Tackling Persistent Misperceptions: Repeated Interactions, Repeated Refutations, and Endogenous Corrections

Research Proposal (Draft)
Instructor: Prof. Sunita Parikh

Weiye (Rex) Deng
Washington University in St. Louis

December 30, 2020

1 Introduction

The prevalence of political misinformation has been increasingly haunting the world (Flynn, Nyhan, and Reifler, 2017). Given that political misinformation can result in undesirable political outcomes including lower willingness of political participation (Jolley and Douglas, 2014), lower trust in government (Einstein and Glick, 2015; Huang, 2017), and higher odds of social protests (Huang, 2017), corrections of such misperceptions become imperative for governments regardless of the regime types. However, studies have found that the corrective measure by providing direct refutations often fails because of persistent misperceptions attached to the misinformation (Nyhan and Reifler, 2010; Flynn, Nyhan, and Reifler, 2017). There are multiple explanations that refer to individual predispositions and psychological mechanisms to account for this counterproductivity, on which suggestions are also based to improve the efficacy of direct refutations (e.g., Kuklinski et al., 2000; Taber and Lodge, 2006; Oliver and Wood, 2014; Fridkin, Kenney, and Wintersieck, 2015; Thorson, 2016; Berinsky, 2017; Flynn, Nyhan, and Reifler, 2017; Flynn and Krupnikov, 2019; Swire-Thompson et al., 2020).

Most of these studies that have illuminated political misinformation as a conundrum and suggested effective ways to tackle it implicitly presumes a *one-shot interaction* with the misinformation in their research design. Under this presumption, these studies evaluate the efficacy of direct refutations by designing an experiment that exposes treatment groups to a single piece of misinformation and the rebuttal (e.g., Nyhan and Reifler, 2010; Thorson, 2016; Berinsky, 2017). This research design may capture a partial causal effect of the direct refutation, but the empirical fact that citizens often *repeatedly interact* with a piece of misinformation or pieces in similar topics may render the experiment of one-shot interaction too simplified and more importantly, it leaves two sets of questions not fully answered. First, even if a direct refutation is not effective to counter the one piece of misinformation, will the respondents continue to adhere to, or gradually abandon their false beliefs when receiving repeated direct refutations to the next pieces of misinformation in a similar topic? Which individual predispositions and psychological mechanisms can explain the results? As will be discussed below, even though most failures of corrective measures can be attributed to the psychological mechanisms of directional motivated reasoning and individuals' predispositions, there is some evidence that shows people are still able to think analytically and pursue accuracy goals so that they can properly reject the misinformation (e.g., Flynn, Nyhan, and Reifler, 2017; Pennycook and Rand, 2019). Therefore, answering this set of questions is meaningful to better assess the effectiveness of direct refutations in a more realistic setting, which allows us to provide suggestions to circumvent the "misinformation trap" and address it more effectively.

Second, it is also prevalent that individuals repeatedly interact with a single piece of misinformation. This fact is especially salient in the era of social media since people are primarily exposed to a subset of information filtered by their social connections (Anspach and Carlson, 2017), which increases the probability of repeated encounters with the same misinformation. For example, if citizens read a piece of misinformation in a news post, it is highly likely that they will read the same misinformation from their friend's repost or comments, since the content they read is closely related to what their social networks endorse (Anspach, 2017). When this type of repeated exposure to a single piece of misinformation is available, the rumor refutations may be not only *exogenously* made by the government or other independent fact-checking institutions, but also arises *endogenously* from the directly involved parties other than the government or third parties. For example, a once widespread rumor in Sina Weibo, one of most popular social media platforms in China, stated that a twelve-year-old girl was raped by two teachers, but "surprisingly", the police and the government totally ignored and even pampered such criminal behaviors. Expectedly, this post greatly instigated the netizens' sympathy to the girl and censures towards the government despite some suspicions on authenticity from other netizens. Nonetheless, this incident was clarified as a rumor by the girl herself several days later, who apologized for causing huge disturbances because she simply made it up "for fun". Then, the government and police followed up with this incident and also refuted the rumor. Such "plot reversal" is a classic example of endogenous corrections of misinformation, where the directly involved party (the girl in this example) instead of the government first corrects the misinformation. If such "plot reversal" indeed happens in some subset of misinformation, do exogenous refutations still necessarily fail? The answer can be no since a trustworthy institution (including the government) can strategically (or inadvertently) refute misinformation to maximize the effect of both endogenous and exogenous correction. That is, if an exogenous correction made by this institution closely follows an endogenous change, direct refutations may be more effective because of a two-fold corrective effect. However, we may expect the effect of direct refutation wanes if only either one type of correction is available within a short period, which resembles the case of one-shot interaction.

This research proposal takes one of the first steps to reevaluate the effect of direct refutations within the setting of multiple interactions. A literature review in the following covers two sections, one concerning how misperceptions are formed, and the other concerning why corrective measures are often counterproductive because of persistent misperceptions. By doing so, I stress both the theoretical and empirical significance of considering the setting of multiple interactions to reevaluate the causal effect of direct refutations and possible mechanisms that constitute my hypotheses under this new experimental setting. The discussion ends with some brief thoughts on data and experimental design.

2 The Origin of Misperceptions

Scholars have extensively investigated the psychological origins of false beliefs in misinformation. One of the earliest and most notable theoretical foundation in the context of political misperceptions is directional motivated reasoning. Kuklinski et al. (2000) argues that people tend to maintain a balance between their prior beliefs and the information they newly receive, on which they base to make reinforcing inferences instead of an accurate one. That is, they stick to a directional motivated reasoning to process information and form attitudes. Later studies support and extend this claim with more experimental evidence and theoretical implementations. Taber and Lodge (2006) finds that directional motivated reasoning drives citizens to not only adhere to their prior attitudes by favoring supporting arguments, but make them possess both disconfirmation bias which leads them to pay extra scrutiny to those information against their prior attitudes, and confirmation bias which leads them to self-select the evidence to support their beliefs. This signifies that even misinformation is likely to be accepted when citizens pursue directional goals, as long as they are congruent with citizens' priors. Flynn, Nyhan, and Reifler (2017) further summarize that individuals' prior attitudes, partisanship, and identity threats are three major sources of directional motivated reasoning (see also Taber and Lodge, 2006; Kahan, 2013, 2017). Moreover, by arguing that pursuing directional goals should precede accuracy goals in affectively charged contexts (see also Redlawsk, 2002), they suggest that political misperceptions resulting from more controversial issues should be attributed to directional motivated reasoning.

While directional motivated reasoning may not be the only origin of false beliefs in misinformation, it may be most influential one because alternative mechanisms are not as feasible in the current social structure. For example, Sunstein and Vermeule (2009) argues that people hold false beliefs such as conspiracy theories because of a lack of information sources in isolated networks, even though information that counters their claims are available in the wider society. However, such argument seems implausible when social media have greatly perpetrated daily life and widened access to and sources of information. Thus, if people adhere to false beliefs, it is more likely that people are subjectively willing to do so (e.g., motivated reasoning) rather than face objective obstacles that impede information acquisition.

In addition, scholars have substantially delved into how individual covariates, sometimes strongly interacting with the psychological mechanism of directional motivated reasoning, play a role in forming political misperceptions, and two main factors they identify are ideological predispositions and political sophistication. For the former, political ideologies (e.g., Jost et al., 2003; Nyhan and Reifler, 2010; Miller, Saunders, and Farhart, 2016; Uscinski, Klofstad, and Atkinson, 2016) and partisanship (e.g., Taber and Lodge, 2006; Kahan, 2013; Flynn, Nyhan, and Reifler, 2017; Kahan, 2017) as two main types of ideological predispositions are intimately related to false beliefs. Generally, these studies suggest that higher level of political conservatism or partisanship positively indicates the level of susceptibility to false beliefs. For the latter, those with higher political sophistication tend to be more prone to false beliefs, possibly because they are more able to make connections between abstract principles and concrete examples and hold stronger prior attitudes, which leads to a stronger directional motivated reasoning (e.g., Taber and Lodge, 2006; Miller, Saunders, and Farhart, 2016).

3 The Persistence of Misperceptions and Evaluation of Corrective Measures

In the above section, I show that the previous literature has substantially discussed the psychological mechanism of directional motivated reasoning and individuals' predispositions as the dominant explanations for the formation of political misperceptions. Furthermore, such mechanisms and individual covariates also largely explain why people stick to their false beliefs even after they are debunked. That is, those individual covariates that contribute to the formation of false beliefs, including prior beliefs, partisanship, ideologies, and political sophistication, are also more likely to make these misperceptions persistent and resistant to corrections. From the standpoint of directional motivated reasoning, people hold persistent false beliefs because the process of forming false beliefs are also a reinforcing one (Kuklinski et al., 2000). Thorson (2016) also links this belief persistence with motivated reasoning by stressing a bidirectional relationship between information and opinion. For individual characteristics, Nyhan and Reifler (2010) find that those most committed to their political ideologies even increase their political misperceptions after they receive a corrective message (i.e., a backfire effect). Swire-Thompson et al. (2020) also show that corrective measures are more effective for non-supporters of Donald Trump or Bernie Sanders than their supporters.

While noting the limitations of direct corrections for false beliefs, scholars provide some suggestions to tackle them by circumventing the traps of reinforcing misperceptions. For example, Lewandowsky et al. (2012) recommend repeated retractions of fake news and rumors to enhance their salience, but he also notes the risk of increasing people's familiarity to the misinformation and hence increase their acceptance (i.e., the effect of psychological fluency). Similarly, Berinsky (2017) also points out the effect of increasing psychological fluency with misinformation by merely repeating itself, and suggests that to override such effect, partisan refutations work more effectively than ones made by non-partisan members. Besides, Cobb, Nyhan, and Reifler (2013) indicate that not all types of misinformation would necessarily lead to belief persistence. They argue that discrediting false positive information that bogusly accredit politicians is easier than false negative information. While some scholars have also questioned the effectiveness of such corrective measures, arguing that people's false beliefs still shape their attitudes even refutations have been effective (i.e., the effect of "belief echoes") (Thorson, 2016).

4 In the Setting of Repeated Interactions

Up to now, studies that have been discussed mostly, if not explicitly, presume that individuals interact with the piece of misinformation in their research design, and few studies have explicitly embedded a setting of repeated interaction with a single piece of misinformation or pieces under a specific topic. However, based on the current literature, I argue that this is not only a methodological refinement that may better gauge the causal effect of direct refutations in a more realistic setting, but also extensively speak to and implement the theoretical foundations of corrective measures for misperceptions established by the current literature. First, even though the mechanism of directional motivated reasoning and psychological fluency indicate that the more frequently people interact with a piece of misinformation, the less likely they accept corrections of their false beliefs, there have been some suggestions that repeated refutations by intentionally emphasizing facts are conducive to contractions of misperceptions (e.g., Lewandowsky et al., 2012). Moreover, there have been also some challenges to the predictions made by theories in motivated reasoning, which necessitates the reevaluation of the efficacy of corrective measures. For example, while highlighting that directional motivated reasoning in affective contexts such as political information, Flynn, Nyhan, and Reifler (2017) also underline the accuracy goals as a great balance. Indeed, recent psychological research has discovered that more repetitions of misinformation are conducive to the retraction of their false beliefs because the explicit corrections makes the falsity of misinformation salient, which leads to a stronger updating upon encoding the corrections (Ecker, Hogan, and Lewandowsky, 2017). Wood and Porter (2019) further provide experimental evidence to show that people also heed corrections of misinformation even such corrections counter their ideological constraints. Meanwhile, Pennycook and Rand (2019) also suggest that the proneness to fake news can be alternatively better explained by a lack of reasoning, instead of motivated reasoning. Therefore, it can be inferred that if people receive multiple pieces of political misinformation in a similar topic with rebuttals to each of them, multiple psychological mechanisms and their own traits will intertwine, so the causal effect of direct refutations need to be gauged in a setting where multiple interactions are available.

Meanwhile, people may also encounter various types of refutations when they repeatedly interact with even only a single piece of misinformation, which has been so prevalent because of people's increasing engagement with social media. For political misinformation that incorporates multiple parties, I classify the refutations as either endogenous or exogenous. Endogenous corrections are made by those directly involved parties in the misinformation, while exogenous corrections are made by independent institutions not involved in the misinformation, such as the government and fact-checking institutions. Such different types of refutations are always available and highly likely to get repeatedly exposed to those who have an interest or knowledge about it in the era of social media. Since some scholars have found that people heed the source when they evaluate the misinformation Swire-Thompson et al. (2020), it is naturally to posit that people also notice the sources of refutations when determining whether to accept them. Therefore, in these cases, the evaluation of the effectiveness of refutations requires an explicit distinguishment of different types of refutations, because whether people encounter a certain type of refutation or the order of encounters could result in completely different conclusions. In the above example where the girl made up the rumor, it can be expected that people who receive both endogenous corrections from the girl and exogenous corrections or suspicions from independent netizens revise revise their false beliefs to a different extent from those who only receive one type of corrections. Analogously, people who only receive endogenous corrections from the girl may not revise their belief to the same extent as those who only receive exogenous corrections from independent netizens or the government. However, most previous studies that adopt an experimental setting of one-shot interaction cannot distinguish this difference, which may lead to estimation bias of the causal effect of corrective effort.

Though it is necessary to design a setting where people repeatedly interact with misinformation to more precisely gauge the causal effect of corrective measures given their methodological and empirical significance, there have been competing mechanisms and results from previous literature, which makes it complicated to hypothesize what will be the results in such a setting. First, for the case where people repeatedly interact with several pieces of misinformation and their rebuttals under the same topic, the model of directional motivated reasoning and psychological fluency predict that people will stick to their persistent misperceptions (e.g., Kuklinski et al., 2000; Taber and Lodge, 2006; Nyhan and Reifler, 2010; Berinsky, 2017), whereas recent studies that cite the psychological model of salience Ecker, Hogan, and Lewandowsky (2017) and inclination to avoid cognitive effort Pennycook and Rand (2019); Wood and Porter (2019) indicate that people can accede to the corrections, especially repeated ones. Drawing the insights from two strands of literature, I find it reasonable that the willingness to correct misperceptions will prevail their attachment to the misinformation primarily because repeated corrections of several pieces of misinformation under the same topic makes the falsity of each misinformation more salient, which can be sufficient to accept corrections, though people subject to ideological constraints may be still less willing to do so. Therefore, I hypothesize that *people will heed and accept corrections of misinformation in different topics as the frequencies of repetitions increase, but such effect will be moderated by people's ideological constraints and partisanship, albeit still positive. More specifically, if the corrections are from the government that shares the same partisan label as the respondents, they are more willing to accept the repeated corrections than the non-partisan counterparts.*

Based on the recent insights on how corrections increase salience and therefore likelihood to reject misinformation, for the case where people repeatedly interact with one piece of misinformation but both endogenous and exogenous corrections are possible, I hypothesize that *people who receive both endogenous and exogenous corrections are more likely to reject the misinformation that those who only receive one type of corrections. Among the latter group, those who only receive endogenous corrections are more willing to reject the misinformation that those who only receive exogenous corrections.* This is because endogenous corrections are more likely to arouse the perceptions of falsity.

5 Experimental Design (Proposed)

I will first evaluate the performance of repeatedly refuting pieces of misinformation in a similar topic. Specifically, I randomly assign respondents into four groups. Half of the first group sequentially read three pieces of unrelated news and one piece of misinformation with a direct refutation from the same partisan government (hereafter combo one), the other half do the same except the direct refutation is from a non-partisan government. The second sequentially read two pieces unrelated news, then combo two, and combo one. The third group sequentially read one piece of unrelated news, then combo three, combo two, and combo one. The last group sequentially read combo four, then combo three, combo two, and combo one. Except the first group, the other three groups all read refutations from same-partisan government. I artificially construct all pieces of misinformation to eliminate the bias from any previous impressions, while the news is chosen from reality and guaranteed to be as objective as possible. I also select several potential moderators based on previous literature, including partisanship, political ideologies, political trust, and other information about individual background, to see to what extent the results can change.

Second, I explore the dynamics of direct refutations by providing a scenario where respondents repeatedly interact with one piece of misinformation. In this article, I present one possible way to create such scenario where direct refutations may make a difference. That is, I allow an endogenous correction made by the implicated parties are available for the treatment group. Specifically, both the treatment and control groups with a piece of misinformation at the first stage. Second, the control group reads an unrelated piece of news, while two treatment groups both read a piece of news that an implicated party of the misinformation refute the misinformation with strong evidence. After that, the control group and the treatment group one read a direct refutation made by an independent institution, but treatment group two reads an unrelated piece of news.

6 Limitations

There are several limitations of this research proposal. First, I cannot incorporate all possible scenarios of repeated interactions with misinformation. In this proposal, I include the repeated interactions with several pieces of misinformation refutations from the same source under the same topic and one specific piece of misinformation with different sources of refutations. However, there can be other possible scenarios where the effect of direct refutations is worth more investigation because they are also realistic but only more complex. For example, people may receive both endogenous and exogenous corrections for one piece of misinformation, but they may only receive endogenous correction for the next piece. What can be the effect of direct refutations in the second case.

The second limitation is about the second experimental design. Since political misinformation always involves the government, its corrective effort is considered as endogenous since it is not independent from the misinformation. In such cases, I have to limit the provider of exogenous corrections within other fact-checking institutions or independent citizens. However, this limitation makes it hard to probe how political trust in the government may interact with people's willingness to reject misinformation, only if the scope of misinformation can be expanded into nonpolitical ones.

7 Bibliography

Anspach, Nicolas M. 2017. “The New Personal Influence: How Our Facebook Friends Influence the News We Read.” *Political Communication* 34(4): 590–606.

Anspach, Nicolas M., and Taylor N. Carlson. 2017. “What to Believe? Social Media Commentary and Belief in Misinformation.” *Political Behavior* 42: 691–718.

Berinsky, Adam J. 2017. “Rumors and Health Care Reform: Experiments in Political Misinformation.” *British Journal of Political Science* 47: 241–262.

Cobb, Michael D., Brendan Nyhan, and Jason Reifler. 2013. “Beliefs Don’t Always Persevere: How Political Figures Are Punished When Positive Information about Them Is Discredited.” *Political Psychology* 34(3): 307–326.

Ecker, Ullrich K.H., Joshua L. Hogan, and Stephan Lewandowsky. 2017. “Reminders and Repetition of Misinformation: Helping or Hindering Its Retraction?” *Journal of Applied Research in Memory and Cognition* 6(2): 185–192.

Einstein, Katherine Levine, and David M. Glick. 2015. “Do I Think BLS Data are BS? The Consequences of Conspiracy Theories.” *Political Behavior* 37: 679–701.

Flynn, D. J., and Yanna Krupnikov. 2019. “Misinformation and the justification of socially undesirable preferences.” *Journal of Experimental Political Science* 6(1): 5–16.

Flynn, D. J., Brendan Nyhan, and Jason Reifler. 2017. “The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics.” *Political Psychology* 38(S1): 127–150.

Fridkin, Kim, Patrick J. Kenney, and Amanda Wintersieck. 2015. “Liar, Liar, Pants on Fire: How Fact-Checking Influences Citizens’ Reactions to Negative Advertising.” *Political Communication* 32(1): 127–150.

Huang, Haifeng. 2017. "A War of (Mis)Information: The Political Effects of Rumors and Rumor Rebuttals in an Authoritarian Country." *British Journal of Political Science* 47(2): 283–311.

Jolley, Daniel, and Karen M. Douglas. 2014. "The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint." *British Journal of Psychology* 105(1): 35–56.

Jost, John T., Jack Glaser, Arie W. Kruglanski, and Frank J. Sulloway. 2003. "Political Conservatism as Motivated Social Cognition." *Psychological Bulletin* 129(3): 339–375.

Kahan, Dan M. 2013. "Ideology, motivated reasoning, and cognitive reflection." *Judgment and Decision Making* 8: 407–424.

Kahan, Dan M. 2017. "Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition." *Cultural Cognition Project Working Paper Series No. 164*: Forthcoming.

Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. 2000. "Misinformation and the currency of democratic citizenship." *Journal of Politics* 62(3): 790–816.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. "The science of fake news." *Science* 359(6380): 1094–1096.

Lewandowsky, Stephan, Ullrich K.H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest, Supplement* 13(3): 106–131.

Miller, Joanne M., Kyle L. Saunders, and Christina E. Farhart. 2016. "Conspiracy Endorsement as Motivated Reasoning: The Moderating Roles of Political Knowledge and Trust." *American Journal of Political Science* 60(4): 824–844.

Nyhan, Brendan, and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32: 303–330.

Oliver, J. Eric, and Thomas J. Wood. 2014. "Conspiracy theories and the paranoid style(s) of mass opinion." *American Journal of Political Science* 58(4): 952–966.

Pennycook, Gordon, and David G. Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188: 39–50.

Redlawsk, David P. 2002. "Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making." *Journal of Politics* 64(4): 1021–1044.

Sunstein, Cass R., and Adrian Vermeule. 2009. "Symposium on conspiracy theories: Conspiracy theories: Causes and cures." *Journal of Political Philosophy* 17(2): 202–227.

Swire-Thompson, Briony, Ullrich K.H. Ecker, Stephan Lewandowsky, and Adam J. Berinsky. 2020. "They Might Be a Liar But They're My Liar: Source Evaluation and the Prevalence of Misinformation." *Political Psychology* 41(1): 21–34.

Taber, Charles S., and Milton Lodge. 2006. "Motivated skepticism in the evaluation of political beliefs." *American Journal of Political Science* 50(3): 755–769.

Thorson, Emily. 2016. "Belief Echoes: The Persistent Effects of Corrected Misinformation." *Political Communication* 33(3): 460–480.

Uscinski, Joseph E., Casey Klofstad, and Matthew D. Atkinson. 2016. "What Drives Conspiratorial Beliefs? The Role of Informational Cues and Predispositions." *Political Research Quarterly* 69(1): 57–71.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359: 1146–1151.

Wang, Chengli, and Haifeng Huang. 2020. “When “Fake News” Becomes Real: The Consequences of False Government Denials in an Authoritarian Country.” *Comparative Political Studies*: Forthcoming.

Wood, Thomas, and Ethan Porter. 2019. “The Elusive Backfire Effect: Mass Attitudes’ Steadfast Factual Adherence.” *Political Behavior* 41: 135–163.