that have emerged in the research regarding the tone and style of debunking messaging that is relevant to all social media platforms. Here, I focus on the actual content of a social media debunk. Obviously, not every approach will work for every corrective

[44] See Gordon Pennycook & David Rand, "The Right Way to Fight Fake News", *New York Times* (24 March 2020), online: <https://www.nytimes.com/2020/03/24/opinion/fake-news-social-media.html>: "The obvious conclusion to draw from all this evidence is that social media platforms should rigorously test their ideas for combating fake news and not just rely on common sense or intuition about what will work."

[45] For the benefits of debunking in the context of a pandemic, see Toni GLA van der Meer & Yan Jin, "Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source" (2020) 35:5 Health Communications 560 at 560: "Results show that, if corrective information is present rather than absent, incorrect beliefs based on misinformation are debunked and the exposure to factual elaboration, compared to simple rebuttal, stimulates intentions to take protective actions." See generally Nathan Walter & Sheila T Murphy, "How to Unring the Bell: A Meta-Analytic Approach to Correction of Misinformation" (2018) 85:3 Communications Monographs 423 at 436. A meta-analysis of existing data concludes that: "corrective attempts can reduce misinformation across diverse domains, audiences, and designs"; Man-pui Sally Chan et al, "Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation" (2017) 28:11 Psychological Science 1531; Brendan Nyhan et al, "Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability" (2019) Political Behaviour [forthcoming], DOI: <10.1007/s11109-019-09528-x>; Victoria L Rubin, "Deception Detection and Rumor Debunking for Social Media" in L Sloan & A Quan-Haase, eds, *The SAGE Handbook of Social Media Research Methods* (London: SAGE, 2017).

[46] Schmid & Betsch, *supra* note 30 at abstract.

message—a tweet is, after all, just 280 characters. But these evidence-informed general principles can help to maximize the impact of efforts to correct online misinformation.

First, use facts. Despite all the concern regarding the impotence of facts to change minds, most studies have found that providing corrective information can be effective,[47] especially if the alterative explanation—the science-informed facts—fills in the gap in understanding caused by the debunk and (when appropriate and possible) provides a causal explanation.[48] This approach can also nudge people to think more critically generally, which may help to shield them against related forms of misinformation.[49]

Second, provide clear, straightforward, and shareable content.[50] Studies have shown that the use of scientific jargon will cause people to disengage, even if explanatory language is also provided in the text.[51]

[47] Leticia Bode & Emily K Vraga, "In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media" (2015) 65:4 J Communication 619 at 630: "Our experimental evidence suggests that attitude change related to GMOs can be achieved with regard to misperceptions by virtue of exposure to corrective information within social media." See also Emily Falk & Molly Crockett, "You Can Help Slow the Virus if You Talk about it Accurately Online", *Washington Post* (28 April 2020), online: <https://www.washingtonpost.com/outlook/2020/04/28/you-can-help-slow-virus-if-you-talk-about-it-accurately-online/>; *ibid*.

[48] See Walter & Murphy, *supra* note 45 at 436: "[C]orrective messages that integrate retractions with alternative explanations (i.e., coherence) emerge as an effective strategy to debunk falsehoods." See also Briony Swire & Ullrich Ecker, "Misinformation and its Correction: Cognitive Mechanisms and Recommendations for Mass Communication" in Brian G. Southwell, Emily A Thorson & Laura Sheble, eds, *Misinformation and Mass Audiences* (Austin: University of Texas Press, 2018): The alternative explanation effectively plugs the model gap left by the retraction. See also Brendan Nyhan & Jason Reifler, "Displacing Misinformation about Events: An Experimental Test of Causal Corrections" (2015) 2:1 J Experimental Political Science 81.

[49] See Ecker et al, *supra* note 39 at 49: "We can thus conclude that embedding a rebuttal in a fact-oriented context has beneficial implications beyond specific belief reduction, fostering a more sceptical and evidence-based approach to the issue at hand."

[50] Samantha Yammine, "Going Viral: How to Boost the Spread of Coronavirus Science on Social Media", *Nature* (5 May 2020), online: <https://www.nature.com/articles/d41586-020-01356-y>.

[51] See e.g. Hillary C Shulman et al, "The Effects of Jargon on Processing Fluency, Self- Perceptions, and Scientific Engagement" (2020) J Language and Social Psychology 1 at 13: "Jargon can then serve as exclusionary language that disengages meaningful relationships between public and expert communities from forming."

Third, use trustworthy and independent sources. Evidence perceived to be removed from an agenda (and the profit motive) is more likely to be trusted and persuasive.[52] While it can be a challenge to find sources that are trusted by all—there has been a significant erosion in trust in many public institutions[53]—public health authorities and independent scientists still retain a relatively high level of trustworthiness, particularly during times of crisis.[54]

Fourth, if applicable and available, emphasize the scientific consensus.[55] Ideally, this tactic should be accompanied by a recognition that science evolves and, as such, the consensus can change.

[52] Susan T Fiske & Cydney Dupree, "Gaining Trust as Well as Respect in Communicating to Motivated Audiences about Science Topics" (2014) 111:4 PNAS 13593.

[53] Timothy Caulfield, "Now More Than Ever, We Must Fight Misinformation. Trust in Science Is Essential", *The Globe and Mail* (20 March 2020), online: <https://www.theglobeandmail.com/opinion/article-now-more-than-ever-we-must-fight-misinformation-trust-in-science-is>. Not surprisingly, studies have found that debunking has a more modest effect if people view the original source of misinformation favourably. But even in this situation, debunking efforts can help. See Jeongwoo Jang, Eun-Ju Lee & Soo Yun Shin, "What Debunking of Misinformation Does and Doesn't" (2019) 22:6 Cyberpsychology, Behavior, & Social Networking 423 at 426: "Overall, the results showed that when the falsehood of information was exposed, participants became less favorable toward the immediate source who shared the misinformation, but their initial source attitude also moderated their reactions by inducing different attribution processes." For another commentary on the impact of low trust, see Mike Caulfield, "Cynicism, Not Gullibility, Will Kill Our Humanity" (27 November 2018), online: *Hapgood* <https://hapgood.us/2018/11/27/cynicism-not-gullibility-will-kill-our-humanity/>.

[54] See Pew Research Centre, "Public Holds Broadly Favorable Views of Many Federal Agencies, Including CDC and HHS" (9 April 2020), online: *Pew Research Centre* <https://www.people-press.org/2020/04/09/public-holds-broadly-favorable-views-of-many-federal-agencies-including-cdc-and-hhs/>: "Currently, 79% of U.S. adults express a favorable opinion of the CDC…"; Hannah Fingerhut, "AP-NORC Poll: High Use, Mild Trust of News Media on COVID-19", *Associated Press* (30 April 2020), online: <https://apnews.com/4e2a20bd01bd2352009c3281b657375d>: "Americans are especially likely to trust information about the coronavirus that comes from the CDC or from personal health care providers," See van der Meer & Jin, *supra* note 45 at 560, where it is summarized that during times of crisis "government agency and news media sources are found to be more successful in improving belief accuracy compared to social peers."

[55] See Sander L van der Linden, Chris E Clarke & Edward W Maibach, "Highlighting Consensus among Medical Scientists Increases Public Support for Vaccines: Evidence from a Randomized Experiment" (2015) 15:1207 BMC Public Health; Jeremy D Sloane & Jason R Wiles, "Communicating the Consensus on Climate Change to College Biology Majors: The Importance of Preaching to the Choir" (2020) 10:2 Ecology and Evolution 594; Sander L van der Linden et al, "The Scientific Consensus on Climate Change as a Gateway Belief: Experimental Evidence" 10:2 PLoS ONE e0118489, DOI: <10.1371/journal.pone.0118489>; and Sander L van der Linden, "Why Doctors Should Convey the Medical Consensus

Fifth, be nice and be authentic. Research has found that an aggressive language style is perceived to be both less credible and less trustworthy.[56] Don't shame, ridicule, or marginalize members of the public who are looking for answers (though I have less patience for those pushing bunk for profit, brand enhancement, and ideological spin).[57] In addition, messaging that comes from someone who is seen to be a unique and authentic individual—that is, not just a talking head associated with an institution—can also enhance trust, credibility, and the persuasiveness of the message.[58]

Sixth, consider using a narrative. Humans are wired to respond to stories.[59] Indeed, there is some evidence that an engaging anecdote can overwhelm our ability to think scientifically.[60] This is one reason that testimonials are such an effective strategy for marketing unproven therapies.[61] But a narrative can also be used to convey science—and information about critical thinking and the scientific process[62]—in a way that is compelling and memorable.[63]

on Vaccine Safety" (2016) 21:3 Evidence Based Medicine 119, DOI: <10.1136/ebmed-2016-110435>.

[56] See Lars König & Regina Jucks, "Hot Topics in Science Communication: Aggressive Language Decreases Trustworthiness and Credibility in Scientific Debates" (2019) 28:4 Public Understanding of Science 401. See also Fisk & Dupree, *supra* note 52.

[57] Anand Ram, "How to (Tactfully) Discourage Spread of False Pandemic Information", *CBC News* (19 April 2020), online: <https://www.cbc.ca/news/canada/covid-19-misinformation-rumour-1.5532302>, where misinformation expert Claire Wardle notes the value of being empathetic and using words that "put yourself in the same perspective."

[58] See Lise Saffran et al, "Constructing and Influencing Perceived Authenticity in Science Communication" (2020) 15:1 PLoS ONE e0226711; Sara Reardon, "Adding a Personal Backstory Could Boost Your Scientific Credibility with the Public", *Nature Career News* (2020), DOI: <10.1038/d41586-020-00857-0>.

[59] Michael F Dahlstrom, "Using Narratives and Storytelling to Communicate Science with Nonexpert Audiences" (2014) 111:4 PNAS 13614.

[60] Fernando Rodriguez et al, "Examining the Influence of Anecdotal Stories and the Interplay of Individual Differences on Reasoning" (2016) 22:3 Thinking & Reasoning 274 at 274: "[A]necdotal stories decreased the ability to reason scientifically even when controlling for education level and thinking dispositions."

[61] Bethany Hawke et al, "How to Peddle Hope: An Analysis of YouTube Patient Testimonials of Unproven Stem Cell Treatments" (2019) 12:6 Stem Cell Reports 1186.

[62] See Michael F Dahlstrom & Dietram A Scheufele, "(Escaping) the Paradox of Scientific Storytelling" (2018) 16:10 PLoS Biology e2006720: "[N]arratives might have most of their power not in conveying facts or building excitement but in rebuilding the foundation of understanding scientific reasoning."

[63] For an overview of the evidence on point, see Timothy Caulfield et al, "Health Misinformation and the Power of Narrative Messaging in the Public Sphere"

Seventh, emphasize the gaps in logic and the flawed strategies used by those pushing misinformation. Several studies have found that using rational arguments, such as highlighting the rhetorical tools used to spread misinformation (for example, relying on conspiracy theories, misrepresenting risks, using false "experts"), can be an effective debunking strategy.[64]

Eighth, make the facts the hook, not the misinformation. While the evidence about whether debunking can inadvertently spread misinformation is mixed, it makes sense to frame debunking in a manner that makes the correct information—not the misinformation, myth, or conspiracy theory—the memorable part of the messaging.[65] Make sure the misinformation is clearly flagged as wrong so the debunk is the key takeaway.

Finally, the audience should be the general public, not the hard-core believer. This should be the case even if the debunk is triggered by information circulated by a hard-core believer or someone who is pushing misinformation for personal gain.[66] It is difficult to change the mind of someone who is heavily invested in a particular myth or conspiracy theory. As noted by the WHO, the probability of changing a vocal science denier is extremely low.[67] For this reason,

(2019) 2:2 Can J Bioethics 52.

[64] See Schmid & Betsch, *supra* note 30; Stephan Lewandowsky & John Cook, *The Conspiracy Theory Handbook* (Fairfax: George Mason University, 2020); Gábor Orosz et al, "Changing Conspiracy Beliefs through Rationality and Ridiculing" (2016) 7:1525 Frontiers in Psychology 8: "[U]ncovering arguments regarding the logical inconsistencies of CT beliefs can be an effective way to discredit them."

[65] Some have called this the "truth sandwich" strategy. See Benkelmam, *supra* note 41 at sum: "There are a number of strategies for reporting on falsehoods without amplifying them. One is the 'truth sandwich,' which involves stating a true fact, then the falsehood, then the true fact again." While this approach makes sense, once again there isn't that much direct empirical evidence on point. And there is some research that suggests order may not be that significant. See Evan R Anderson, William S Horton & David N Rapp, "Hungry for the Truth: Evaluating the Utility of 'Truth Sandwiches'" (July 2019), online: *ResearchGate* <www.researchgate.net/publication/334491502_Hungry_for_the_Truth_Evaluating_the_Utility_of_Truth_Sandwiches_as_Refutations>, where it was found that "the truth sandwich structure did not significantly affect the likelihood of readers' endorsing false claims relative to a more typical refutation structure."

[66] I will often use a pop culture moment—the spread of misinformation by a celebrity, for example—as an opportunity to create sharable content about science and the problems associated with the spread of health misinformation.

[67] World Health Organization, "Best Practices Guidance: How to Respond to Vocal Vaccine Deniers in Public" (Copenhagen: Regional Office for Europe of the World Health Organization, 2016): "Rule 1: The general public is your target audience, not the vocal vaccine denier."

the correct information should be framed as if the general public is the audience.

Empowering Users

Fighting the spread of misinformation will, of course, require more than just carefully crafted debunks on social media. We need to come at this issue from every angle.[68] We need, for instance, social media platforms to adopt evidence-informed strategies that will both remove the most harmful content and heighten user vigilance. Studies have found, for example, that the use of warning tags—such as those "rated false"—on social media posts can be an effective strategy to inform the public about potential problems with accuracy with specific content.[69] And we need a more robust policy response against individuals who are pushing unproven products and ideas on social media platforms in a manner that infringes existing laws and regulations.[70]

Perhaps the most important strategy will be to empower people with the tools necessary to be more critical consumers of information. This should incorporate teaching both critical thinking skills and media literacy,[71] including inoculating (or "pre-bunking") people

[68] See e.g. Kate Starbird, "Disinformation's Spread: Bots, Trolls and All of Us" (2019) 571 Nature World View 449, DOI: <10.1038/d41586-019-02235-x>: "But effective disinformation campaigns involve diverse participants; they might even include a majority of 'unwitting agents' who are unaware of their role."

[69] Katherine Clayton et al, "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media" (2019) Political Behaviour at abstract, " … indicate that false headlines are perceived as less accurate when people receive a general warning." While warning tags seem to have a role to play, they need to be deployed sensibly. Research has found, for example, that general warnings telling readers to beware of misinformation can have an unintended spillover of effect of decreasing "belief in the accuracy of true headlines…" Pennycook et al, *supra* note 43 highlights that using warning tags can lead to an inappropriate implication that posts without warnings are *more* accurate. See also Melanie Freeze et al, "Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect" (2020) Political Behaviour, DOI: 10.1007/s11109-020-09597-3>.

[70] For an example of regulatory action, see Health Canada, *Health Products that Make False or Misleading Claims to Prevent, Treat or Cure COVID-19 May Put Your Health at Risk,* (Advisory RA-72659) (Ottawa: Health Canada, 27 March 2020); Federal Trade Commission, Press Release, "FTC Sends 45 More Letters Warning Marketers to Stop Making Unsupported Claims That Their Products and Therapies Can Effectively Prevent or Treat COVID-19" (7 May 2020).

[71] See e.g. Michelle A Amazeen & Erik P Bucy, "Conferring Resistance to Digital Disinformation: The Inoculating Influence of Procedural News Knowledge"

against misinformation[72] and simply reminding them to think about accuracy before sharing.[73] A growing body of literature has found that, in general, people want to be accurate and want to share only factual material.[74] Most users do not fall for or share misinformation due to a malevolent agenda or, even, a partisan bias.[75] If we can nudge people to think about accuracy before they share social media content, we may be able to have a significant impact on the spread of misinformation.[76] A 2020 study that specifically looked at misinformation in the context of the coronavirus found exactly this effect, concluding

(2019) 63:3 J Broadcasting & Electronic Media 415 at 429: "[A]dditional educational campaigns to inform citizens about mainstream news media operations could yield significant benefits." See also Viren Swami et al, "Analytic Thinking Reduces Belief in Conspiracy Theories" (2014) 133:3 Cognition 572.

[72] See e.g. Jon Roozenbeek & Sander van der Linden, "The New Science of Prebunking: How to Inoculate against the Spread of Misinformation" (7 October 2019), online (blog): *BMC On Society* <http://blogs.biomedcentral.com/on-society/2019/10/07/the-new-science-of-prebunking-how-to-inoculate-against-the-spread-of-misinformation/>; Jon Roozenbeek & Sander van der Linden, "Fake News Game Confers Psychological Resistance against Online Misinformation" (2019) 5:65 Palgrave Communications at abstract, DOI: 10.1057/s41599-019-0279-9>: "We provide initial evidence that people's ability to spot and resist misinformation improves after gameplay [which is teaching about misinformation], irrespective of education, age, political ideology, and cognitive style."

[73] Bence Bago, David G Rand & Gordon Pennycook, "Fake News, Fast and Slow: Deliberation Reduces Belief in False (But Not True) News Headlines" J Experimental Psychology: General, Advance online publication, online: *NCBI* <https://www.ncbi.nlm.nih.gov/pubmed/31916834> at abstract: "Our data suggest that, in the context of fake news, deliberation facilitates accurate belief formation and not partisan bias."

[74] Emma Young, "Most People Who Share 'Fake News' Do Care About the Accuracy of News Items—They're Just Distracted" (16 January 2020), online: *Research Digest (The British Psychological Society)* <https://digest.bps.org.uk/2020/01/16/most-people-who-share-fake-news-do-care-about-the-accuracy-of-news-items-theyre-just-distracted/>.

[75] Gordon Pennycook & David G Rand, "Lazy, Not Biased: Susceptibility to Partisan Fake News is Better Explained by Lack of Reasoning Than By Motivated Reasoning" (2019) 188 Cognition 39 at abstract: "Our findings therefore suggest that susceptibility to fake news is driven more by lazy thinking than it is by partisan bias per se—a finding that opens potential avenues for fighting fake news."

[76] See e.g. Lisa Fazio, "Pausing to Consider Why a Headline is True or False Can Help Reduce the Sharing of False News" (10 February 2020), online: *Misinformation Review* <https://misinforeview.hks.harvard.edu/article/pausing-reduce-false-news/>: "This research suggests that forcing people to pause and think can reduce shares of false information"; Gordon Pennycook et al, "Understanding and Reducing the Spread of Misinformation Online" (25 November 2019) at abstract [working paper], online: <https://psyarxiv.com/3n9u8/>: "we find that subtly inducing people to think about the concept of accuracy increases the quality of the news they share."

that “nudging people to think about accuracy is a simple way to improve choices about what to share on social media.”[77]

Conclusion

There is a growing body of research on both the phenomenon of online misinformation and the best way counter it. While the data remain complex and, at times, contradictory, there is little doubt that efforts to correct misinformation are worthwhile. In fact, fighting the spread of misinformation should be viewed as a critical health and science policy priority.

Citation: Timothy Caulfield, “Does Debunking Work? Correcting COVID-19 Misinformation on Social Media” in Colleen M. Flood, Vanessa MacDonnell, Jane Philpott, et al., eds., Vulnerable: The Law, Policy and Ethics of COVD-19 (University of Ottawa Press, 2020) 183-200.

[77] Gordon Pennycook, “Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy Nudge Intervention” (2020) [working paper], online: <https://psyarxiv.com/uhbk9/>.