Leveraging Psychological Heuristics to Combat Disinformation

[A behaviorally informed proposal produced for Carnegie Mellon University]

Author: Jessica Young
Date: 11/29/2018

DISINFORMATION & PRIOR POLICY APPROACHES

The spread of disinformation online has become rampant[i], and with it, the issue that content consumers are unable to discern the "real" from the "fake." A 2016 post-election survey by Ipsos found "fake news headlines fool American adults about 75 percent of the time" and "'fake news' was remembered by a significant portion of the electorate and… seen as credible."[ii] Similarly, Stanford research found even among highly educated students, the majority failed to detect false or biased news.[iii]. Despite these findings, Pew Research shows over 80% of Americans feel "somewhat" or "very confident" they can in fact distinguish fabricated from factual news.[iv] Such research stresses the need for policy interventions that help individuals become aware of, and critically assess, content validity.

To date, we have seen four key types of policy solutions adopted: content removal or demotion; labels for questionable content; referrals to factual content; and media literacy education. Each of these has its own limitations. First, a heavy-handed approach of automatic content removal meets objections of censorship. A subtler approach using AI to minimize disinformation's reach[v] [vi] (or burying it below known factual pieces[vii]) minimizes the harmful effects of exposure, but it fails to build an audience that can critically assess veracity. Second, labeling questionable content with "disputed news" tags, as several media platforms have done, achieves limited results. Yale research found "'disputed' tags made participants just 3.7 percentage points more likely to correctly judge headlines as false."[viii] This shows that pointing to potential inaccuracies may not suffice in a "fake news" era where nearly all content is disputed by *someone*.[ix] Third, waiting until after readers have consumed disinformation to direct them to factual content (or show "related articles") will achieve limited efficacy due to confirmation bias. These strategies have reduced the sharing of disinformation to some extent but not click-through rates[x], and research shows "related articles" are perceived more negatively when they counter pre-existing beliefs.[xi] Finally, while media literacy programs have recently been adopted from California to Nigeria, randomized control trial (RCT) results analyzing their efficacy have yet to be published. Despite receiving an education on best practices for discernment, participants will still face barriers when content evokes "PAIN" triggers (e.g. disinformation on topics that are personal, involve abrupt changes, the immoral, and the now/present)[xii]; in-group versus out-group fears; and empathy gap biases. For instance, despite WhatsApp media literacy ads in India, mob violence ensued when disinformation stoked primal fears.[xiii]

BEHAVIORAL CONSIDERATIONS

An intervention building on the above to attain improved results will need to account for the primary cognitive biases and psychological considerations at play. First and foremost, the policy should seek to intervene before disinformation has been consumed. Given confirmation bias, it will be much more difficult to discount the credibility of information after it has been internalized. In some cases, the utility gained from confirming one's priors could exceed utility gained from knowing the truth.[xiv] Further, case studies from Myanmar, India, and the US show ex-post attempts to debunk disinformation have led to surges in disinformation sharing and even mob violence.[xv] Empathy gap bias also underscores the

need to intervene at the point of exposure; even if viewers anticipate consuming news according to best practices, they may act differently in the moment – especially if the PAIN factors mentioned above are triggered. A successful intervention should also limit the cognitive burden placed on individuals to personally research and verify each news item appearing in their feeds. Finally, the intervention should be designed bearing in mind current political polarization and current levels of trust in mainstream media which hover as low as 20%.[xvi]

PROPOSED INTERVENTION

I propose testing a content management policy requiring each news article be served with a color-coded spectrum that signals the piece's veracity to the viewer. Deeper levels of confidence would be invoked with deep green, while deep red would invoke a conditioned response to stop or proceed with caution. Behavioral research shows traffic-light-colored ratings have driven beneficial behavior in active choice environments[xvii] and displaying standard and comparative information decreases cognitive burden.[xviii] To account for the viewer's potentially distrustful state, those who wished to learn more could hover over the displayed color to see a selection of factors contributing to the determination (much like Gmail's Boomerang app does[xix] to inspire confidence and gain user trust in AI-generated ratings). Factors could include: content source/diffusion by a bot or unverified account, location/IP address of origin, historical counts of account/site disinformation, trending disinformation topic, or flags by third parties such as PolitiFact and/or members of one's trusted peer network). This would serve to veer away from polarized reactions to "political" or biased assessments and allow users to assess objective criteria detected behind the scenes that would interest individuals concerned with manipulation - regardless of political affiliation. Importantly, the spectrum would display simultaneously with headlines - before disinformation is internalized, or further shared.

Benefits would include the following. This intervention would potentially deter viewers from opening disinformation flagged in deep red, or at least prime them to approach it with skepticism and a critical eye. This embraces a libertarian paternalism approach to disinformation that permits active choice in consumption, but with a visually-cued heuristic for cautious assessment and/or quick scanning for the most reliable content. It also simultaneously serves an educational purpose by displaying the key criteria that are important for veracity upon hovering; this would function much like media literacy efforts, but while viewers are in a present, actionable moment.

IMPLEMENTATION & IMPACT

Prior to deploying this intervention, the hover factors to be displayed should be pre-tested to see which combinations are perceived as the most powerful/credible among a diverse audience representative of US social media (e.g. Facebook) users. Pre-testing should also assess the optimal quantity of factors to display that will achieve this aim while still limiting information overload. Ideally, such an RCT would then be deployed on a platform

such as Facebook given roughly half of American adults consume news via this platform[xx]. With (US) active account users as the randomization unit, A/B testing would serve the color-coded display to the treatment group while the control received Facebook's current display of "related articles." Efficacy should be measured by the percent of users in each group who (a) click and (b) share the false content served. If click rates see a statistically significant reduction and/or share rates are reduced relative to the control, the color-coded option should be adopted. Internal validity would be high as the implementation would mirror typical conditions for Facebook user interface changes and responses; external validity would also be high for similar social media platforms in the US but limited for platforms in diverse international contexts. If this intervention did prove successful, a follow-up study would be recommended to evaluate if among the treated there was an uptick in flagging content that was later verified to be false and removed. This would provide further evidence of more informed, critical media consumers – an audience that could be incentivized to assist platforms in detecting disinformation given contexts where AI detection is still limited.[xxi]

This intervention is feasible as Facebook runs extensive A/B testing on a continual basis, and because AI is capable of detecting the hover factors at scale. Costs for designing the display would be minimal and benefits significant if it was successful; this is evident given the high cost to society of internalizing manipulative disinformation and the large investments platforms like Facebook are pouring into content moderation solutions.[xxii] Potential backfiring might occur due to active information avoidance (e.g. if seeing a 'red' rating by a favorite news source would cause a welfare loss) or if users actively sought out those labeled red. However, the latter will likely be minimized via engrained responses to the heuristic, and this would likely be a small subset of the population that would seek actively seek such content regardless. In addition to the above, unintended results could occur due to misinterpretation by color-blind users but this could be minimized through accessible design.

---

[i] Edkins, B. 2016. Americans Believe They Can Detect Fake News. Studies Show They Can't. *Forbes*, December 20, 2016.

[ii] West, D. 2017. How to Combat Fake News and Disinformation. *Brookings*, December 18, 2017.

[iii] Funke, D. 2018. Study: Fake News Is Making College Students Question All News. *Poynter*, October 16, 2018.

[iv] Barthel, M., Mitchell, A., and Holcomb, J. 2018. Many Americans Believe Fake News Is Sowing Confusion. *Pew Research*, April 26, 2018.

[v] BBC. 2017. Google Search Changes Tackle Fake News and Hate Speech. *BBC News*, April 25, 2017.

[vi] Ingram, D. 2017. Facebook Changes Algorithm to Curb 'Tiny Group' of Spammers. *Reuters*, June 30, 2017.

[vii] Funke, D. 2018. A Guide to Anti-misinformation Actions around the World. *Poynter*, October 31, 2018.

[viii] Marshall, G. 2014. Don't Even Think About It: Why Our Brains Are Wired to Ignore Climate Change. August 19, 2014.

[ix] Ibid.

[x] Ong, T. 2017. Facebook Found a Better Way to Fight Fake News. *The Verge*, December 21, 2017.

[xi] Poynter. 2018. Related Links on Facebook Could Help Correct Misinformation. *Poynter*, November 29, 2018.

[xii] Marshall, G. 2014. Don't Even Think About It: Why Our Brains Are Wired to Ignore Climate Change. August 19, 2014.

[xiii] Saldanha, A., Rajput, P., and Hazare, J. 2018. Child-Lifting Rumours: 33 Killed In 69 Mob Attacks Since Jan 2017. *IndiaSpend*, July 9, 2018.

[xiv] Allcott, H., and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31, 211-236.

[xv] Young, J., Swamy, P., and Danks, D. 2018. Beyond AI: Responses to Hate Speech and Disinformation.

[xvi] Allcott, H., and Gentzkow, M. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31, 211-236.

[xvii] Downs, J.S., Wisdom, J., and Loewenstein, G. 2015. Helping consumers use nutrition information: Effects of format and presentation. *American Journal of Health Economics*, 1(3), 326-344.

[xviii] Loewenstein, G., Sunstein, C., and Golman, R. 2013. Disclosure: Psychology Changes Everything. SSRN Electronic Journal. 6. 10.2139/ssrn.2312708.

[xix] Google. 2018. Boomerang for Gmail: Scheduled Sending and Email Reminders.

[xx] Edkins, B. 2016. Americans Believe They Can Detect Fake News. Studies Show They Can't. *Forbes*, December 20, 2016.

[xxi] Young, J., Swamy, P., and Danks, D. 2018. Beyond AI: Responses to Hate Speech and Disinformation.

[xxii] Ibid.