IPS Working Papers

FAKE NEWS, FALSE INFORMATION AND MORE:

COUNTERING HUMAN BIASES

CAROL SOON
and
SHAWN GOH

September 2018

IPS Working Papers No. 31

Institute of
Policy Studies

About Institute of Policy Studies (IPS)

The Institute of Policy Studies (IPS) was established in 1988 to promote a greater awareness of policy issues and good governance. Today, IPS is a think-tank within the Lee Kuan Yew School of Public Policy (LKYSPP) at the National University of Singapore. It seeks to cultivate clarity of thought, forward thinking and a big-picture perspective on issues of critical national interest through strategic deliberation and research. It adopts a multi-disciplinary approach in its analysis and takes the long-term view. It studies the attitudes and aspirations of Singaporeans which have an impact on policy development and the relevant areas of diplomacy and international affairs. The Institute bridges and engages the diverse stakeholders through its conferences and seminars, closed-door discussions, publications, and surveys on public perceptions of policy.

IPS Working Papers No. 31

FAKE NEWS, FALSE INFORMATION AND MORE:

COUNTERING HUMAN BIASES

CAROL SOON

Senior Research Fellow

Institute of Policy Studies

carol.soon@nus.edu.sg

and

SHAWN GOH

Research Assistant

Institute of Policy Studies

shawn.goh@nus.edu.sg

September 2018

CONTENTS

Sections

FAKE NEWS, FALSE INFORMATION AND MORE:

COUNTERING HUMAN BIASES

1. INTRODUCTION

Despite the various measures adopted by the public, private and people sectors in the past 18 months to counter fake news and various types of disinformation, concerns among the public remain high. According to the 2018 Reuters Institute Digital News Report, over half (54%) agree or strongly agree that they were concerned about what is real and fake on the Internet. This was highest in countries like Brazil (85%), Spain (69%), and the US (64%) where politics are polarised and social media use is high. The countries that saw less concern among the public are Germany (37%) and the Netherlands (30%) where recent elections were largely untroubled by concerns over fake content.[1] In Singapore, a BBC Global News Survey conducted in 2017 found that 84% of respondents were concerned over fake news, and 59% found it difficult to distinguish between real and fake news online.[2]

[1] Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., & Nielsen, R. K. (2018). Reuters Institute digital news report 2018. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf.

[2] Chua, J. (2017, May 25). Most S'poreans concerned about fake news: BBC study. *TODAY*. Retrieved from https://www.todayonline.com/singapore/most-sporeans-concerned-about-fake-news-bbc-study.

In an earlier report published last year, titled "What Lies Beneath the Truth: A Literature Review on Fake News, False Information and More", as well as in our Written Representation to the Select Committee on Deliberate Online Falsehoods, we presented the various factors that contribute to the problem of false information (which includes fake news, misinformation and disinformation). One of the factors is human factors — individuals' limited time, cognitive resources and motivations make it difficult for them to sort truth from falsehoods, fact from fiction. This problem is aggravated by the vast amount of content published and shared online, and is more pronounced when it comes to complex topics such as scientific findings and politics.[3] Given the deluge of information that confronts individuals every day, people tend to rely on heuristics and social cues, such as perceived trustworthiness and attractiveness of the information source, their past experiences, as well as what others think, to assess the information they encounter online. As a result, individuals typically do not interpret information in a rational, neutral and objective manner, but succumb to their own biases when processing information instead.[4] [5] [6]

[3] Swire, B., & Ecker, U. K. H. (2018). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. In *Misinformation and mass audiences* (pp. 195–211). Austin, Texas: University of Texas Press.

[4] Lau, R. R., & Redlawsk, D. P. (2001). Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science*, *45*(4), 951-971.

[5] Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, *60*(3), 413-439.

[6] Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

In addition to influencing how people assess information, human biases also have an effect on the debunking of false information. Research has shown that false information continues to influence people's memory, reasoning and decision-making even after people have received corrections. Furthermore, debunking efforts may at times reinforce the false information one seeks to correct, and inaccurate beliefs that are formed can be firmly entrenched in people's minds.[7] [8]

Since the publication of the abovementioned report, new research and developments in countering false information have emerged. This working paper focuses on the various strategies and interventions that have been used to overcome human biases, as well as their limitations. In the next section, we identify key human biases that have been shown to affect information processing, evaluation and debunking efforts. Following which, we present three approaches, namely technocognitive solutions, effective communication, and education and cultivating literacy, and review recent measures taken on these fronts to counter falsehoods.

[7] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

[8] Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420-1436.

2. COGNITIVE BIASES AND OTHER HUMAN FACTORS

As mentioned, people tend to rely on cognitive shortcuts to assess the information they encounter online, often making them vulnerable to their biases as a result. In this section, we identify key human biases and explain how they affect information processing, evaluation and debunking efforts.[9]

2.1 Confirmation bias and motivated reasoning

Confirmation bias refers to people's tendency to embrace information consistent with their pre-existing beliefs and reject information that contradicts them. The more consistent a piece of new information is with what an individual already assumes to be true, the more likely this information will be accepted as true, even if it is false. From a cognitive-consistency perspective, this happens because less effort, motivation and cognitive resources are required for an individual to assimilate a piece of information that has logical compatibility with what he or she already believes to be true.[10] [11] [12]

[9] Note: The biases presented in this section are by no means an exhaustive list of factors that affect people's information consumption behaviours. This paper presents those that are more salient and relevant to the problem of falsehoods.

[10] Holton, B., & Pyszczynski, T. (1989). Biased information search in the interpersonal domain. *Personality and Social Psychology Bulletin*, *15*(1), 42-51.

[11] Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*(6), 636-653.

[12] Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220.

This phenomenon is perhaps best demonstrated in a classic experiment conducted by Lord, Ross and Lepper (1979) at Stanford University, where participants with either strong pro- or anti-death penalty views were presented with evidence that supported either the continuation or abolition of the death penalty i.e., evidence either confirmatory or disconfirmatory of their pre-existing attitudes towards the death penalty. The study found that the nature of the evidence presented did not matter as much as people's pre-existing beliefs.[13] [14]

Besides having the tendency to embrace information consistent with their pre-existing beliefs, people may even actively seek information to justify their desired beliefs and reduce cognitive dissonance, thus allowing them to believe what they choose to. This process is known as motivated reasoning.[15] [16] [17] Research has shown that people also engage in motivated reasoning to preserve self-identity and group identity (e.g., political identity), especially when they come into contact

[13] People who were against the death penalty rated anti-death penalty evidence as highly convincing and rated pro-death penalty evidence as unconvincing. The reverse was true for those who were supportive of the death penalty. Furthermore, those who were against the death penalty to begin with became more anti-death penalty when shown pro-death penalty evidence (and vice versa), suggesting evidence of a "backfire effect" as well (see *Section 2.6: Worldview backfire effect*).

[14] Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098-2109.

[15] Festinger, L. (1956). *When prophecy fails: A social and psychological study of a modern group that predicted the destination of the world*. New York City, New York: Harper.

[16] Hastorf, A. H., & Cantril, H. (1954). They saw a game; a case study. *Journal of Abnormal and Social Psychology*, *49*(1), 129-134.

[17] Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480-498.

with counter-evidence.[18] [19] [20] In other words, motivated reasoning explains why people accept and reject false information differently, depending on how the information resonates with their deeply held beliefs and identities.

Confirmation bias and motivated reasoning also underscore the formation of online echo chambers and filter bubbles because of people's tendency to form groups based on common interests or beliefs. This is explained further in *Section 2.7.2: Homophily in social networks*.

2.2 Continued influence effect

Research has found that corrections are rarely fully effective because despite being given a correction and acknowledging it, people often continue to rely partially on information that they know is false. Known as the continued influence effect, this phenomenon has been observed across various studies.[21] [22]

[18] Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, *38*, 127-150.

[19] Sunstein, C. R. (2014). *On rumors: How falsehoods spread, why we believe them, and what can be done*. Princeton, New Jersey: Princeton University Press.

[20] Sunstein, C. R. (2014). *Conspiracy theories and other dangerous ideas*. New York City, New York: Simon and Schuster.

[21] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

[22] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369.

The continued influence effect is probably best demonstrated in classic experiments by Ross, Lepper and Hubbard (1975) at Stanford University, [23] and in another study that exposed people to false information about a fictitious warehouse fire and then gave corrections clarifying that parts of the story were incorrect. In both studies, researchers found that despite remembering and accepting the corrections given, people still referred to the false information when answering questions about the study.[24] [25]

Research has also found that continued influence effects, also known as “belief echoes”, can be created either affectively or deliberatively. Affective belief echoes are created when a piece of false information has a larger impact than its correction, especially when it is vivid and produces a strong affective charge. On the other hand, deliberative belief echoes are created through a conscious chain of reasoning where an

[23] Researchers presented two groups of participants with pairs of suicide notes and asked them to distinguish the genuine notes from the fake ones. After completing this activity, researchers told one group of participants that they have correctly identified 24 out of 25 suicide notes, but told the other group that they have only correctly identified 10 out of 25 notes. However, this was a deception for both groups. Researchers subsequently revealed to each group of participants that they were no more discerning than participants from the other group, and asked participants to estimate how many genuine suicide notes they thought they really identified. Interestingly, participants who were initially told that they accurately identified 24 out of 25 notes consistently gave higher estimates of “correctly identified notes” than participants who were initially told that they only correctly identified 10 out of 25 notes, suggesting that participants continued to rely on a piece of information (either that they “fared better” or “fared worse”) which they had been told was false.

[24] Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420-1436.

[25] Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology*, *32*(5), 880-892.

individual deliberates on a correction and reasons that the existence of a piece of false information makes it more likely that other pieces of negative information are true. The different processes of how belief echoes are created have an implication on how partisanship magnifies the continued influence effect – partisanship plays a smaller role in amplifying affective belief echoes because they are not created through conscious deliberation. Instead, partisanship plays a bigger role in amplifying deliberative belief echoes because individuals consciously recall the correction, and the inferences they draw are thus influenced by their political attitudes.[26] [27]

Regardless of the processes that create belief echoes, the continued influence effect suggests that impressions, once formed, may be difficult to change or correct. Thus, once a piece of information is out in the open, it may be too late to retract or blunt its influence. While fact checking efforts can meticulously point out falsehoods, they may be limited in their ability to undo the beliefs and impressions created when people first encounter the false information.

[26] Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*(3), 460-480.

[27] Ecker, U. K., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory and Cognition*, *38*(8), 1087–1100.

2.3 Illusory truth effect

Research has found that repeated exposure to false information increases its likelihood of being accepted as true. Scholars first discovered this illusory truth effect when they observed that people consistently rated repeated statements as truer than new statements.[28] [29] Recent studies looking at fake news stories circulated on social media during the 2016 US Presidential Election also found that people's belief in fake news stories became stronger as their exposure to them increased in frequency.[30] This occurred even for stories that contained highly implausible and partisan statements.[31] Such an effect also happened with people who were knowledgeable about a topic.[32] In other words, repeated exposure to a falsehood increases its chances of being accepted as true because people rely on familiarity as a heuristic in their cognitive processing, and repeated false information feels more familiar and truer.

[28] Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107-112.

[29] Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*(4), 446.

[30] Goodwin-Ortiz de Leon, C. (2017). Fake news on social media: Illusory truth and the 2016 presidential election [Bachelor thesis]. Retrieved from https://www.researchgate.net/publication/316418350_Fake_News_On_Social_Media_Illusory_Truth_and_the_2016_Presidential_Election.

[31] Pennycook, G., Cannon, T. D., & Rand, D. G. (2017). Prior exposure increases perceived accuracy of fake news [Research paper]. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2958246.

[32] Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993-1002.

2.4 Familiarity backfire effect

Related to the illusory truth effect is the familiarity backfire effect, which operates on a similar principle — increased exposure and familiarity increases the likelihood of a piece of information being accepted as true. However, the familiarity backfire effect is usually observed in the context of fact checking and retractions.

In order to debunk a falsehood, one will typically mention it. However, this repetition of the falsehood may also inadvertently increase people's familiarity with it.[33] [34] For instance, research has found that people who were shown a flyer that debunked common myths about flu vaccines were able to distinguish myths from facts immediately after reading the flyer, but did more poorly at identifying myths 30 minutes after reading the flyer.[35] [36] This is because familiarity increases the chances of accepting a falsehood as true, especially when details of the retraction

[33] Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*(4), 713-724.

[34] Swire, B., Ecker, U. K., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1948-1961.

[35] Cook, J., & Lewandowsky, S. (2011). The debunking handbook. Retrieved from https://www.skepticalscience.com/docs/Debunking_Handbook.pdf.

[36] Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, *2*(2), 81-94.

have faded from their memory. This effect has also been found to be particularly strong among older adults.[37] [38]

2.5 Overkill backfire effect

An overkill backfire effect occurs because processing many arguments requires more cognitive effort than just considering a few. When a falsehood is corrected with information that contains a complex explanation, people may reject the correction in favour of the former, which is a simpler account. This is because a simple myth is more cognitively attractive than an over-complicated correction. Hence, providing too many counterarguments can potentially backfire.[39] [40]

Related to this is also the issue of “satisficing” on the part of information consumers, especially when there is information overload. Satisficing is the process of selecting information that is “good enough” to satisfy basic needs. It could also mean selecting the first “acceptable answer” to a question or a solution to a problem, even if it means accepting a lower quality or quantity of information. Factors such as intellectual laziness,

[37] Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, *39*, 127-161.

[38] Bastin, C., & van der Linden, M. (2005). The effects of aging on the recognition of different types of associations. *Experimental Aging Research*, *32*(1), 61-77.

[39] Cook, J., & Lewandowsky, S. (2011). The debunking handbook. Retrieved from https://www.skepticalscience.com/docs/Debunking_Handbook.pdf.

[40] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

being unwilling or unable to deal with information overload, or not having the requisite information evaluation skills to reliably source information, explain why people may engage in satisficing.[41]

2.6 Worldview backfire effect

In some circumstances, people's belief in false information may increase when a correction challenges their worldviews, especially when they hold strong positions regarding an issue.[42] [43] For instance, Republicans are more likely than Democrats to strengthen their belief that Iraq possessed weapons of mass destruction despite the presence of retractions.[44] Recent studies have found that this worldview backfire effect seems stronger among the political right, where attitude-dissonant retractions were found to be consistently ineffective. This supports the view that conservative minds are particularly prone to worldview backfire effects when processing contentious corrective information.[45] [46]

[41] Cooke, N. A. (2018). *Fake news and alternative facts: Information literacy in a post-truth era*. Chicago, Illinois: American Library Association.

[42] Cook, J., & Lewandowsky, S. (2011). The debunking handbook. Retrieved from https://www.skepticalscience.com/docs/Debunking_Handbook.pdf.

[43] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

[44] Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303-330.

[45] Ecker, U., & Ang, L. C. (2017). Political attitudes and the processing of misinformation corrections. Retrieved from https://www.emc-lab.org/uploads/1/1/3/6/113627673/ecker.2018.polpsy.pdf.

[46] Ecker, U. K. (2017). Why rebuttals may not work: the psychology of misinformation. *Media Asia*, *44*(2), 79-87.

Research has also found evidence of similar backfire effects for non-political messages relating to climate change and vaccinations.[47] [48]

Besides content, subtle contextual cues can also affect the efficacy of a correction when they activate misinformation-congruent mental models. For instance, a picture of an Imam in Middle-Eastern attire reduces the efficacy of a message attributed to that person compared to when the same message is accompanied by a picture of the Imam dressed in Western attire.[49] [50]

In short, retractions that contradict one's worldviews are often perceived as less familiar, less coherent, and cognitively more difficult to process. They are also often less supported in one's social network (also see *Section 2.7: Other human factors*) and are more likely to be seen as coming from an untrustworthy source.[51]

[47] Hart, P. S., Nisbet, E. C., & Shanahan, J. E. (2011). Environmental values and the social amplification of risk: An examination of how environmental values and media use influence predispositions for public engagement in wildlife management decision making. *Society and Natural Resources*, *24*(3), 276-291.

[48] Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: A randomized trial. *Pediatrics*, *133*(4), e835-e842.

[49] Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, *63*(4), 617-637.

[50] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" Era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369.

[51] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

2.7 Other human factors

2.7.1 False consensus effect

Research has found that people tend to believe in false information when a significant number of others appear to believe in it as well. When people lack information of their own, they tend to rely on cues from others (which take the form of words and actions), on top of relying on heuristics and cognitive shortcuts as mentioned in the preceding sections.[52] Thus, people tend to hold firm to beliefs that they think are widely shared, even when it is not the case in reality. This discrepancy between actual and perceived prevalence of an opinion is known as the false consensus effect.[53]

For instance, a survey on attitudes towards climate change found that although the proportion of respondents who were climate change deniers was small (between 5% and 7%), this minority group of respondents felt that their opinion (that climate change is not real) was shared by between 43% and 49% of the population.[54] Furthermore, research has also found that when people believe that their opinion is widely shared, they are also more resistant to correction, and are more

[52] Price, V., Nir, L., & Cappella, J. N. (2006). Normative and informational influences in online political discussions. *Communication Theory*, *16*(1), 47-74.

[53] Krueger, J., & Zeiger, J. S. (1993). Social categorization and the truly false consensus effect. *Journal of Personality and Social Psychology*, *65*(4), 670-680.

[54] Leviston, Z., Walker, I., & Morwinski, S. (2013). Your opinion on climate change might not be as common as you think. *Nature Climate Change*, *3*(4), 334-337.

likely to insist that their views prevail.[55] This false consensus effect is particularly powerful when it pertains to one's reference group, making repetition in online echo chambers and filter bubbles especially influential.[56]

2.7.2 Homophily in social networks

Research has also found that homophily – the tendency for people to aggregate and form groups based on common interests and beliefs – plays a role in the spreading of false information online.[57]

A study which looked at the circulation of fake news stories on social media during the 2016 US Presidential Election found that individuals' voting patterns were strongly correlated with their exposure to fake news websites (i.e., users who tend to visit fake news websites more also tend to vote for Trump). Furthermore, the study also found that homophily on individuals' social networks explained this correlation between voting patterns and exposure to fake news websites. In other words, individuals tend to be connected to other users who share similar political beliefs on

[55] Miller, C. T. (1993). Majority and minority perceptions of consensus and recommendations for resolving conflicts about land use regulation. *Personality and Social Psychology Bulletin*, *19*(4), 389-398.

[56] Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279-301.

[57] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415-444.

social media, which in turn influences the degree to which they are likely to be exposed to fake news websites on online.[58]

Another study which looked at 12 million Italian Facebook users' consumption of scientific and conspiracy news found that an individual's engagement with a specific type of content (either scientific news or conspiracy news) correlated with the number of friends that engaged in similar content consumption patterns (i.e., homophily). In other words, whether consuming scientific news or conspiracy news, users tend to consume information in a manner similar to others in their social network. The study also concluded that homophily in social networks could serve as a key metric to identify online communities where false information is more likely to spread.[59]

2.7.3 Disbelieving facts altogether

Some scholars have also argued that being bombarded with falsehoods can cause people to stop believing in facts altogether and disengage from public discourse.[60] This was demonstrated in a study that looked at

[58] Fourney, A., Racz, M. Z., Ranade, G., Mobius, M., & Horvitz, E. (2017, November). Geographic and temporal trends in fake news consumption during the 2016 US Presidential Election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 2071-2074). ACM.

[59] Bessi, A., Petroni, F., Del Vicario, M., Zollo, F., Anagnostopoulos, A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015, May). Viral misinformation: The role of homophily and polarization. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 355-356). ACM.

[60] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369.

the changes in people’s attitudes on climate change after they were exposed to false information on the issue. Researchers found that presenting people with a message on the scientific consensus on climate change resulted in a positive influence on perceived scientific agreement on the matter. However, when false information that suggested no scientific consensus on climate change was presented alongside the consensus message, people did not experience any change in their beliefs in climate change.[61] Other studies have also found similar results which suggest that the presence of false information can cancel out the influence of facts.[62] Furthermore, this “disbelieving facts altogether” effect is particularly pronounced when false information is packaged as a conspiracy theory, where mere exposure to conspiratorial discourse makes people less likely to believe official information.[63] [64]

In short, the result of rampant false information is confusion, cynicism and information fatigue. When trust in facts is eroded, people begin to perceived facts as “unknowable”, irrelevant, and eventually start disbelieving facts altogether. As Lewandowsky *et al*. (2017) put it,

[61] van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2). 1-7.

[62] McCright, A. M., Charters, M., Dentzman, K., & Dietz, T. (2016). Examining the effectiveness of climate change frames in the face of a climate change denial counter-frame. *Topics in Cognitive Science*, *8*(1), 76-97.

[63] Einstein, K. L., & Glick, D. M. (2015). Do I think BLS data are BS? The consequences of conspiracy theories. *Political Behavior*, *37*(3), 679-701.

[64] Jolley, D., & Douglas, K. M. (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS one*, *9*(2), e89177.

“Misinformation is therefore not just about being misinformed. It is also about the overall intellectual well-being of a society”.

3. TECHNOCOGNITIVE SOLUTIONS

As seen in *Section 2: Cognitive biases and other human factors*, humans are arguably “biologically hardwired” to be prone to believing in falsehoods. However, the technology (e.g., the Internet and social media) we use in our daily lives for information consumption often act on our biases, and play a role in amplifying the spread of falsehoods as well. In Singapore for instance, a poll conducted by REACH in 2018 found that among respondents who came across inaccurate online news, 50% came across it on WhatsApp and 46% did so on Facebook.[65]

Thus, it unsurprising that many surveys have found that people generally feel that technology companies have a responsibility to address the spread of online falsehoods, and that they are still not doing enough to address the problem:

- The 2018 Reuters Institute Digital News Report found that most respondents believed that publishers (75%) and platforms (71%) have the biggest responsibility to fix problems of fake and unreliable news.[66]

[65] Findings of poll on attitudes towards fake news. (2018, March 26). Retrieved from https://www.reach.gov.sg/~/media/2018/press-release/media-release-on-findings-of-fake-news-poll-26-mar-2018.pdf.

[66] Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., & Nielsen, R. K. (2018). Reuters Institute digital news report 2018. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf.

- In a survey conducted by Gallup and Knight Foundation in 2018, 76% of US adults said that major Internet companies have an obligation to identify misinformation that appears on their platforms, and 48% strongly felt that Internet companies are not doing enough to stem the spread of misinformation.[67]
- Another poll conducted by Huffington Post and YouGov in August 2018 found that 48% of Americans felt that social media sites are currently not strict enough in regulating what is posted on these platforms.[68]

In their defense, technology companies have been rolling out various countermeasures to combat the scourge of online falsehoods on their platforms. For instance, Facebook has experimented with initiatives ranging from flagging articles disputed by third-party fact checkers, to its recent initiative that rates how trustworthy Facebook users are at reporting false news.[69] Twitter, on the other hand, has been actively cracking down on bots and fake accounts on its platform.[70] The recent

[67] Americans' views of misinformation in the news and how to counteract it. (2018, June 20). Retrieved from https://www.knightfoundation.org/reports/americans-views-of-misinformation-in-the-news-and-how-to-counteract-it?utm_source=link_newsv9&utm_campaign=item_235796&utm_medium=copy.

[68] Edwards-Levy, A. (2018, August 20). Most people say social media sites should crack down on harassment, fake news: Poll. *The Huffington Post*. Retrieved from https://www.huffingtonpost.com/entry/social-media-harassment-fake-news-poll-alex-jones_us_5b7b1c53e4b0a5b1febdf30a.

[69] Kastrenakes, J. (2018, August 21). Facebook begins rating users on how trustworthy they are at flagging fake news. Retrieved from https://www.theverge.com/2018/8/21/17763886/facebook-trust-ratings-fake-news-reporting-score.

[70] Kastrenakes, J. (2018, July 27). Twitter reports a million fewer users as a result of ongoing crackdown on bots. Retrieved from https://www.theverge.com/2018/7/27/17620440/twitter-q2-2018-earnings-1-million-mau-fall-in-spam-crackdown.

ban of notorious conspiracy website, Infowars, by YouTube and Facebook has reduced its reach. Analysis from web data firms Tubular Labs and SimilarWeb found that traffic to Infowars' website fell from 1.4 million visits to 715,000 visits, suggesting that technology companies exert significant influence over the spread of information online.[71]

However, research has found that technological interventions can at times be ineffective or counterproductive. For instance, some studies have found that interacting with fact checks on Facebook can ironically lead to greater engagement with conspiracy-related posts, and exposure to warnings tagged to fake news may increase cynicism about news articles in general.[72] [73]

Thus, some scholars have argued that in order for technological solutions to be most effective, they should incorporate the principles and findings of psychological research, particularly on the psychology of how people communicate and consume information online. This discipline is known as "technocognition", which is defined by Lewandowsky *et al*.

[71] Nicas, J. (2018, September 4). Alex Jones said bans would strengthen him. He was wrong. *The New York Times*. Retrieved from https://www.nytimes.com/2018/09/04/technology/alex-jones-infowars-bans-traffic.html.

[72] Zollo, F., Bessi, A., Del Vicario, M., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., & Quattrociocchi, W. (2017). Debunking in a world of tribes. *PloS one*, 12(7), e0181821.

[73] Pennycook, G., & Rand, D. G. (2017). The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384.

(2017) as *"an interdisciplinary approach that combines research findings from psychology, critical thinking approaches from philosophy, and behavioural economics principles, in the design of information architectures deployed via scaleable, technological solutions, to nudge against the spread of misinformation"* (p. 362).[74]

Adopting a technocognitive approach to combat online falsehoods would include asking questions such as, "How can technology help nudge people out of their filter bubbles?"; "When people are nudged out of their filter bubbles, what type of 'alternative information' is considered 'ideal' to be presented to them?"; "How far can 'alternative information' sit outside of people's comfort zone?"; "What is the effect of indicators of trustworthiness (e.g., verifications awarded by independent fact checkers) on the perceived credibility of websites and online information?" This section will examine how effective existing technological interventions have been at mitigating human biases, and how taking a technocognitive approach to the problem can further improve the quality of information shared online.

[74] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369.

3.1 “Bursting” people’s filter bubbles

Research has shown that one strategy to counter confirmation bias is to get people to consider perspectives apart from their own.

As explained in *Section 2.1: Confirmation bias and motivated reasoning*, an experiment where people were presented with evidence that supported either the continuation or abolition of the death penalty found that people who were against the death penalty rated anti-death penalty evidence as highly convincing and rated pro-death penalty evidence as unconvincing. The reverse was true for those who were supportive of the death penalty.

The researchers repeated the experiment, but with two added sets of instructions. The first set of instructions told participants to be “as objective and unbiased as possible” by imagining that they were judges tasked to “weigh all of the evidence in a fair and impartial manner". The second set of instructions told participants to think about how they would assess the evidence if the results had pointed the opposite way. This was coined as the **“consider the opposite”** strategy. The study found that participants who received the second set of instructions seemed to perform better at overcoming their confirmation bias – while they did not rate disconfirmatory evidence as more convincing than confirmatory evidence, they also did not become more entrenched or more extreme in their views at the very least.[75] In short, it might be insufficient to simply tell people to be fair and objective in their assessment of information and views. Instead, people should be told to consider views apart from their own.

Applying the same principles, platform companies can help users mitigate the effects of confirmation bias by "bursting" users' filter bubbles and by exposing them to a more **diverse range of information types**. Currently, platform companies allow users to personalise the content they see according to individual preferences. Algorithms then infer from users' preferences and further recommend content that is consistent with their likes. As a result, people are being put in an online filter bubble where their worldviews are constantly reinforced.[76] Instead, platform companies should attempt to increase the social distance between suggested content and users' preferences by designing algorithms to provide users with suggestions to follow pages and accounts that give different types of content.[77]

[75] Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231-1243.

[76] Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London, UK: Penguin.

[77] Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Retrieved from https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Such technological tools have already been designed by various stakeholders in the recent past. Platform companies themselves have been actively working on solutions to "burst" people's filter bubbles. For instance, Google has been looking into improving its "Featured Snippets" function. Currently, Google provides users with a descriptive box at the top of their Google search results to make it easier to for users to access information they are looking for. However, these "Featured Snippets" may sometimes include misleading information, unsupported conspiracy theories or even unexpected offensive results. Thus, Google is experimenting with surfacing multiple snippets when users perform a search in order to offer varying points of view.[78] [79] Twitter, on the other hand, has recently been testing out a function that prompts users about accounts they might want to unfollow.[80]

Third-party developers have also been designing mobile apps and browser extensions that seek to "burst" people's filter bubbles. One example is a new app called Burst, which targets filter bubbles on Reddit. Reddit, an online forum and content aggregator platform, also has a serious issue with filter bubbles as users are allowed to follow subreddits

[78] Hao, K. (2018, February 1). Google is finally admitting it has a filter-bubble problem. Retrieved from https://qz.com/1194566/google-is-finally-admitting-it-has-a-filter-bubble-problem/.

[79] Sullivan, D. (2018, January 30). A reintroduction to Google's featured snippets. Retrieved from https://www.blog.google/products/search/reintroduction-googles-featured-snippets/.

[80] Summers, N. (2018, August 30). Twitter tests personalised 'unfollow' recommendations. Retrieved from https://www.engadget.com/2018/08/30/twitter-test-personalized-unfollow-recommendations/.

– essentially forming subcommunities – and only hear their own viewpoints.[81] Other tools that aim to increase people's exposure to diverse content include Chrome extension PolitiEcho, Twitter plug-in FlipFeed, and Facebook plug-in Escape Your Bubble.[82] Besides players in the technology industry, news organisations have also leveraged technology to "burst" people's filter bubbles. For instance, The Times launched a Facebook messenger bot called Filter-bubble Buster which provided people with balanced information leading up to the 2017 UK General Election.[83]

On top of designing technological tools that "burst" filter bubbles, another way to diversify people's exposure to online content is to encourage them to make their voices heard so that the discourse will not be dominated by few similar opinions. Research has found that the comments on news articles and blog posts can affect other readers' impressions of the issue being discussed.[84] [85] Furthermore, the tone of the discussion can affect people's attitudes towards the topic, and just a

[81] Perez, S. (2018, April 5). Burst breaks you out of your filter bubble on Reddit. Retrieved from https://techcrunch.com/2018/04/04/burst-breaks-you-out-of-your-filter-bubble-on-reddit/.

[82] Hess, A. (2017, March 3). How to escape your political bubble for a clearer view. *The New York Times*. Retrieved from https://www.nytimes.com/2017/03/03/arts/the-battle-over-your-political-bubble.html.

[83] Davies, L. (2017, May 30). UK pubs enlist bots to fight against filter bubbles ahead of the UK election. Retrieved from https://digiday.com/media/uk-pubs-enlist-bots-fight-filter-bubbles-ahead-uk-election/.

[84] Stavrositu, C. D., & Kim, J. (2015). All blogs are not created equal: The role of narrative formats and user-generated comments in health prevention. *Health Communication*, *30*(5), 485-495.

[85] Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior*, *58*, 461-470.

few dissenting voices can shift the perceived social norm. For instance, a study found that even when people read a neutral article on nanotechnology, exposure to uncivil comments and other expressions of incivility polarised their views of nanotechnology.[86] This suggests that platform companies can "burst" filter bubbles (as well as mitigate the false consensus effect – see *Section 2.7.1: False consensus effect*) through active **content and comment moderation** to set the tone of the discourse, and establish online norms that will encourage more people to express their opinions. One example is a Norwegian news site that requires readers to pass a brief comprehension quiz about the article before being allowed to post comments in order to reduce rants and to establish a common ground for debate.[87]

Others have also suggested that social network companies should change the terminology used on their platforms. For instance, "to follow" implies a kind of agreement that may create an emotional resistance against other opinions, causing people to remain in their online filter bubbles at a subconscious level. Thus, neutral labels should be used instead.[88]

[86] Runge, K. K., Yeo, S. K., Cacciatore, M., Scheufele, D. A., Brossard, D., Xenos, M., Anderson, A., Choi, D., Kim, J., Li, N., Stubbings, M., Leong, Y. S., & Liang, X. (2013). Tweeting nano: How public discourses about nanotechnology develop in social media environments. *Journal of Nanoparticle Research*, *15*(1), 1381.

[87] Schimdt, C. (2017, August 11). Remember that Norwegian site that made readers take a quiz before commenting? Here's an update on it. Retrieved from http://www.niemanlab.org/2017/08/remember-that-norwegian-site-that-makes-readers-take-a-quiz-before-commenting-heres-an-update-on-it/.

[88] Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Retrieved from

While these efforts look promising, the impact of these tools is not known and more research needs to be done. Furthermore, a mere exposure to alternative views and information does not guarantee attention as people may still consciously overlook content that does not interest them. In fact, it may even induce cynicism of "the other side". Thus, when designing tools to "burst" filter bubbles, technology companies should ensure that these tools and initiatives remain as politically neutral as possible.[89] [90]

3.2 Curbing the spread of falsehoods

As mentioned in *Section 2.3: Illusory truth effect*, increased exposure to and (thus) increased familiarity with a falsehood increases its likelihood of being accepted as true. Hence, technology companies can mitigate the illusory truth effect by leveraging their technological expertise (e.g., tweaking algorithms to reduce the visibility of false information) to curb the spread of falsehoods online and prevent them from going viral.

For instance, Google announced a Google News Initiative that aims to help journalism thrive in the digital age. Part of the initiative focuses on combating false information by training Google's systems to recognise

https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

[89] Bode, L., Vraga, E. K., & Troller-Renfree, S. (2017). Skipping politics: Measuring avoidance of political content in social media. *Research & Politics*, *4*(2), 1-7.

[90] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369.

and elevate more authoritative content and demote low quality and misleading content. The initiative also aims to work with news organisations to help information consumers distinguish fact from fiction especially during breaking news situations and election periods.[91] Given that hundreds of millions of people rely on Google's Chrome browser and search engine for factual data, Google is also considering developing a browser extension that functions like a false information detector.[92]

Facebook, on the other hand, has adopted an "ecosystem" approach which comprises actions taken at various steps – from account creation, false content creation, to distribution – to curb the spread of falsehoods on its platform. Specific to targeting the distribution of falsehoods on its platform, Facebook has removed "Trending" in June 2018 after criticisms which questioned the way it had curated stories that appeared in "Trending" – some alleged that "Trending" contributed to the spread of inaccurate and offensive news, and routinely suppressed conservative stories as well.[93] [94] In its place, Facebook rolled out "Breaking News

[91] Schindler, P. (2018, March 20). The Google News Initiative: Building a stronger future for news. Retrieved from https://www.blog.google/outreach-initiatives/google-news-initiative/announcing-google-news-initiative/.

[92] Timmons, H. (2018, February 8). Google executives are floating a plan to fight fake news on Facebook and Twitter. Retrieved from https://qz.com/1195872/google-facebook-twitter-fake-news-chrome/.

[93] Cook, James. (2018, June 1). Facebook to remove 'trending' news stories section following years of controversy. *The Telegraph*. Retrieved from https://www.telegraph.co.uk/technology/2018/06/01/facebook-remove-trending-news-stories-section-following-years/.

[94] Kastrenakes, J. (2018, June 1). Facebook will remove the Trending topics section next week. Retrieved from https://www.theverge.com/2018/6/1/17417428/facebook-trending-topics-being-removed.

Label" for breaking news situations, and "Today In", which provides people with updates from their local publishers, officials and organisations.[95] Facebook has also partnered third-party fact checking organisations – it currently has 25 partners in 14 countries – to identify false information and prevent their spread. Once an article is rated as false by these certified fact checkers, Facebook will deprioritise the article in NewsFeed, reducing its future views by an average of 80%. On top of that, Facebook has also been using machine learning to help **predict content that might be false news** and prioritise material that they send to their third-party fact checker partners for verification (see *Section 3.3: Flagging falsehoods and promoting verified content* for more details).[96] [97]

Automated accounts such as bots also perpetuate the illusory truth effect by amplifying the spread of online falsehoods at high volumes. Furthermore, they are also used to give the false impression that a large number of people are talking about a certain topic, thus contributing to the false consensus effect as well. Research has found that Twitter accounts that actively spread false information are significantly more likely to be bots, and that about 9% to 15% of active Twitter accounts

[95] Hardiman, A. (2018, June 1). Removing Trending from Facebook. Retrieved from https://newsroom.fb.com/news/2018/06/removing-trending/.

[96] Lyons, T. (2018, June 21). Increasing our efforts to fight false news. Retrieved from https://newsroom.fb.com/news/2018/06/increasing-our-efforts-to-fight-false-news/.

[97] Lyons, T. (2018, June 14). Hard questions: How is Facebook's fact-checking program working? Retrieved from https://newsroom.fb.com/news/2018/06/hard-questions-fact-checking/.

are bots.[98] Bots often play a significant role in the early spreading phases of viral fake news, and tend to target influential users (i.e., accounts that are highly connected nodes on social networks).[99] This suggests that **clamping down on bots and bot users** may also be another effective strategy for curbing the spread of online falsehoods.

In response to the problem of bots, Twitter has been actively identifying bot-based manipulation and spam, and developing tools that can be used to spot and shut down fake accounts on its platform. For instance, in January 2018, Twitter made changes to its TweetDeck and API to prohibit users from performing coordinated actions across multiple accounts in their services.[100] Continuing its efforts to clean up the site, Twitter announced in July 2018 that it will be removing tens of millions of suspicious and fake accounts to prevent attempts to manipulate conversations on the platform (e.g., Russian meddling in the 2016 US Presidential Election). As a result, Twitter's malicious spam removal in 2018 was up by 214% as compared to in 2017.[101] [102] Lastly, Twitter also

[98] Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization [Research paper]. Retrieved from https://arxiv.org/pdf/1703.03107.pdf.

[99] First evidence that social bots play a major role in spreading fake news. (2017, August 7). Retrieved from https://www.technologyreview.com/s/608561/first-evidence-that-social-bots-play-a-major-role-in-spreading-fake-news/.

[100] Roth, Y. (2018, February 21). Automation and the use of multiple accounts [Blog post]. Retrieved from https://blog.twitter.com/developer/en_us/topics/tips/2018/automation-and-the-use-of-multiple-accounts.html.

[101] Liao, S. (2018, June 17). Twitter's spam removal is up 214 percent compared to 2017. Retrieved from https://www.theverge.com/2018/6/27/17510582/twitters-spam-removal.

[102] Kastrenakes, J. (2018, July 27). Twitter reports a million fewer users as a result of ongoing crackdown on bots. Retrieved from

plans to improve machine-learning technology that can help the company detect suspicious and automated account activities.[103] [104]

Facebook has also been actively tackling the problem of fake accounts on its platform after investigations revealed that a Russian troll farm had used the social media site to try to influence the 2016 US Presidential Election. Relying largely on algorithms and machine learning (as well as a small proportion being reported by users) to identify and eliminate fake accounts, Facebook announced in May 2018 that it had disabled 583 million accounts in the first quarter of 2018, many of which with the intent of spreading spam or conducting illicit activities like scams.[105] [106] On top of removing fake accounts in general, Facebook had also eliminated multiple pages, groups and accounts for coordinated inauthentic behaviour. In August 2018, Facebook announced that it had removed 652 fake accounts and pages with connections to Russia and Iran that attempted to exert political influence in the US, UK, Middle East and

https://www.theverge.com/2018/7/27/17620440/twitter-q2-2018-earnings-1-million-mau-fall-in-spam-crackdown.

[103] Roth, Y., & Harvey, D. (2018, July 26). How Twitter is fighting spam and malicious automation. Retrieved from https://blog.twitter.com/official/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html.

[104] Stewart, E. (2018, July 11). Twitter's wiping tens of millions of accounts from its platform. Retrieved from https://www.vox.com/2018/7/11/17561610/trump-fake-twitter-followers-bot-accounts.

[105] Community standards enforcement preliminary report. (2018). Retrieved from https://transparency.facebook.com/community-standards-enforcement#fake-accounts.

[106] Wagner, K., & Molla, R. (2018, May 15). Facebook has disabled almost 1.3 billion fake accounts over the past six months. Retrieved from https://www.recode.net/2018/5/15/17349790/facebook-mark-zuckerberg-fake-accounts-content-policy-update.

Latin America. The perpetrators behind these influence campaigns often used tactics such as creating networks of accounts to mislead others about who they were and what they were doing.[107] [108] In September 2018, Facebook CEO Mark Zuckerberg said that Facebook's efforts in this aspect have better prepared the platform for election meddling in the future.[109]

While such active efforts to remove automated bot accounts and fake accounts are crucial in mitigating the illusory truth and false consensus effects, experts, and even technology companies themselves, have recognised that such measures have their limitations. Many producers of falsehoods are often well-funded and are constantly changing tactics. This places technology companies in an arms race against falsehood producers as fraudulent pages are being created as fast as platform companies can delete them.[110] Facebook has acknowledged that it would have to invest heavily in better technology to improve and continue preventing bad actors from misusing its platform.[111]

[107] Taking down more coordinated inauthentic behavior. (2018, August 21). Retrieved from https://newsroom.fb.com/news/2018/08/more-coordinated-inauthentic-behavior/.

[108] Solon, O. (2018, August 22). Facebook removed 652 fake accounts and pages meant to influence world politics. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/aug/21/facebook-pages-accounts-removed-russia-iran.

[109] Zuckerberg, M. (2018, September 13). Preparing for elections. Retrieved from https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/10156300047606634/.

[110] Vaidhyanathan, S. (2018, September 5). Why Facebook will never be free of fakes. *The New York Times*. Retrieved from https://www.nytimes.com/2018/09/05/opinion/facebook-sandberg-congress.html.

[111] Removing bad actors on Facebook. (2018, July 31). Retrieved from https://newsroom.fb.com/news/2018/07/removing-bad-actors-on-facebook/.

3.3 Flagging falsehoods and promoting verified information

In March 2018, a study by MIT researchers found that false news spreads more rapidly than real news on Twitter, and that humans, not bots, were primarily responsible for the spread of misleading information.[112] The findings of this study suggest that as much as it is important to curb the spread of falsehoods by bots, measures must also be put in place to ensure that real users do not inadvertently contribute to this spread. As such, technology companies have been – often in collaboration with fact checkers – working on tools that flag falsehoods on their platforms, promoting verified information and authoritative content, and ensuring that corrective information is designed in a specific way that debunks falsehoods while minimising any potential backfire effects. This last point is essential to mitigate the backfire effects mentioned earlier such as the worldview backfire effect (*Section 2.6: Worldview backfire effect*).

Promoting verified information

Technology companies like Facebook and Google have been revising their algorithms to **feature information from authoritative sources more prominently**. Amid the rash of conspiracy theories, YouTube has also announced that it would be promoting videos from vetted news sources on its Top News and Breaking News sections to better support

[112] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

trusted news providers.[113] It will also start to add previews and links to trusted news articles, and incorporate text from third parties such as Wikipedia and Encyclopedia Britannica on falsehood-ridden subjects (e.g., the moon landing and the Oklahoma City bombing).[114] Experts have also proposed for social networks and search engines to **highlight contextual details and leverage visual indicators** especially when promoting verified information. For instance, technology companies could surface context information and metadata – e.g., automatically showing when a website was registered or running a reverse Google image search to see whether an image is an old one – that would provide users with more cues to ascertain the truth of a piece of content. Visual indicators such as a blue verification tick can also serve as a helpful visual indicator if it is consistent across platforms. Thus, **technology companies could collaborate** to build a consistent set of visual indicators for these contextual details. Furthermore, such visual language should be developed in collaboration with cognitive psychologists to ensure efficacy.[115]

[113] Hern, A. (2018, July 10). YouTube to crack down on fake news, backing 'authoritative' sources. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/jul/09/youtube-fake-news-changes.

[114] Castillo, M. (2018, July 9). YouTube will use six popular YouTube stars to educate kids about fake news. *CNBC*. Retrieved from https://www.cnbc.com/2018/07/09/youtubes-plan-to-fight-fake-news-includes-more-support-article-links.html

[115] Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. Retrieved from https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Flagging falsehoods

Facebook first rolled out the "Disputed Flags" feature in an attempt to bring to people's attention whenever they came across an article that had been disputed by third-party fact checkers. However, in December 2017, Facebook announced that it would no longer be using "Disputed Flags" to identify false news, citing research which showed that placing an attention-grabbing cue, like a red flag, next to an article may actually backfire and entrench deeply held beliefs. An implied truth effect, which researchers found was more pronounced among Trump supporters and young adults, may also result because false stories that are not tagged with a disputed flag may be seen as validated and accurate.[116] Moreover, the "Disputed Flags" feature did not tell people why fact checkers had disputed an article, which is important especially when most users do not bother clicking on links to additional information.[117] [118] Thus, Facebook replaced "Disputed Flags" with "**Related Articles**", which aims to give people more context about an issue to help them decide what is true and what is false for themselves. Facebook also cited

[116] Pennycook, G., & Rand, D. G. (2017). The implied truth effect: Attaching warning to a subset of fake news stories increases perceived accuracy of stories without warnings. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384.

[117] Shu, C. (2017, December 21). Facebook will ditch Disputed Flags on fake news and display links to trustworthy articles instead. Retrieved from https://techcrunch.com/2017/12/20/facebook-will-ditch-disputed-flags-on-fake-news-and-display-links-to-trustworthy-articles-instead/.

[118] However, some research also suggested that warning labels do have an effect on people's perceived accuracy of a piece of information, except that using the term "Rated False" seems to work better than the term "Disputed". This suggests the need for further research into how to most effectively counter fake news using warning labels [Blair, S., Busam, J. A., Clayton, K., Forstner, S., Glance, J., Green, G., & Zhou, A. (2017). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check banners in reducing belief in false stories on social media. Retrieved from https://www.dartmouth.edu/~nyhan/fake-news-solutions.pdf.]

research that supported the idea that directly surfacing related stories to correct a post containing false information can significantly reduce people's misperceptions, and led to fewer shares of the false news story than when a disputed flag is shown.[119] [120]

It is clear that technology companies are in the process of testing and trying which interventions work and which do not. Given the complexities involved – understanding users' cognition; the changing nature of falsehoods; and increasingly sophisticated deployment of technology by falsehood producers – this process is likely to be an ongoing one.

One potential area is to explore the **effects of social context on people's propensity to fact check** the information they encounter online, and how certain aspects of information technologies (e.g., user interface) can be re-designed to address these issues. For instance, research has found people were less likely to fact check when they evaluated claims in a collective and social setting (e.g., on social media) than when in an individual setting. This suggests that perceived highly social contexts (e.g., online environments with "likes" and "shares") may impede fact checking or lower people's guard in terms of being sceptical

[119] Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, *65*(4), 619-638.

[120] Lyons, T. (2017, December 20). Replacing Disputed Flags with Related Articles. Retrieved from https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/.

about the information they receive.[121] Interestingly however, other research has also found that strong social connections between fact checkers and rumour spreaders tend to encourage the latter to accept corrections and share accurate information. A study that examined political corrections on Twitter found that people were more likely to accept corrections from individuals whom they follow (and are followed by) than when a being corrected by a stranger. Researchers of this study argued that this is because when there are strong social connections, individuals feel more accountable to their friends and community, and are thus more likely to share collective interests with them.[122] In sum, these two studies suggest that social context plays an important part in people' fact checking behaviour and tendency to accept correction, and that more research needs to be done to better understand the nuances such that these principles can be applied when designing technological solutions.

4. EFFECTIVE COMMUNICATION

On top of drawing from findings of psychological research to design technological solutions (i.e., technocognitive solutions), there are also specific messaging strategies that we can learn from both psychology and communications research – especially when presenting corrective

[121] Jun, Y., Meng, R., & Johar, G. V. (2017). Perceived social presence reduces fact-checking. *Proceedings of the National Academy of Sciences*, *114*(23), 5976-5981.

[122] Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, *35*(2), 196-219.

information to audiences during fact checking – to minimise any of the backfire effects mentioned in *Section 2: Cognitive biases and other human factors*.

Existing research underscores the importance of effective messaging in the context of debunking and correcting falsehoods. For instance, the headline of a debunking message with the myth in big and bold letters is less effective than communicating the facts in the headline. An effective debunking messaging should increase people's familiarity with the facts by starting with and emphasising the facts; keeping the content lean; using simple language and short sentences; and using graphics to illustrate wherever possible. Ideally, the debunking message should also end with a strong and simple message to increase its recall and stickiness, so that people will share the message with their friends.[123]

Besides tailoring the various elements of a debunking message, research also suggests that equal importance should be given to the ways in which information is presented in a message. For instance, research in consumer psychology that examined the effects of one-sided, two-sided non-refutational, and two-sided refutational messages on consumer response found that two-sided refutational appeals were positively correlated with greater acceptance of the communicator's

[123] Cook, J., & Lewandowsky, S. (2011). The debunking handbook. Retrieved from https://www.skepticalscience.com/docs/Debunking_Handbook.pdf.

position than one-sided and two-sided non-refutational appeals. Two-sided refutational appeals were also generally more effective in increasing perceived source truthfulness and believability.[124] [125]

Related to that is another promising solution derived from the inoculation theory, which prepares individuals for potential false information by exposing logical fallacies a priori. The term "inoculation" is biological metaphor first suggested by McGuire in 1961, who proposed that an individual's beliefs could be inoculated against persuasive attacks the same way our immune system could be immunised against viruses.[126] Similar to how vaccines work – by injecting a weakened form of a virus into the body so that the body produces antibodies that will protect itself from a stronger form of the virus in future – McGuire argued that exposing individuals to information containing weakened arguments can help individuals develop resistance against more persuasive attacks in future. Hence, this application of inoculation theory functions like a "psychological vaccination", where individuals are inoculated against false information that they may encounter in future.[127]

[124] Kamins, M.A., & Assael, H. (1987). Two-sided versus one-sided appeals: A cognitive perspective on argumentation, source derogation, and the effect of disconfirming trial on belief change. *Journal of Marketing Research*, *24*, 29-39.

[125] Rucker, D.D., Petty, R.E., & Brinol, P. (2008). What's in a frame anyway? A meta-cognitive analysis of the impact if one versus two-sided message framing on attitude certainty. *Journal of Consumer Psychology*, *18*, 137-149.

[126] McGuire, W. J. (1961). The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. *Sociometry*, *24*(2), 184-197.

[127] Compton, J., Jackson, B., & Dimmock, J. A. (2016). Persuading others to avoid persuasion: Inoculation theory and resistant health attitudes. *Frontiers in Psychology*, *7*(122). 1-9.

4.1 Filling the gap: Providing an alternative explanation

Research suggests that **providing an alternative explanation**, instead of simply negating false information, may mitigate the continued influence effects of false information.

As mentioned in *Section 2.2: Continued influence effect*, people have a tendency to continue to rely partially on false information even after being given corrections. This is largely because they build mental models of the world and prefer their mental models to be complete even if they might be incorrect. In other words, people incorporate false information they encounter into their mental models as it provides a particular explanation and helps them understand the world. When this myth is debunked, a gap is left in their mental models. In the absence of a better explanation, people may continue to opt for the wrong account, which explains why people sometimes continue to rely partially on a myth even after retractions are given.[128] [129]

Thus, the most effective way to counter the continued influenced effects of false information is to provide an alternative explanation to fill the gap

[128] Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420-1436.

[129] Ecker, U. K. (2017). Why rebuttals may not work: the psychology of misinformation. *Media Asia*, *44*(2), 79-87.

in people's mental models. To increase the effectiveness of a retraction, debunking messages should ideally: [130] [131] [132]

- Explain why the false information was thought to be correct in the first place.
- Explain the underlying motivations of those who promote the false information.
- Expose the rhetorical techniques used to misinform (e.g., cherry picking of evidence, quoting of fake experts etc.).
- Be plausible, possess internal coherence, and account for all the important causal qualities and observed features in the false information.

Unfortunately, this approach may have some limitations, namely when no good alternative explanation is available and thus there is nothing to fill the gap in people's mental models with.[133] [134] Even in situations where a factual alternative is available, it may sometimes be very complicated and difficult for people to understand, or may motivate

[130] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369.

[131] Cook, J., & Lewandowsky, S. (2011). The debunking handbook. Retrieved from https://www.skepticalscience.com/docs/Debunking_Handbook.pdf.

[132] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

[133] Swire, B., & Ecker, U. K. H. (2018). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. In *Misinformation and mass audiences* (pp. 195–211). Austin, Texas: University of Texas Press.

[134] Ecker, U. K. (2017). Why rebuttals may not work: the psychology of misinformation. *Media Asia*, *44*(2), 79-87.

people to reject it because it is not consistent with their pre-existing attitudes and worldviews.[135] The subsequent sections will elaborate on some other principles that are worth adhering to in order to maximise the effectiveness of retractions and debunking messages.

4.2 Using explicit pre-exposure warnings

Research has found that **providing a pre-exposure warning**[136] that details the possibility of the continued influence effect before providing a factual alternative that debunks a myth greatly reduced people's reliance on the false information. In fact, research found that the pre-exposure warning itself was as effective as the factual alternative in mitigating the continued influence effects of false information. Researchers involved in the study argued that this is because the pre-exposure warning allows individuals to more effectively tag the misinformation as false and also facilitates easier recall of the tag, and thus, easier recall of the retraction.[137] [138]

[135] *ibid.*

[136] One example of a pre-exposure warning is the statement, *"Research has shown that people continue to rely on outdated information even when it has been retracted or corrected. Please read the following carefully."*

[137] Swire, B., & Ecker, U. K. H. (2018). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. In *Misinformation and mass audiences* (pp. 195–211). Austin, Texas: University of Texas Press.

[138] Ecker, U. K., Lewandowsky, S., & Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087-1100.

4.3 Emphasising facts

As mentioned earlier in *Section 2.3: Illusory truth effect* and *Section 2.4: Familiarity backfire effect*, increased exposure to and familiarity with a piece of false information can increase the likelihood of it being accepted as true. Ideally, the familiarity backfire effect can be avoided by not mentioning the myth altogether during retractions. However, this is sometimes not a practical option. In fact, research has found that it is safe to repeat the false information once in order to refute it, although it certainly should not be repeated more than necessary.[139] Thus, it is more important to focus on emphasising facts when presenting corrective information instead.

One way to emphasise facts is to **start with the facts** before presenting the myths in debunking messages. A common mistake made in retractions is that they often repeat the myth before pointing out that it is false – for instance, *"**Flight MH370 was hijacked** – FALSE"*. In such cases, people who read the statement may potentially believe it, only to be told that it is false afterwards. They then have to backtrack and revise their understanding, which might instead boost the familiarity of the false information and risk the chance of the false information being mistakenly remembered as true subsequently. Hence, it is better to start with the facts, warn people that what is upcoming is false, and only then mention

[139] Ecker, U. K., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, *6*(2), 185-192.

the myth in the correction.[140] [141] [142] Headlining debunking messages with the myth in big, bold letters should also be avoided. Instead, the facts should be communicated and emphasised in the debunking headline.[143]

Another way to emphasise facts is to **repeat retractions (without reinforcing the myth)**. Research has found that the effectiveness of debunking messages can be enhanced if they are repeated – repeated retractions alleviated the effects of false information even though they did not eliminate it. This technique is important in the domain of social media, where false information is often disseminated quickly and widely. Thus, the retractions and debunking messages must be circulated with equal (if not greater) vigor in order to counter the effects of false information[144] [145].

[140] Swire, B., & Ecker, U. K. H. (2018). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. In *Misinformation and mass audiences* (pp. 195–211). Austin, Texas: University of Texas Press.

[141] Ecker, U. K. (2017). Why rebuttals may not work: the psychology of misinformation. *Media Asia*, *44*(2), 79-87.

[142] Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, *39*, 127-161.

[143] Ecker, U. K. (2017). Why rebuttals may not work: the psychology of misinformation. *Media Asia*, *44*(2), 79-87.

[144] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

[145] Ecker, U. K., 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*(3), 570-578.

In short, the goal of emphasising facts in retractions and debunking messages is to increase people's familiarity with the facts.

4.4 Keeping it simple

As mentioned in *Section 2.5: Overkill backfire effect*, a simple myth is often cognitively more attractive than an over-complicated correction, which may cause people to reject the correction for the former. Thus, it is important to generate only a few counterarguments and ensure that corrective information is **lean, mean, and easy to read**. This is because information that is cognitively easier to process is more likely to be accepted as true. This can be done through techniques such as using simple language; short sentences; subheadings and paragraphs; and using graphics wherever possible to illustrate your points. It is also ideal to end with a strong and simple message so that people will remember the message and share it with their friends.[146] [147]

4.5 Affirming worldview

As mentioned in *Section 2.6: Worldview backfire effect*, debunking messages that contradict people's worldviews are often ineffective because such retractions are perceived as less familiar, less coherent

[146] Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106-131.

[147] Cook, J., & Lewandowsky, S. (2011). The debunking handbook. Retrieved from https://www.skepticalscience.com/docs/Debunking_Handbook.pdf.

and cognitively more difficult to process; they may even ironically increase people’s belief in the falsehood.

Existing research has found that **framing solutions to a problem in worldview-congruent terms** can enhance its acceptance by people who would typically reject it. For instance, research found that conservatives were more likely to accept evidence of climate change if it was presented as a business opportunity for the nuclear industry.[148] Similarly, people who opposed nanotechnology on grounds of being eco-friendly might be less likely to dismiss evidence of its safety if the use of nanotechnology was presented as part of an effort to protect the environment. In other words, debunking messages and retractions must be tailored to their specific audience to reduce its perceived threat, especially when an issue is contentious or politically sensitive.[149] [150]

Furthermore, research also found that **credible sources** can significantly augment the effects of a worldview-congruent framing. For instance, a study found that participants who were presented with arguments on the impacts of crime and violence were 19% more likely

[148] Feygina, I., Jost, J. T., & Goldsmith, R. E. (2010). System justification, the denial of global warming, and the possibility of “system-sanctioned change”. *Personality and Social Psychology Bulletin*, *36*(3), 326-338.

[149] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “post-truth” era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369.

[150] Swire, B., & Ecker, U. K. H. (2018). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. In *Misinformation and mass audiences* (pp. 195–211). Austin, Texas: University of Texas Press.

to support gun control measures if the message came from a New York Times journalist than if it came from a spokesperson with a perceived bias.[151] Thus, worldview congruence can also be conveyed through an appropriate choice of messenger to leverage the effects of source credibility.

On top of using worldview-congruent frames and credible sources, research also found that **self-affirmation** can make people less defensive against counter-attitudinal information. For instance, research found that people who were given an opportunity to affirm their basic values – e.g., write a few sentences about a time where they felt good because they acted on a value that was important to them – were more likely to respond positively to evidence that challenged their views on issues like the death penalty and abortion.[152] [153]

In sum, the strategy to counter the worldview backfire effect involves presenting retractions in worldview-congruent frames by credible sources and affirming people's self-identity in order to deliver corrective information without emotional challenge. That said, the worldview

[151] Callaghan, K., & Schnell, F. (2009). Who says what to whom: Why messengers and citizen beliefs matter in social policy framing. *The Social Science Journal*, *46*(1), 12-28.

[152] Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., & Ross, L. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *Journal of Personality and Social Psychology*, *93*(3), 415-430.

[153] Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and Social Psychology Bulletin*, *26*(9), 1151-1164.

backfire effect is often strongest among those fixed in their views. Thus, it would be strategic to target these efforts at the undecided majority instead of the unswayable minority as there would be a greater chance of correcting false information among the former.[154] [155]

4.6 "Pre-bunking" through inoculation

Often, a "firehose of falsehoods" cannot be countered by the "squirt gun of truths".[156] Instead of debunking each individual falsehood, it would be more effective to step back and expose the manipulation and hidden agenda. Inoculation can be used as a "pre-bunking" tool as it exposes the techniques and strategies used by those propagating false information, as well as the logical flaws inherent in the false information.[157]

As briefly mentioned at the start of *Section 4: Effective communication*, inoculation functions like a **"psychological vaccine"** where individuals are inoculated against false information that they may encounter in future by exposing logical fallacies a priori. Inoculation messages involve two

[154] Cook, J., & Lewandowsky, S. (2011). The debunking handbook. Retrieved from https://www.skepticalscience.com/docs/Debunking_Handbook.pdf.

[155] Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353-369.

[156] Paul, C., & Matthews, M. (2016). The Russian "firehose of falsehood" propaganda model. Retrieved from https://www.rand.org/content/dam/rand/pubs/perspectives/PE100/PE198/RAND_PE198.pdf.

[157] Ecker, U. K. (2017). Why rebuttals may not work: the psychology of misinformation. *Media Asia*, *44*(2), 79-87.

main components – 1) an explicit warning about an impending threat, and 2) a refutation of a pre-empted argument. The “threat” component signals to the individual that his or her position on an issue is susceptible to persuasion and change, while the “refutation” component provides information that they can mobilise to strengthen their attitudes and resist the persuasion. For instance, an inoculation message might include – 1) an explicit warning that there are attempts to cast doubt on the scientific consensus on climate change, and 2) that one of the methods of doing so is by referencing fake experts to feign a lack of consensus. Thus, the false information (i.e. lack of consensus on climate change) is being delivered to individuals in a weakened form. [158]

Research has demonstrated the effectiveness and potential of inoculation theory at combating false information. In a study by van der Linden *et al*. (2017), researchers tried to understand if it was possible to inoculate people’s beliefs about climate change.[159] The study involved three different messages – 1) a message that communicated scientific consensus on climate change, 2) false information that communicated no scientific consensus on climate change, and 3) an inoculation message which warned that politically and financially motivated groups were trying to convince the public that there was no scientific consensus

[158] Compton, J., Jackson, B., & Dimmock, J. A. (2016). Persuading others to avoid persuasion: Inoculation theory and resistant health attitudes. *Frontiers in Psychology*, *7*(122). 1-9.

[159] van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2). 1-7.

on climate change. The study found that presenting subjects with the consensus message alone resulted in a positive influence of perceived scientific agreement, and presenting the false information alone resulted in a negative influence of perceived scientific agreement. When the false information was presented after the consensus message, positive influence of perceived scientific agreement from the consensus message was largely negated by the false information. However, when the inoculation message was presented after the consensus message but before the false information, up to two-thirds of the positive influence of perceived scientific agreement from the consensus message was preserved. Since then, other studies have also found evidence that inoculating messages helped neutralise the adverse effects of false information about climate change. Furthermore, research has also found that a specific type of “reason-based inoculation” – one that is based on general critical thinking methods – offers a distinct advantage of being accessible to and effective for people who lack expertise (e.g., in climate science).[160] [161]

[160] Cook, J., Lewandowsky, S., & Ecker, U. K. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one*, *12*(5), e0175799.

[161] Cook, J., Ellerton, P., & Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, *13*(2), 024018.

In short, the findings from the aforementioned studies suggest that inoculation could be a useful approach to protect the public from false information.[162]

5. EDUCATION AND CULTIVATING LITERACY

In an age in which everyone is a publisher, where tweets and Facebook statuses are being reported as news, there is an increasingly need for Internet users to be competent and intelligent users of information. Measures such as the use of legislation, fact checking and removal of content, while important, happen after the falsehood. Thus, it is also necessary to strengthen critical literacy among citizens. Equipping citizens with critical thinking skills will have sustainable benefits in long run as it would boost their "immunity" to the different types of falsehoods circulating in our information ecology and future challenges. However, many studies have pointed to a general lack of literacy among the public:

- According to a survey done by Pew Research Center in 2017, 64% of US adults said that "fabricated news stories caused a great deal of confusion about the basic facts of current issues and events".[163]
- A study conducted by Stanford History Education Group found dismaying results regarding middle-school, high-school and college

[162] van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2). 1-7.

[163] Barthel, M., Mitchell, A., & Holcomb, J. (2016, December 15). Many Americans believe fake news is sowing confusion. Retrieved from http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowingconfusion/.

students' ability to assess online sources of information despite being digitally savvy. The study found that students were not able to distinguish fake accounts from real ones, differentiate information from activist groups versus information from neutral sources, and tell apart advertisements from articles.[164]

- In the UK, a 2017 YouGov survey commissioned by Channel 4 found that only 4% of people could correctly identify fake news.[165]
- Globally, the 2018 Edelman Trust Barometer found that 63% agreed that the average person did not know how to tell good journalism from rumours or falsehoods, and 59% agreed that it is harder to tell if a piece of news was produced by a respected media organisation.[166]
- In Singapore, the 2018 REACH poll found that only 1 in 2 respondents were confident of their own ability to discern fake news. Of those who felt they had seen fake news, 70% were not always able to discern falsehoods at the time they read information.[167]

According to Cooke (2018), a combination of critical information literacy, digital literacy and ultimately meta-literacy would enable information

[164] Donald, B. (2016, December 22). Stanford researchers find students have trouble judging the credibility of information online. Retrieved from https://ed.stanford.edu/news/stanford-researchers-find-students-have-trouble-judgingcredibility-information-online.

[165] C4 study reveals only 4% surveyed can identify true or fake news. (2017, February 6). *Channel 4 News*. Retrieved from http://www.channel4.com/info/press/news/c4-study-reveals-only-4-surveyed-can-identify-true-or-fake-news.

[166] 2018 Edelman trust barometer global report. Retrieved from https://cms.edelman.com/sites/default/files/2018-01/2018%20Edelman%20Trust%20Barometer%20Global%20Report.pdf.

[167] Findings of poll on attitudes towards fake news. (2018, March 26). Retrieved from https://www.reach.gov.sg/~/media/2018/press-release/media-release-on-findings-of-fake-news-poll-26-mar-2018.pdf.

consumers to seek, find, and use appropriate and quality information. Critical information literacy demands that users evaluate a piece of information in relation to the underlying power structures that shape all information, and apply this skill to different domains in their lives, Digital literacy, on the other hand, stresses on being *"skilled at deciphering complex images and sounds as well as syntactical subtleties of words"* (p. 18). Meta-literacy is an overarching framework that integrates various technologies and brings together different types of literacy, and placing emphasis on producing and sharing information in participatory digital environments.[168]

5.1 Informing people of their own biases

Although media, news and digital literacy have been recognised as an integral part of the solution to the problem of online falsehoods, literacy curricula could be strengthened by teaching psychological perspectives and designing activities that prompt students to reflect on their prejudices and biases, as well as their susceptibility to being trapped in online filter bubbles.[169] In particular, individuals must recognise the need to confront and grapple with one's own biases, predictions and worldviews, which is something increasingly difficult to engage in when one's views and

[168] Cooke, N. A. (2018). *Fake News and alternative facts: Information literacy in a post-truth era*. Chicago, Illinois: American Library Association.

[169] Nekmat, E., & Soon, C. (2017, September 22). Fake news mind traps. *The Straits Times*. Retrieved from https://www.straitstimes.com/opinion/fake-news-mind-traps.

convictions are often mirrored and echoed in online filter bubbles.[170] Others have also suggested that another way to overcome confirmation bias is to encourage people to engage in diverse social mixing so that people get exposed to a diverse range of views and meet others who challenge their personal assumptions.[171] [172] Furthermore, research has found that social endorsements can play a part in empowering individuals to broaden their social circles and improve quality of information flows to help overcome effects of selective exposure online.[173] [174]

Another important aspect relating to educating people about their own personal biases is to consider the role that technology companies play in mediating information, and thus, how they potentially exacerbate people's cognitive biases (as mentioned in *Section 3: Technocognitive solutions*).

[170] Gibson, C., & Jacobson, T. E. (2018). Habits of mind in an uncertain information world. *Reference & User Services Quarterly*, *57*(3), 183-192.

[171] Patel, D. (2017, May 7). How to overcome cognitive bias and use it to your advantage. *The Huffington Post*. Retrieved from https://www.huffingtonpost.com/entry/how-to-overcome-cognitive-bias-and-use-it-to-your-advantage_us_5900fff3e4b00acb75f1844f.

[172] Shah, J. (2016, December 1). How to use psychology to overcome your biases. Retrieved from https://www.forbes.com/sites/theyec/2016/12/01/how-to-use-psychology-to-overcome-your-biases/#2784c09e6d9f.

[173] Messing, S., & Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research*, *41*(8), 1042-1063.

[174] Vraga, E. K., & Bode, L. (2017). Leveraging institutions, educators, and networks to correct misinformation: A commentary on Lewandosky, Ecker, and Cook. *Journal of Applied Research in Memory and Cognition*, *6*(4), 382-388.

Some critical questions to ask include:[175]

- Why do people see different stories in their NewsFeed? Why and how should that matter?
- How do the technological processes that determine the content of what users see interact with their cognitive biases?
- How can information consumers be made more aware of their own cognitive biases to mitigate the possibility of technology companies exploiting these biases for commercial interests?

More literacy effort is certainly needed in this area, especially after a survey conducted by Pew Research Centre in 2018 found that a significant portion of Facebook users lacked a clear understanding of how Facebook's NewsFeed operates. The survey found that 53% of adults in the US do not understand why certain posts are included in their NewsFeed, with older users (50 years old and above) being more likely to say that they do not understand how NewsFeed works. Although Facebook offers a number of tools to help users customise the content that they see in their NewsFeed, only 14% of respondents felt that they have a lot of control over the content that appears there; 28% felt that they have no control at all. Again, older users were more likely to feel this way. Lastly, the survey also found that only about a third of respondents have actively tried to influence the content they see on their

[175] Lim, S. S. (2018, March 5). Submission to Select Committee on deliberate online falsehoods. Retrieved from https://www.parliament.gov.sg/docs/default-source/sconlinefalsehoods/written-representation-110.pdf.

NewsFeed – e.g., followed or unfollowed specific pages or people; changed their ad preferences etc. – with older users being less likely to do so.[176]

In sum, the findings of this survey demonstrate the importance of educating online users about how the algorithms influence the content surfaced to them. Without such educational and literacy efforts, it will remain difficult for information consumers to be conscious of the hidden biases in the news and information they encounter online.

5.2 Implementing misconception-based learning

Relating to the techniques of providing an alternative explanation (*Section 4.1: Filling the gap: Providing an alternative explanation*) and inoculation (*Section 4.6: "Pre-bunking" through inoculation*) is the concept of misconception-based learning. Misconception-based learning offers a powerful and practical way to apply inoculation in an educational setting. Similar to the principles espoused in *Section 4: Effective communication*, the technique involves teaching scientific concepts by explaining the misconceptions and how they distort science, and critiquing the techniques employed to mislead.[177] Research in the

[176] Smith, A. (2018, September 5). Many Facebook users don't understand how the site's news feed works. Retrieved from http://www.pewresearch.org/fact-tank/2018/09/05/many-facebook-users-dont-understand-how-the-sites-news-feed-works/.

[177] Cook, J. (2017). Understanding and countering climate science denial. In *Journal and Proceedings of the Royal Society of New South Wales* (Vol. 150, No. 465/466, p. 207). Royal Society of New South Wales.

field of education on correcting students' scientific misconceptions in the classroom has found that corrections were most successful when they included sufficient explanations of why a piece of misinformation was false and why the facts are true.[178] Instead, conventional strategies such as non-refutational explanations that simply presented the correct information without explaining the misconception were often successful only in the short term.[179]

The technique of misconception-based learning has been found to result in greater and longer-lasting learning gains; improve argumentative and critical thinking skills; and is more engaging to students as well.[180] [181] [182] One specific example of misconception-based learning (to evaluate climate change-related false information for instance) involves a six-step critical thinking process:[183] [184]

[178] Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly*, *28*(2), 117-159.

[179] Guzzetti, B. J. (2000). Learning counter-intuitive science concepts: What have we learned from over a decade of research?. *Reading & Writing Quarterly*, *16*(2), 89-98.

[180] McCuin, J. L., Hayhoe, K., & Hayhoe, D. (2014). Comparing the effects of traditional vs. misconceptions-based instruction on student understanding of the greenhouse effect. *Journal of Geoscience Education*, *62*(3), 445-459.

[181] Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological Science*, *22*(4), 545-552.

[182] Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education*, *93*(1), 26-55.

[183] Nuccitelli, D. (2018). Humans need to become smarter thinkers to beat climate denial. *The Guardian*. Retrieved from https://www.theguardian.com/environment/climate-consensus-97-per-cent/2018/feb/06/humans-need-to-become-smarter-thinkers-to-beat-climate-denial.

[184] Cook, J. (2017). Understanding and countering climate science denial. In *Journal and Proceedings of the Royal Society of New South Wales* (Vol. 150, No. 465/466, p. 207). Royal Society of New South Wales.

- **Step 1: Identify the claim being made.** For example, *"Earth's climate has changed naturally in the past, so current climate change is natural."*
- **Step 2: Construct the argument by identifying the premises leading to that conclusion.** In this case, the first premise is that Earth's climate has changed in the past through natural processes, and the second premise is that the climate is currently changing.
- **Step 3: Determine whether the argument is deductive** – i.e., starts out with a general statement and reaches a definitive conclusion. In this case, 'current climate change is natural' qualifies as a definitive conclusion.
- **Step 4: Check the argument for validity** – does the conclusion follow from the premises? In this example, it does not follow that current climate change must be natural because climate changed naturally in the past. Instead, the conclusion can be rephrased to, *"The current climate change may not be the result of human activity"*, which no longer refutes human-caused global warming.
- **Step 4a: Identify hidden premises.** By adding an extra premise to make an invalid argument valid, we can gain a deeper understanding of why the argument is flawed. In this example, the hidden assumption is *"if nature caused climate change in the past, it must always be the cause of climate change."* Adding this premise makes it clear why the argument is false – it commits single-cause fallacy, assuming that only one thing can cause climate change.
- **Step 5: Check to see if the argument relies on ambiguity.** For example, the argument that human activity is not necessary to explain

current climate change because natural and human factors can both cause climate change is ambiguous about the 'climate change' in question. Not all climate change is equal as the rate of current change is more than 20 times faster than natural climate changes. Therefore, human activity is necessary to explain current climate change.

- **Step 6: If the argument has not yet been ruled out, determine the truth of its premises.** For example, the argument that *"if something was the cause in the past, it will be the cause in the future"* is invalid if the effect has multiple plausible causes or mechanisms (as with climate change). This is where the myth most obviously falls apart.

Finally, research also shows that misconception-based learning or inoculation techniques may potentially be gamified for better engagement. For instance, an experimental pilot study found that the implementing the process of inoculation via an educational fake news game reduced people's perceived reliability and persuasiveness of fake news articles exposed to them. This finding suggests that educational games may serve as a promising vehicle to inoculate the public against falsehoods.[185]

[185] Roozenbeek, J., & van der Linden, S. (2018). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, 1-11.

5.3 Fostering scepticism

Research has found that certain attitudes can safeguard individuals against the effects of false information. One particular example of such an attitude is “scepticism”.

Contrary to popular belief, scepticism is not a form of evidence-denial that is driven by motivated reasoning (*Section 2.1: Confirmation bias and motivated reasoning*). Instead, scepticism is an awareness of potential hidden agendas and a desire to accurately and critically understand the evidence presented.[186] For instance, a study found that scepticism of the overall context of a piece of information led to more accurate processing of information presented and recognition of correct information, but yet did not translate into cynicism or a blanket denial of all information.[187] One reason why scepticism can mitigate the effects of falsehoods is because it leads to the allocation of more cognitive resources to the task of assessing the veracity of both the falsehood and the fact. When people are tasked to fact check, spot inconsistencies and correct inaccuracies, this increased deliberation nudges people to rely less on the false information.[188]

[186] Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, *26*(1), 283-327.

[187] Lewandowsky, S., Stritzke, W. G., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: The Iraq War 2003. *Psychological Science*, *16*(3), 190-195.

[188] Rapp, D. N., Hinze, S. R., Kohlhepp, K., & Ryskin, R. A. (2014). Reducing reliance on inaccurate information. *Memory & Cognition*, *42*(1), 11-26.

Furthermore, research has found that scepticism is a quality and skill that can be induced, taught and honed. For instance, a study found that a negative mood can increase scepticism and improve the ability to accurately detect deceitful communication.[189] Some scholars have also argued that scepticism can be taught in educational settings through critical thinking and evidence-based evaluation. One example would be to design activities that get students to identify pseudoscience or examine real-world false advertising claims in the media to highlight the prevalence of falsifiable claims in the public sphere. Another example would be to design activities that get students to create their own pseudoscience in order to demonstrate the ease by which "evidence" can be fabricated.[190]

In short, the ability to maintain doubt, question and scrutinise evidence – even when evidence is consistent with one's pre-existing beliefs and worldviews – can help people avoid unconsciously relying on falsehoods.

[189] Forgas, J. P., & East, R. (2008). On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology*, *44*(5), 1362-1367.

[190] Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, *6*, 888.

6. CONCLUSION

Since the publication of the first report, What Lies Beneath the Truth: A Literature Review on Fake News, False Information and More, there have been many developments in countering falsehoods on various fronts (e.g., by technology companies and fact checking organisations). In addition, the findings from recently conducted polls and surveys have shed light on the impact and effectiveness of various measures, which range from the flagging of falsehoods by technology companies, the design of retractions, leveraging social networks to spread debunking messages, to education and literacy initiatives.

In this working paper, we presented the salient measures, their strengths and limitations, specifically in overcoming the human biases at play during information seeking and processing. The paper underscores the need for a technocognitive approach where the design of information architectures is informed by disciplines such as psychology and behavioural economics, paying close scrutiny to minute details in the design and communication of corrective information, and raising the game when developing education and literacy programmes.

While the preceding sections have highlighted the limitations of some of these measures, the key takeaway is not to give up on them but find means to improve them, and design even more effective interventions. For instance, Facebook's now-defunct "Disputed Flags" and corrective messages that were not designed carefully contributed to the familiarity and worldview backfire effects. Cases such as these demonstrate the need for continual trial and improvement that is informed by testing and research. Research has found that identifying and understanding the specific process that creates belief echoes can help optimise the type of retraction that should be used to debunk the false information.

Some of the recommendations highlighted in the paper can be considered for initiatives that are already in place. They include targeting specific groups which could be more open to corrective information for debunking (e.g., the undecided majority as opposed to those who occupy extreme ends of the political spectrum), increasing public awareness for some of the tools that are already in the market for use (e.g., “filter bubble bursting” apps and browser extensions), and tailoring messages accordingly and using contextual cues for different groups. More attention can also be given to inoculating the public in anticipation of false content, especially when it comes to communicating policies that will generate high public interest (e.g., health, immigration policies, and taxation).

As for technology companies such as Facebook, Google, YouTube and Twitter, they are testing and rolling out initiatives at a quick pace. Some of the recommendations highlighted in this paper include content and comment moderation, use of more neutral labels, and collaboration among the companies to create a consistent set of intervention mechanisms (e.g., visual indicators to flag false information). While some of the suggestions put forth may contravene the companies' mandate and commercial interests (e.g., raising questions such as the extent to which technology companies should intervene in public discourse and the feasibility of creating a shared set of visual indicators), they are worthy of consideration, in light of empirical evidence.

In the report published last year, we had presented several initiatives pertaining to increasing media literacy among online users. The research we have highlighted in this working paper points to the need for programmes that go beyond discerning the message to specifically targeting users' psychology, attitudes and prejudices. Educators and literacy programme designers should look into content that is geared towards increasing awareness among people of their own personal biases, leveraging misconception-based learning, and cultivating scepticism. What is required in an increasingly complex information ecosystem is for people to consume information from diverse sources and confront views that are different from their own. Ultimately, a vigilant information consumer is the best defence against false information.

About IPS Working Paper Series

The IPS Working Papers Series is published in-house for early dissemination of works-in-progress. This may be research carried out by IPS researchers, work commissioned by the Institute or work submitted to the Institute for publication.

The views expressed in the Working Papers are strictly those of the author(s) alone and do not necessarily reflect the views of the IPS.

Comments on the Working Papers are invited. Please direct your comments and queries to the author(s).

IPS Working Papers are available from the IPS at $7.00 each (before GST). Postage and handling charges will be added for mail orders.

For more information, please visit www.lkyspp.nus.edu.sg/ips or
contact email: ips@nus.edu.sg or tel: 6516-8388 or fax: 6777-0700.

Institute of Policy Studies
Lee Kuan Yew School of Public Policy
National University of Singapore
1C Cluny Road House 5
Singapore 259599

Tel: (65) 6516 8388 Fax: (65) 6777 0700
Web: www.lkyspp.nus.edu.sg/ips
Registration Number: 200604346E