A Signal-Detection Framework for Misinformation Interventions

Bertram Gawronski, Lea S. Nahon, Nyx L. Ng
University of Texas at Austin

We propose a conceptual framework for misinformation interventions based on Signal Detection Theory. We highlight that different factors can lead people to fall for misinformation and call for interventions to be tailored to these factors.

Misinformation is widely regarded as a major threat to individuals and society. To mitigate the dangers of misinformation, researchers across disciplines have devoted considerable effort to develop interventions that reduce people's propensity to believe and share misinformation.[1] Yet, despite major advances, studies on the effectiveness of these interventions have used different research designs and different evaluation criteria, leading to conflicting conclusions about what can be done to reduce susceptibility to misinformation.[2] To address this problem, we propose a signal-detection framework for the development and evaluation of person-centered misinformation interventions. Signal Detection Theory (SDT) is a framework for measuring how people differentiate between patterns that bear information and those that are random. While SDT has originally been developed for research in psychophysics,[3] the theory can be applied to any decision problem involving bipolar responses to two stimulus classes. Expanding on a growing body of research using SDT to study responses to true and false information, a central proposition of the proposed framework is that three proximal factors can lead people to fall for misinformation. Because the relative impact of these factors can differ across applications (e.g., belief in versus sharing of misinformation), content domains (e.g., mundane vs. contentious topics), and contextual variables (e.g., degree of societal polarization), it is important to (1) determine why the focal population accepts misinformation and (2) develop interventions that are tailored to the nature of the problem. While prior work has developed SDT as a framework for research on misinformation susceptibility,[4] we aim to expand on this approach by explicitly discussing the implications of SDT for the development and evaluation of misinformation interventions.

A Signal-Detection Framework

To illustrate the core ideas of SDT as applied to misinformation susceptibility, consider the four potential cases concerning the acceptance versus rejection of true versus false information: (1) acceptance of true information; (2) acceptance of false information; (3) rejection of true information; and (4) rejection of false information.[4] Research on misinformation susceptibility is essentially concerned with the acceptance of false information: why do people believe or share false information and what can be done about it?[2]

According to SDT, there are two potential reasons why people may accept false information.[4] First, people may be unable to distinguish between true and false information. In this case, people would mistakenly accept a lot of false information and, at the same time, mistakenly reject a lot of true information. Second, people may have a general tendency to accept information regardless of whether it is true or false. In this case, people would mistakenly accept a lot of false information and, at the same time, correctly accept a lot of true information. Using SDT's terminology, the first case can be described as reflecting low truth sensitivity (or low discernment); the second case can be described as reflecting a low acceptance threshold.

Beyond these two basic factors, research suggests that people show lower acceptance thresholds for information that is congruent with their beliefs compared to information that is incongruent with their beliefs.[5] This finding represents an instance of a broader phenomenon known as *myside bias*: the tendency to evaluate information in a manner biased toward one's personal beliefs.[6] For example, in the political domain, Americans who identify as Democrats have been found to show a lower acceptance threshold for information with a pro-Democrat slant compared to information with a pro-Republican slant, while Americans who identify as Republicans showed the opposite effect.[5] Although myside bias is likely more pronounced for polarized compared to mundane topics, differential thresholds for belief-congruent and belief-incongruent information can occur for any issue on which a person holds prior beliefs.[6]

Together, these considerations suggest that three proximal factors can lead people to believe or share false information: (1) low truth sensitivity, (2) low overall threshold, and (3) lower threshold for belief-congruent compared to belief-incongruent information (i.e., myside bias). While other factors are important as well (e.g., digital literacy), these factors can be described as distal in the sense that their impact occurs via their effect on one (or more) of the three proximal factors (e.g., digital literacy influencing misinformation susceptibility via truth sensitivity). For the current analysis, it is also worth noting that the three proximal factors are independent in that each can vary without the other. For example, cognitive reflection has been found to increase truth sensitivity without affecting overall threshold or myside bias.[7] Conversely, greater subjective confidence in the accuracy of one's beliefs has been found to be associated with greater myside bias, while being unrelated to truth sensitivity.[5] Likewise, overall threshold has been found to be higher for sharing decisions compared to judgments of truth, but the higher threshold for sharing decisions was not associated with greater truth sensitivity or reduced myside bias.[5] Together, these results suggest that truth sensitivity, overall threshold, and myside bias have distinct psychological underpinnings, and therefore require different types of interventions. Hence, misinformation interventions will likely be most effective if they target the underlying reasons for why people believe or share false information: is it because they (1) are unable to distinguish between true and false information, (2) have a low overall threshold for accepting information, and/or (3) show a strong myside bias?

Relevance of the Three Factors

The significance of this argument can be illustrated with the results of a study that investigated the differential roles of the three proximal factors in veracity judgments and sharing decisions for political (mis)information in a sample of American participants who identified as either Democrat or Republican.[5] When participants were asked to judge whether the information presented is true or false (i.e., veracity judgments), truth sensitivity was quite high, indicating that participants were able to discern true from false information with a high degree of accuracy. In contrast, when participants were asked if they would share the information presented (i.e., sharing decisions), truth sensitivity was not significantly different from chance level, indicating that participants were as likely to share true information as they were to share false information. Myside bias, on the other hand, was very strong for both veracity judgments and sharing decisions. Interestingly, although truth sensitivity was substantially lower for sharing decisions compared to veracity judgments, overall threshold was much higher for sharing decisions than veracity judgments, indicating that greater reluctance in accepting information is not necessarily associated with greater accuracy. Applied to the current question, these results suggest that, in the domain of the study, (1) low truth sensitivity is a greater problem for sharing decisions than veracity judgments, (2) low overall threshold is a greater problem for veracity judgments than sharing decisions, and (3) myside bias is a significant problem for both veracity judgments and sharing decisions.

While these conclusions suggest that interventions have to be designed differently depending on whether they target belief in versus sharing of misinformation, it is worth noting that the relative impact of the three proximal factors may also depend on content-related and contextual variables.[8] For example, truth sensitivity may be more important in domains where people have relatively little knowledge compared to domains where people have considerable knowledge. Moreover, while myside bias may be a strong contributor to misinformation susceptibility for contentious topics and in highly polarized societies, myside bias might play a weaker role for mundane topics and in less polarized societies. Although it seems desirable to have universal interventions that reduce misinformation susceptibility irrespective of content domains and contextual variables, the effectiveness of any intervention likely depends on its fit to the nature of the focal problem.

Evaluating Interventions

Together, these considerations suggest that, prior to the development of any misinformation intervention, researchers should identify which of the three proximal factors (or combination of factors) is responsible for the acceptance of misinformation in their area of application: do people believe or share misinformation because they (1) are unable to distinguish between true and false information, (2) have a low overall threshold for accepting information in general, and/or (3) show a strong myside bias? Arguably, a given intervention will be more effective if it is tailored to the reason underlying why people believe or share false information. Interventions are likely less effective if they target a factor irrelevant to the problem one aims to address.

For the subsequent evaluation of misinformation interventions, we suggest a two-step approach that aligns with what can be deemed two levels of analysis in misinformation research (see Figure 1).[5] In a first step, researchers should test whether an intervention reduces acceptance of misinformation, as reflected in people's belief in and sharing of false information. In a second step, researchers should test whether the intervention affects truth sensitivity, overall threshold, and/or myside bias. Ideally, misinformation interventions would reduce acceptance of false information without reducing acceptance of true information.[2] That is, regardless of the nature of the focal problem, misinformation interventions should never reduce (and ideally increase) truth sensitivity.

However, despite the unequivocal importance of truth sensitivity in the evaluation of misinformation interventions, truth sensitivity should not be used as the sole target in the development of misinformation interventions. Such an approach could lead to attempts to increase people's ability to distinguish between true and false information even when low truth sensitivity is not the focal problem. Thus, while effects on truth sensitivity are always important to consider in the evaluation of misinformation interventions, a focus on truth sensitivity may or may not be relevant for the development of misinformation interventions. Moreover, in cases where low truth sensitivity is not part of the problem, an intervention may be effective even when it does not increase truth sensitivity (e.g., when it effectively reduces acceptance of false information via reduced myside bias without affecting truth sensitivity).

The proposed steps for the development and evaluation of misinformation interventions require designs that include both true and false information.[2] In addition, both types of information should vary in terms of whether it is congruent or incongruent with participants' beliefs. In this design, truth sensitivity is reflected in the difference between responses to true and false information; overall threshold is reflected in the overall acceptance of both true and false information; and myside bias is reflected in the difference between thresholds for information that is incongruent versus congruent with participants' beliefs. SDT provides an established and well-suited analytic approach to quantify the three factors, with SDT's *d'* index as a measure of truth sensitivity, SDT's *c* index as a measure of overall threshold, and the difference between *c* indices for belief-incongruent and belief-congruent information as a measure of myside bias.[4]

One example illustrating the value of SDT for research on misinformation interventions involves the effectiveness of gamified inoculation interventions.[9] While these interventions have been found to effectively reduce acceptance of misinformation, a reanalysis using SDT suggests that the observed reductions are driven by increased overall thresholds for accepting information rather than increased truth sensitivity.[10] However, research in this area has not identified which of the three proximal factors is responsible for the acceptance of misinformation in the focal domains. The reanalysis also did not include myside bias as a criterion (because the original studies did not manipulate belief-congruence of the presented information), entailing the possibility that these interventions may increase (rather than decrease) myside bias. Addressing these questions requires additional work to identify whether acceptance of misinformation in the focal domains is rooted in low truth sensitivity, low overall thresholds, or myside bias. In addition, studies on the effectiveness of gamified inoculation interventions should include true and false information that is congruent or incongruent with participants' beliefs, which is critical for gauging the impact of these interventions on myside bias.

While our recommendations may be easier to implement for interventions that pre-emptively target broader causes of misinformation susceptibility (e.g., inoculation) compared to interventions that retroactively target specific ideas or narratives (e.g., debunking),[1] we hope that the proposed framework will aid research on misinformation interventions by supporting (1) the identification of the specific reason(s) why people fall for misinformation in the focal area of application, (2) the development of interventions that are tailored to the nature of the focal problem, and (3) the evaluation of these interventions in terms of the problem one is trying to address.

References

(1) Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., ... & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13-29.

(2) Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour*, *7*(8), 1231-1233.

(3) Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Wiley.

(4) Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*, *17*(1), 78-98.

(5) Gawronski, B., Ng, N. L., & Luke, D. M. (2023). Truth sensitivity and partisan bias in responses to misinformation. *Journal of Experimental Psychology: General, 152*(8), 2205–2236.

(6) Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259-264.

(7) Sultan, M., Tump, A. N., Geers, M., Lorenz-Spreen, P., Herzog, S. M., & Kurvers, R. H. (2022). Time pressure reduces misinformation discrimination ability but does not alter response bias. *Scientific Reports*, *12*(1), Article 22416.

(8) Van Der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460-467.

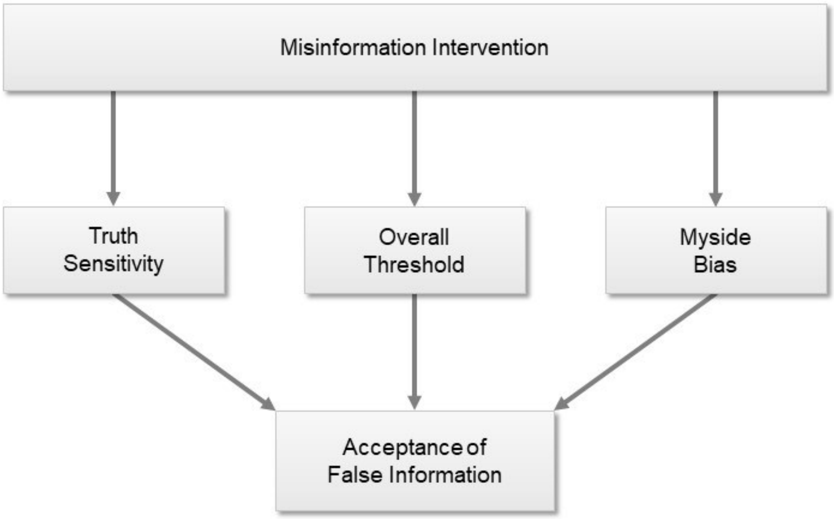
(9) van der Linden, S. (2024). Countering misinformation through psychological inoculation. *Advances in Experimental Social Psychology*, *69*, 1-58.

(10) Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General*, *152*(9), 2411-2437.

Acknowledgements

This work was supported by the National Science Foundation (BCS-2040684), the Swiss National Science Foundation (P500PS_214298), and the John Templeton Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. The funders had no role in the decision to publish or the preparation of this manuscript. The authors declare no competing interests.

Figure 1. General framework for the development and evaluation of misinformation interventions. Acceptance of false information can arise from low truth sensitivity, low overall threshold, or myside bias. An intervention may reduce acceptance of false information by increasing truth sensitivity, increasing overall threshold, or reducing myside bias. The effectiveness of a given intervention is assumed to depend on whether it is targeting the critical factors underlying acceptance of false information, which may differ across applications of the intervention (e.g., for reducing belief in versus sharing of false information), content domains (e.g., contentious versus mundane topics), and contextual variables (e.g., political polarization).


Misinformation Intervention
Truth Sensitivity
Overall Threshold
Myside Bias
Acceptance of False Information