Protecting Free Speech While Countering Internet Misinformation: The Rebuttal Strategy

ABSTRACT

Censorship as a way to protect the integrity of political debate and thwart malign forces has ancient problems. The alternative is to find ways to reconstruct the public sphere in order to curtail the effects of propaganda, misinformation, and disinformation. Most fundamentally the strategy should be to give truth (or at least rebuttals) wings equal to the falsehoods they pursue. The starting point for a solution is a new option to tag a news story, webpage, tweet, blog post or a link shared on social media with a 'rejoinder' or 'rebuttal'. This study evaluates the rebuttals proposal as a potential means of reducing the radicalization and polarization engendered by fake news and opinion bubbles on social media using simulations to evaluate the implications of rebuttals for the spread of fake news in an agent-based SIR model. Rebuttals that accompany the misinformation dramatically reduce the spread of misinformation even when the rebuttal is less likely to persuade recipients than the misinformation.

Jesse Richman

Associate Professor of Political Science and International Studies

Old Dominion University

jrichman@odu.edu

Introduction

The problem of misinformation is hardly new. More than three hundred years ago, the humorist Jonathan Swift wrote “Falsehood flies, and the Truth comes limping after it.” (Quote Investigator 2014) And so it has proved to be in social media. Vosoughi et. al. found that “False news reached more people than the truth; the top 1% of false news cascades diffused to between 1000 and 100,000 people, whereas the truth rarely diffused to more than 1000 people. Falsehood also diffused faster than the truth.” (2018, p. 1146). Swift was right. The attention-getting falsehood goes viral, especially the falsehood that spreads animosity for out-groups (Rathje et. al. 2021) and carries high emotional intensity (Brady et. al. 2017). The factual version often fails to follow, and all too often no one finds the fact check unless they go looking for it.

For the last several years social media and internet companies have grappled with the challenge of how to manage or contain the spread of disinformation and misinformation on their sites (Barrett et. al. 2021). The stakes in this struggle are high as current platform configurations appear to be corroding democracy and trust, especially in established democracies (Lorenz-Spreen et. al. 2022). Haidt (2022) argues that social media virality is rendering American life “uniquely stupid.” This paper focuses on a proposal for a crowd-sourced approach to connecting truth with falsehood as it spreads, and evaluates through simulations this technological proposal to reduce the spread of misinformation – incorrect information that is spread by people who think it is true. A promising potential implementation of an idea similar to the one proposed is being explored by Twitter, which has named the initiative @Birdwatch.

Misinformation is present when an individual “firmly [holds] the wrong information.” Kuklinski et al. (2000, p. 792). In a review of the literature on political misinformation Jerit and

Zhao (2020) note that misinformation leads individuals to sometimes take incorrect actions, and as these actions cumulate, it can undermine democratic decision-making if it systematically favors one side (p78) by leading to incorrect aggregate choices. In the current literature, misinformation is applied not only to individual beliefs, but also to information sources stemming from elite debate or media (Jerit and Zhao 2020).

Belief in misinformation has both informational / cognitive and motivational aspects, which makes combatting it difficult. Often individuals believe misinformation that is consistent with other attitudes, identities, or commitments (Jerit and Zhao 2020). They are therefore likely to be motivated to avoid accepting the correction. Brian Weeks writes "It is clear that corrections work in some circumstances but not others. What is not apparent is why or how corrections succeed or fail when one is attempting to challenge partisan-based claims. This is a critical question that must be answered" (Weeks 2018, p. 148).

The literature suggests that one of the most effective strategies to combat misinformation is providing a factually based counterargument. Jerit and Zhao write

> "Research in cognitive psychology has led to specific recommendations about how to correct misinformation. For example, providing an alternative factual account is particularly effective because a person can replace the debunked misinformation with the alternative explanation." (2020, p. 82)

But they note that efforts to empirically test this approach have met mixed results, arguing that these differences can be explained partly in terms of the extent to which political identity / ideology is tightly linked to the misinformation, and partly on the basis of the credibility of the source. The tightest links to identity may produce "backfire" effects (Nyhan and Reifler 2010) but in instances where identity is less salient and the source is most credible, correction is more likely to succeed. Most efforts to replicate the Nyhan and Reifler result have failed. Wood and Porter (2019) sum up the results of a failed 52-issue effort to replicate the effect as follows: "By and large, citizens heed factual information, even when such information challenges their ideological commitments." Thus, providing rebuttals that correct false beliefs is an effective strategy.

Building beyond such laboratory studies, then, the fundamental challenge facing free societies in both the United States and around the world, is how to reconfigure the public sphere to reduce the spread of misinformation. How to keep falsehood from flying? Or how to help truth keep up with it.

Currently, censorship is being tried by some as a strategy to block disinformation and limit the spread of misinformation. The implementation of this solution began in force after the 2016 presidential election, and became much more intense by the 2020 presidential election. It has involved the selective identification of accounts for removal or suspension, and attaching fact checks to posts, linked articles, and tweets that have been identified as potential vectors of fake news. It has recently expanded to the wholesale removal of apps and websites believed to be vectors of propaganda, radicalization, and misinformation (e.g. the delisting of Parler by Google and Apple, followed by the cancelling of the entire Parler website by Amazon in January 2021. This strategy for coping with the problem of misinformation involves censorship of specific platforms and voices.

Faced with the problem of falsehood flying, censorship as a way to protect the integrity of political debate and thwart malign forces has long been a go-to strategy for elites, and it has ancient well-known problems. It often entails collateral damage.[1] And it can all too easily be turned to the service of elite agendas instead of the truth. Critics of censorship efforts in social media cite prominent examples of the censors potentially getting the story wrong, as with the censorship of stories concerning the possibility that the Covid-19 pandemic resulted from a lab leak, and the discussion of censorship and suppression of stories about the Hunter Biden laptop based upon claims it was disinformation that have turned out to be incorrect. Even if censors are well informed and well intentioned, they will occasionally get stories wrong, and this will corrode their credibility and their ability to effectively accomplish their goals. Studies suggest that Republicans in particular have become wary of "fact check" organizations, perhaps reflecting the tendency for those organizations to give Republican politicians lower ratings (Richman and Richman 2021).

In Federalist 10, perhaps the most profound of the Federalist Papers, James Madison wrestled with what is in part the problem of misinformation -- the problem of factions adverse to the rights of other citizens on the long term interests of the community. Madison noted that one possible strategy for dealing with these evil actors is to try to label and suppress them: to censor them and destroy their liberty. But he rejected this as antithetical to the entire project of liberty – the basic foundations of freedom and self government. The fundamental challenge once this road is taken is that one person's 'faction' may be another person's speaking truth to power.

[1] A minor personal example: in December 2020 I was forced to switch from Google to the typically inferior rival search engine DuckDuckGo in order to find fact checks of a viral story someone shared with me about election vote totals being changed because the search filters that seemed to be in place on Google to protect me from finding falsehoods about the election were so aggressive that they also kept me from finding a USA Today fact check of those falsehoods.

Most questions in politics deal with issues which are partly based on fact and partly arguable rather than being capable of definitive resolution as facts. Like the government agents who labeled Martin Luther King a dire threat, this approach risks misidentifying threats and undermining liberty. Madison wrote

> "Liberty is to faction what air is to fire, an aliment without which it instantly expires. But it could not be less folly to abolish liberty, which is essential to political life, because it nourishes faction, than it would be to wish the annihilation of air, which is essential to animal life, because it imparts to fire its destructive agency."

Censorship as a strategy for curtailing misinformation carries the risk that actions taken with the very best goals, such as protecting democracy, may in the end undermine the free exchange of ideas and democratic debate.

This paper argues that there is an alternative approach that could help thwart radicalization and better connect the public sphere without need for censorship.[2] The solution should find ways to reconstruct the public sphere in order to curtail the effects of factious misinformation and fake news instead of attempting to eliminate the causes. Most fundamentally, the solution should be to give truth (or at least rebuttals) wings equal to the falsehoods they pursue. In this way, fact can balance falsehood more fully, and citizens can exercise more effective reasoned judgment. The key goal is to connect the public sphere – to get people to argue with and refute each other, and to ensure that those receiving misinformation simultaneously receive the best refutation of it available.

[2] The approach taken here is most relevant to conditions of legitimate political debate and misinformation. Deliberate disinformation campaigns and fake accounts are potentially a different matter.

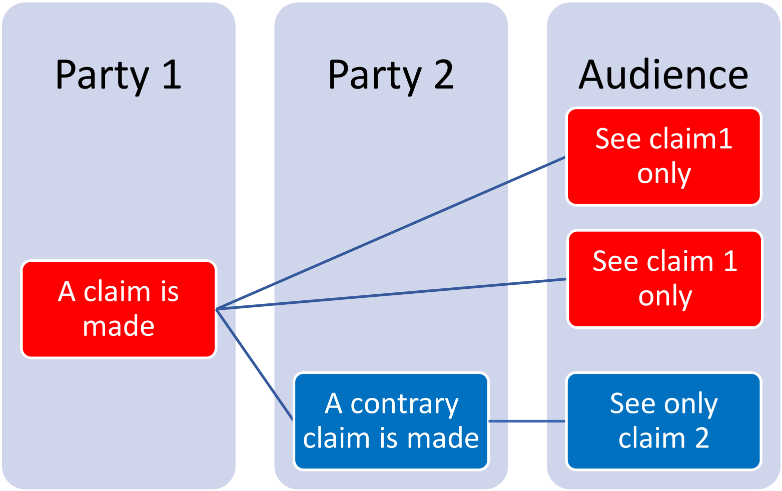
Tagged Rebuttals: A pro-free-speech solution to misinformation

The problem

The essential element of an argument is a controversy. Party 1 asserts a claim about the world that is arguable – that can potentially be contested, rebutted, or fact-checked. Such claims are an inevitable part of politics. Party 2 disagrees with the claim. The media ecology then shapes what happens to the audience – to citizens – when such a controversy takes place.

In the current media environment what tends to happen is that supporters of each side tend to only listen to and experience one side in the controversy, and this problem seems likely to become worse as censorship leads to a fragmentation of social media networks along ideological lines. Ideologically polarized media combine with ideologically polarized social network friend groups, and personally adapted suggestions to make it likely that many members of the audience will be selectively exposed to only one side of the story: they will often see only one viewpoint on controversial claims, and they make consequently make the mistake of accepting such clams, including the associated misinformation, as fact. Figure 1 illustrates this process: a controversial claim spreads, with many members of the audience primarily exposed to only one viewpoint.

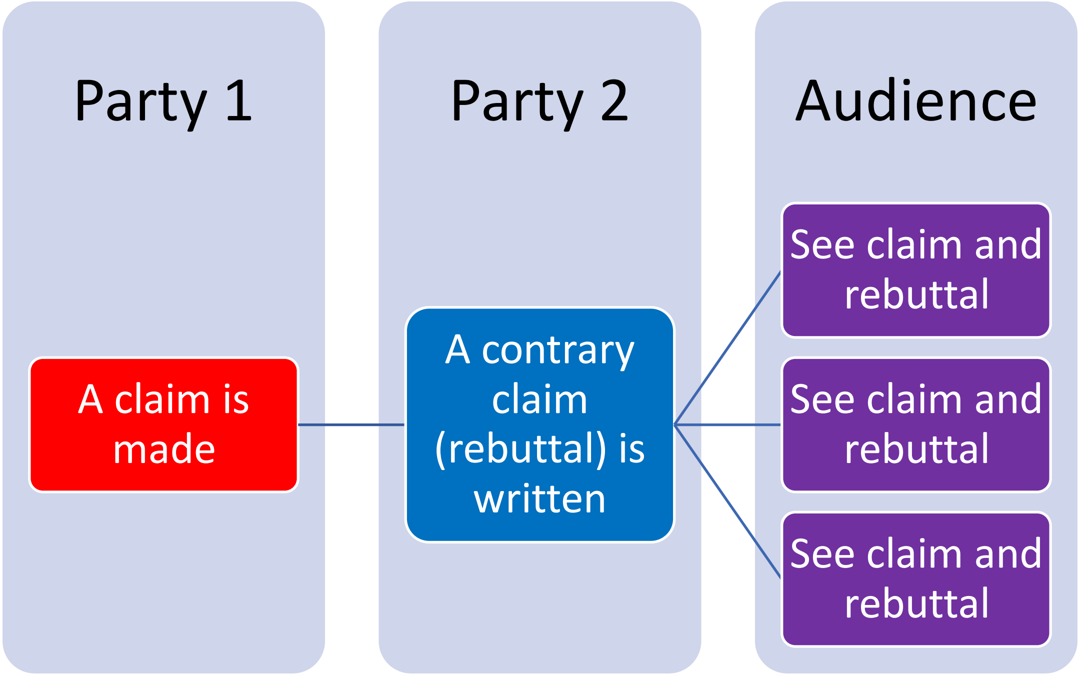
Figure 1: Common Current Pattern of Debate



Under the current pattern of selective attention, the sides of the argument often don't really end up speaking to each other much. The argument instead gets heard only partially by most observers. If my friends, my YouTube suggestions, and my favorite media outlets tend to take one side of the issue, then I will tend to hear that side of the debate echoed back to me much more than any rebuttal of it. I may become increasingly radicalized as a result. As falsehood and distortion flies around the world, the results are polarization and unreality (Barrett et. al. 2021).

In an ideal public sphere, the audience – the rest of us – would observe the whole controversy and come to a conclusion about which argument is stronger, and we could then act on that basis. This preferable pattern is illustrated in Figure 2.

Figure 2: Pattern of Debate with Rebuttals



In figure 2, the audience sees both the claim and the rebuttal (and potentially the rebuttal to the rebuttal…) and then makes up its mind having seen both. Instead of ending up hewing to either the Party 1 or Party 2 line, the audience is more likely to make a reasoned judgment concerning the merits of the relative claims, albeit with the necessary caveat that motivated reasoning will sometimes lead some members of the audience to fail to adjust. At the very least, the audience is forced to acknowledge the existence of the rebuttal, and to make a conscious choice of which argument to accept, knowing that there are alternatives.

A Solution

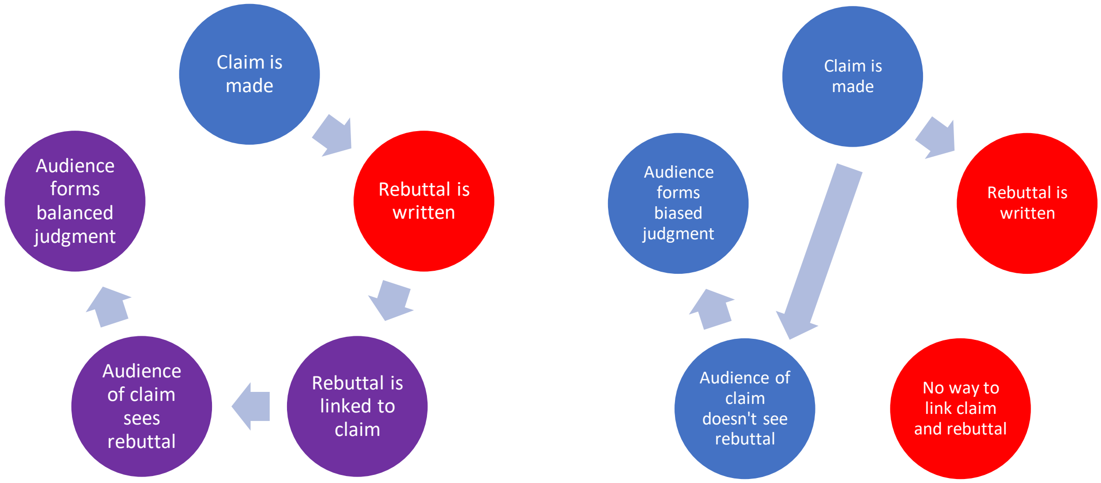
How can we move from the world of Figure 1 to the world of Figure 2? The starting point for a solution is a new option to tag a rebuttal or rejoinder to a news story, webpage, tweet, post or link. This is the 'rejoinder'. Or in the jargon currently being explored by Twitter, a

"Birdwatch note". A rejoinder is a counterargument or rebuttal. Any post or link tagged with a rejoinder will be shared with the rejoinder paired with it. When there are multiple rejoinders or links, a system of up and down voting of rejoinders should adjudicate between alternatives, potentially with an editorial role from fact checkers, social media companies, and others to ensure that the system isn't abused, or manipulated by bots. A key element is to use diverse viewpoints of different readers to identify responses that are credible to a broad audience.

The core contribution is the "tag rebuttal" or "tag rejoinder" or "Birdwatch note" option. In accord with the dictionary definition, a rebuttal is intended to be a refutation or contradiction. It might correct a misstatement of fact. It might also correct a weak argument that draws inappropriate conclusions from a body of facts. It might merely point out the limitations, flawed assumptions, or partial perspectives offered in the original piece. Once one or more rebuttals have been tagged for a specific post or link, all instances in which that link or posting is shared will appear paired with the link to at least one of the rebuttals.

Figure 3 illustrates the different processes of information spread with and without a "tag rejoinder" option.

Figure 3. Process of Idea Spread With and Without Tagged Rebuttal Option


Claim is made
Rebuttal is written
Rebuttal is linked to claim
Audience of claim sees rebuttal
Audience forms balanced judgment
Claim is made
Rebuttal is written
Audience forms biased judgment
Audience of claim doesn't see rebuttal
No way to link claim and rebuttal

The rejoinder idea is very simple, but very powerful. In the introduction I quoted Jonathan Swift's observation that "Falsehood flies, and the Truth comes limping after it." Updated for the 21st Century, we could say that "Falsehood goes viral, and the fact check gets many fewer shares," an observation well supported by sharing patterns online (Vosoughi et. al. 2018). The tag-rejoinder idea overcomes the 'limp' and the fewer shares of the truth/rebuttal by granting it a metaphorical tow rope connected to the falsehood by which means it can glide after it. If the viral story caries with it to every reader a link to the leading rebuttals or rejoinders targeting it, then these rebuttals have a much better chance of getting considered.

In addition to the 'tow rope' of the link to the rejoinder itself, one of the additional strengths of this "tag rejoinder" idea is that if implemented correctly it could provide a profit motive and energy to the process of rebutting arguments, transforming the fact checking process from a burden on social media platforms to a source of revenue for a burgeoning global industry. Through this motive, and the ability to rely upon users to rate rebuttals and proposed rebuttals, this approach would provide the scale so badly needed to effectively manage the process of connecting specious, biased, opinionated, or incomplete arguments with appropriate rejoinders, counterarguments, and rebuttals. Writing good rejoinders could become a profitable business.

Technical Details

There are obviously many important details underlying getting the rejoinder system to work properly. Most importantly, the key would be to have a system for upvoting and downvoting suggested rejoinders along with metrics concerning quality of sources to prioritize rejoinders from sources with better credibility. While the system could be mostly automated, this does not preclude the value of some management from fact-checkers and editors working with social network companies and other internet gate keepers.

To suggest a rebuttal, users can choose the “propose rebuttal” / “tag rejoinder” option in order to propose a particular link or post as a rebuttal to another post or link. AI could also potentially be employed to suggest possible rebuttals based upon similar postings or links that already have a rebuttal.

Once a rebuttal has been proposed, the rebuttal might be evaluated in several ways. (1) Active editing: staff hired by the social media company might review the original posting and the proposed rebuttal to evaluate whether the rebuttal is in fact a credible rebuttal. (2) User-sourced. A subset of the individuals who had shared or posted the item tagged for rebuttal might receive an automatic request to evaluate the proposed rebuttal. Several recent studies demonstrate that the well known wisdom of crowds phenomenon applies to the evaluation of the truth of claims: groups of non-experts can achieve high levels of accuracy (Allen et. al. 2021, Bhuiyan et. al. 2021, Resnick et. al. 2022). (3) Automatically posted if by credible sources. A list of sources considered to be of high quality (e.g. USA today fact checkers) could receive automatic approval of their proposed rebuttals. (4) Rating of user's prior rebuttal proposals. Users who frequently suggest rebuttals that win high ratings by readers and good evaluations from editors might have their rebuttals subject to less scrutiny of the forms outlined above, as compared to users who have a record of suggesting rebuttals that are weak or off topic. (5) Directly posted without evaluation. This might be most appropriate for postings or links with relatively few shares. In any event, once one or more rebuttals have been linked to a post, users will have the option to up or down vote a rebuttal as credible or not credible. This tagging of the credibility of the rebuttal could be used to (1) identify rebuttals that need to be manually evaluated, and (2) to rank rebuttals along with metrics concerning the frequency with which rebuttals are viewed.

How will users who wish to make it their business to write rebuttals find opportunities? Perhaps social media companies could maintain a "rejoinder/rebuttal needed list" -- a list of popular posts and popular shared links which have no rebuttal identified that can be searched by keyword. This list might be supplemented by a user-request feature that would allow those viewing social media posts and shares to tag a post as needing a rebuttal or rejoinder.

Simulations to Test the Proposed Rebuttals Solution

To evaluate the impact of the tagged rebuttals solution to the spread of disinformation on social media, I simulate the spread of disinformation using a Netlogo-based implementation of a virus spread epidemiological model. This model simulates the spread of disinformation on a social network. This is because, network models are best suited to analysis of (dis)information spread (Ji et al., 2017).

The underlying framework for the simulation model is the widely applied Susceptible-Infected-Recovered (SIR) epidemiological model (Kermack & McKendrick, 1927). I follow

Stonedahl and Wilensky (2008) in reinterpreting the ‘R’ term in the model to be ‘resistant’ which captures the possibility that resistance could be acquired through exposure to fact checks or rebuttals as well as through a process of recovery. Furthermore, like SIS models, the model here allows for the possibility that recovery from belief in a falsehood may fail to produce resistance to re-infection, again following Stonedahl and Wilensky (2008).

In this SIR model there are three primary populations of agents: individuals may be susceptible, infected, or resistant. Susceptible agents are potentially receptive to a (false or incomplete) claim. Infected agents believe the claim and are potentially spreading it. Resistant agents have been persuaded by or have independently come up with a rebuttal to the claim. Transition probabilities shape the likelihood that an infected agent will spread the infection to a susceptible agent, the probability of recovery, etc.

SIR models have been used for many years to study the spread of information. Goffman and Newill (Goffman & Newill, 1964) and Daley and Kendall (Daley & Kendall, 1964) were among the first to note that the spread of ideas and rumors could be modeled using the SIR framework. Since then, this approach has been applied across many fields. For example, Robert J. Schiller (2019), applied SIR type models to economic narratives. Several scholars have used these models to examine the ‘viral’ spread of ideas, videos, and memes in social networks (Weng et al., 2012; Weng et al., 2013; Zhao et al., 2013; Bauckhage et al., 2014; Bauckhage, et al., 2015; Beskow et al., 2020).

To explore the consequences of rebuttals, I explore a modified version of the agent-based network-structure SIR model by Stonedahl and Wilensky (2008) in which agents who have developed resistance to misinformation or disinformation that is spreading like an epidemic can resist the spread of that disinformation. Table 1 describes the conditions that are explored.

Table 1: Summary of Model Conditions

Condition	Development of Resistance	Actions available to resistant agents
Baseline 1 Model. Viral spread model. (Stonedahl and Wilensky 2008 model)	When they conduct fact checks (at fixed intervals) agents independently recover with a fixed probability after becoming infected. Agents who have recovered have a fixed probability of becoming resistant.	Resistant agents cannot do anything to resist the spread of disinformation beyond not themselves becoming infected and not themselves spreading.
Baseline 2 Model. Current network spread pattern (Richman et. al. 2022 model)	Agents who are connected in the network to resistant agents are more likely to fact check and to recover and develop resistance.	Resistant agents can prompt fact checks and the development of resistance but only among their network neighbors.
Tagged Rebuttal Model	Once one agent has developed resistance, all agents exposed to infection are simultaneously exposed to fact check and may develop resistance. Key parameter r/i is the ratio of the development of infection versus resistance in exposed agents.	Once at least one agent has developed resistance, in the simulation of this proposal, every effort at infection is paired with a fact check – all agents immediately fact check when exposed by being neighbors with an infected agent, leading some to develop resistance.

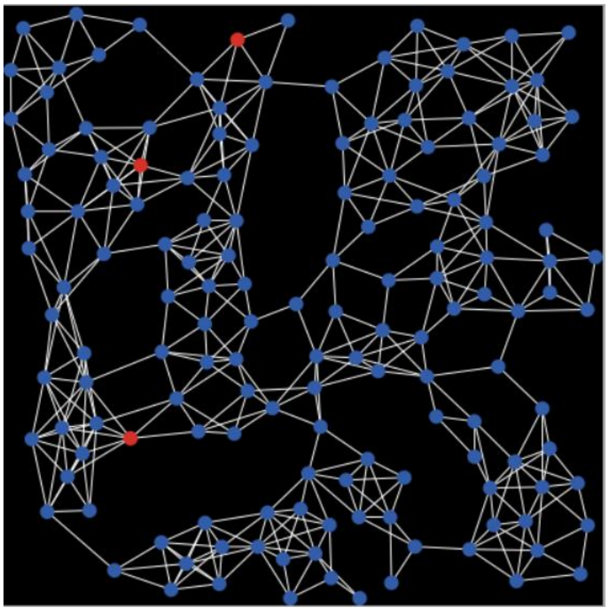
The Baseline 1 Model is the "virus on a network" model by Stonedahl and Wilensky (2008). Agents interact in a network structure, as illustrated in the screen shot below. Each tick, infected agents may trigger an infection in their neighbors with a probability equal to the "disinformation spread chance" slider. Once infected, agents have a probability of recovery and resistance development governed by (1) the frequency of fact checks – the frequency with which the agent investigates and potentially discovers that it has fallen for misinformation or disinformation. When a fact check occurs, there is a probability (given by the recovery chance slider) that the agent recovers. And among agents who have recovered, there is another probability (given by the gain resistance chance slider) that the agent will not only recover but also gain resistance to the disinformation such that they will not fall for it again. In the illustrative runs shown below, the infection chance, recovery chance, and gain resistance chance are all set to their defaults in the Stonedahl and Wilensky (2008) model.

The Baseline 2 Model is an adaptation of the "virus on a network" model of Stonedahl and Wilensky to more accurately reflect the nature of disinformation spread in social networks. This model is drawn from (Richman et. al. 2022). This model adds a resistance fact check chance slider which "Represents the probability that a node which has become immune will "push back" against disinformation by causing neighboring infected nodes to fact check." (Richman et. al. 2022). We set this probability at 25 and 50 percent. All other aspects are unchanged from Stonedahl and Wilensky (2008). This captures the possibility that individuals who have encountered false or misleading information may push back against that information with their network contacts, potentially triggering additional fact checks by those network contacts.

The Tagged Rebuttals Model allows agents who have developed resistance to interfere more generally with the spread of disinformation. Once at least one agent has developed resistance to the disinformation, they can begin interfering with the spread of that disinformation across the entire network through the tagged rebuttal. Specifically, and in line with the rebuttal tags idea developed above, in this model once an agent has developed resistance their rebuttal of the disinformation reduces the ability of the disinformation to continue spreading by causing all susceptible agents who are threatened with infection by an infected neighboring node to potentially develop resistance by engaging in a fact check.

The primary new parameter added to the model is the "Rebuttal-gain-resistance-probability" which is the probability that instead of believing the misinformation, an agent threatened with possible infection will instead become resistant through contact with the rebuttal, a process which can only happen once resistance has developed – this possibility is available only once at least one agent has become resistant. The critical factor for the spread of disinformation that will be varied in the experiments here is the strength of the rebuttal relative to the disinformation: the probability of developing resistance (r) divided by the probability of developing an infection (i) among exposed agents.

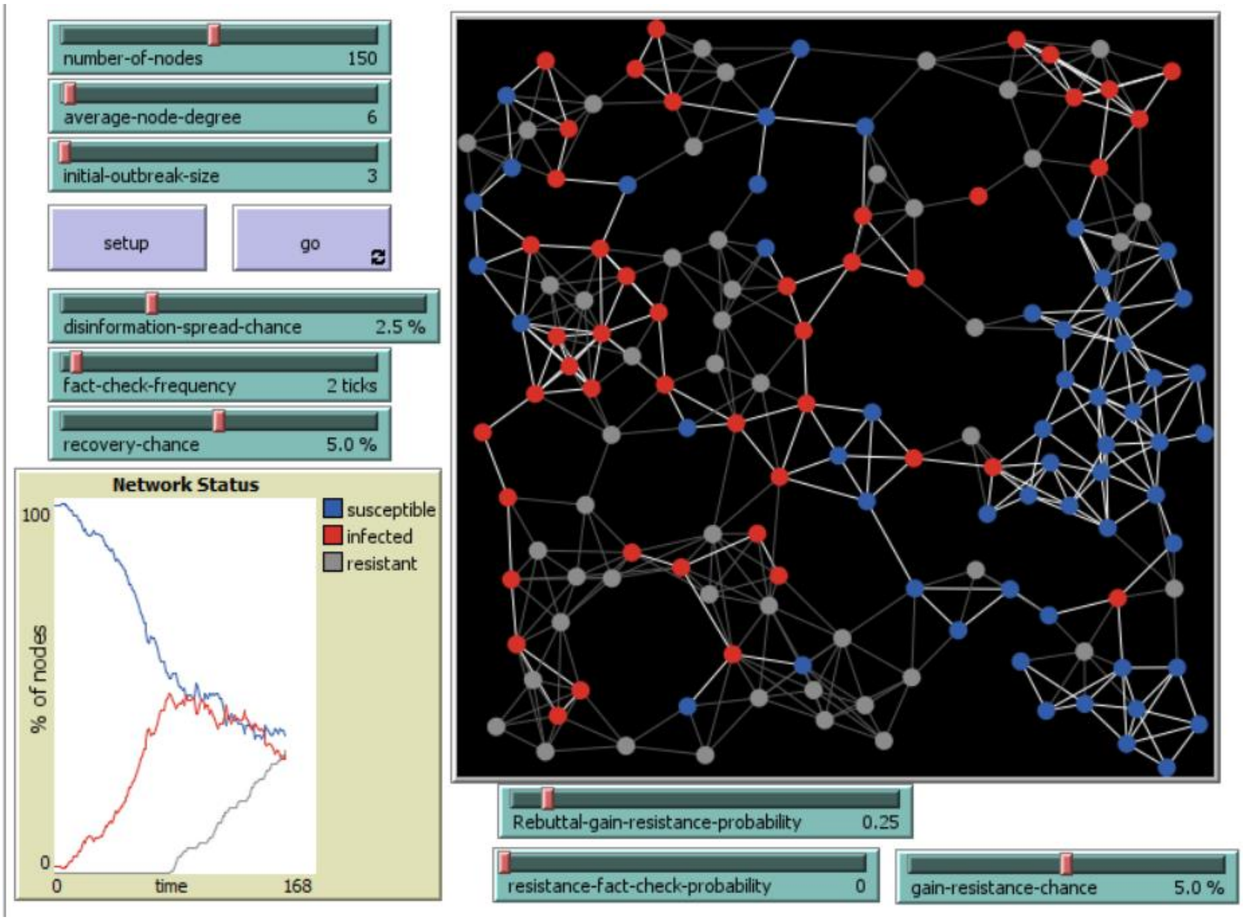
Figure 4: Example of Initial Network Structure



All of the experiment runs start with the same parameters for initial network structure based upon the default values in the Stonedahl and Wilensky (2008) model. An instance of the initial conditions for the model run is shown in Figure 4. Here (as in all runs) three initial infected notes (red) will begin to spread infection to susceptible network neighbors (blue). Once nodes develop resistance, they will change color (to gray) in the visual representation.

Figure 5 shows a screen shot of a model in mid-run. The sliders control model parameters. The network display shows nodes which are infected (red), susceptible (blue) and resistant (gray). The graph in the bottom right shows the history of the number of infected, susceptible, and resistant nodes.

Figure 5: Baseline 2 Plus Tagged Rebuttals Model in Mid-Run



number-of-nodes 150

average-node-degree 6

initial-outbreak-size 3

disinformation-spread-chance 2.5 %

fact-check-frequency 2 ticks

recovery-chance 5.0 %

Rebuttal-gain-resistance-probability 0.25

resistance-fact-check-probability 0

gain-resistance-chance 5.0 %

The ratio of two model parameters: “Rebuttal-gain-resistance-probability” / “disinformation-spread-chance” defines what one might think of as rebuttal persuasiveness. When rebuttal persuasiveness is 1, then agents threatened with infection by having a neighbor who is infected are equally likely to become resistant and become infected. When rebuttal persuasiveness is 0.5, then agents with an infected neighbor are twice as likely to become infected as to become resistant. In the screen shot shown in Figure 5, “rebuttal-gain-resistance-probability” is 0.25, and “disinformation-spread-chance” is 2.5, so Rebuttal Persuasiveness is

0.1, indicating that in any given turn, an agent exposed to an infected neighbor is ten times more likely to believe the disinformation than they are to become resistant to it as a result of exposure to the rebuttal.

Experiment 1: Baseline 1 with Rebuttals

The first experiment involves varying the persuasiveness of the tagged rebuttal compared to the viral misinformation under baseline 1. Throughout the experiment the probability that a susceptible agent with an infected neighbor will themselves become infected in any given turn is fixed at a 2.5 percent probability. By varying the "rebuttal-gain resistance-probability" I vary the relative probability of becoming resistant and becoming infected in order to produce a variety of values of rebuttal persuasiveness ranging from zero (the original baseline 1 model) through 1.4 (the rebuttal is forty percent more persuasive than the misinformation).

Figure 6. Average Steps to Elimination of Misinformation (Experiment 1: 500 runs per condition)

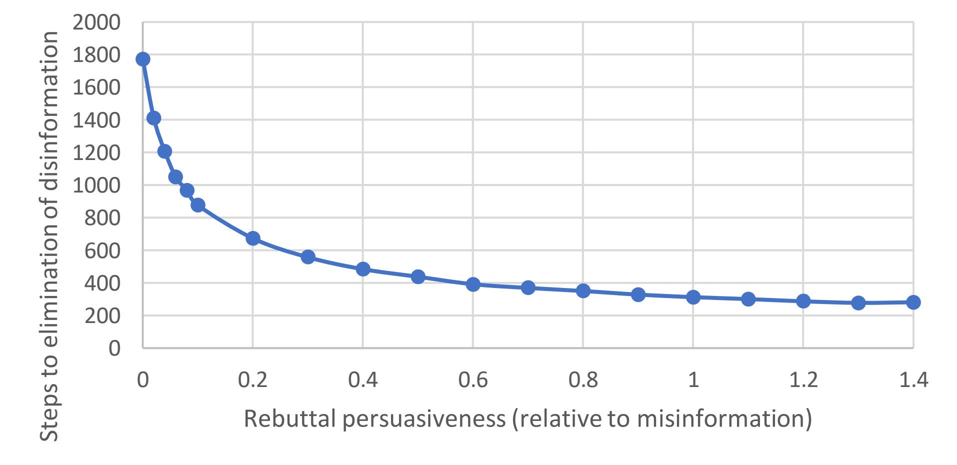


Figure 6 shows the average number of experimental 'tics' or 'steps' required for full elimination of disinformation. The key lesson from Figure 6 is that the most rapid payoff in terms of reduced time to elimination of misinformation spread occurs early on – a substantial portion of the gains relative to the baseline model occur when the misinformation remains substantially more persuasive than the rebuttal. Blocking the spread of disinformation or misinformation doesn't require a surefire rebuttal that is more likely to attract support than the rebuttal – it merely requires that the rebuttal have it's chance to convert those being exposed to the disinformation or misinformation.

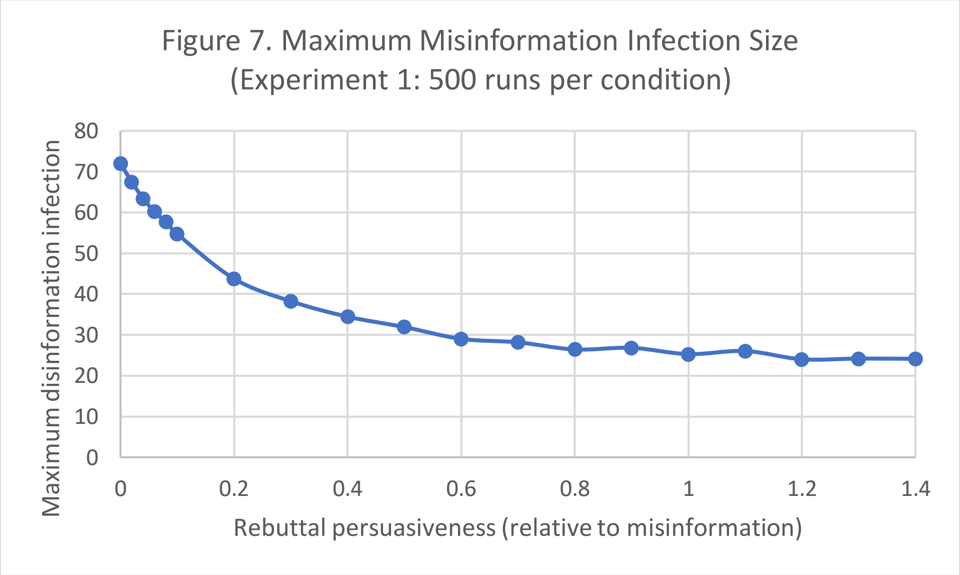
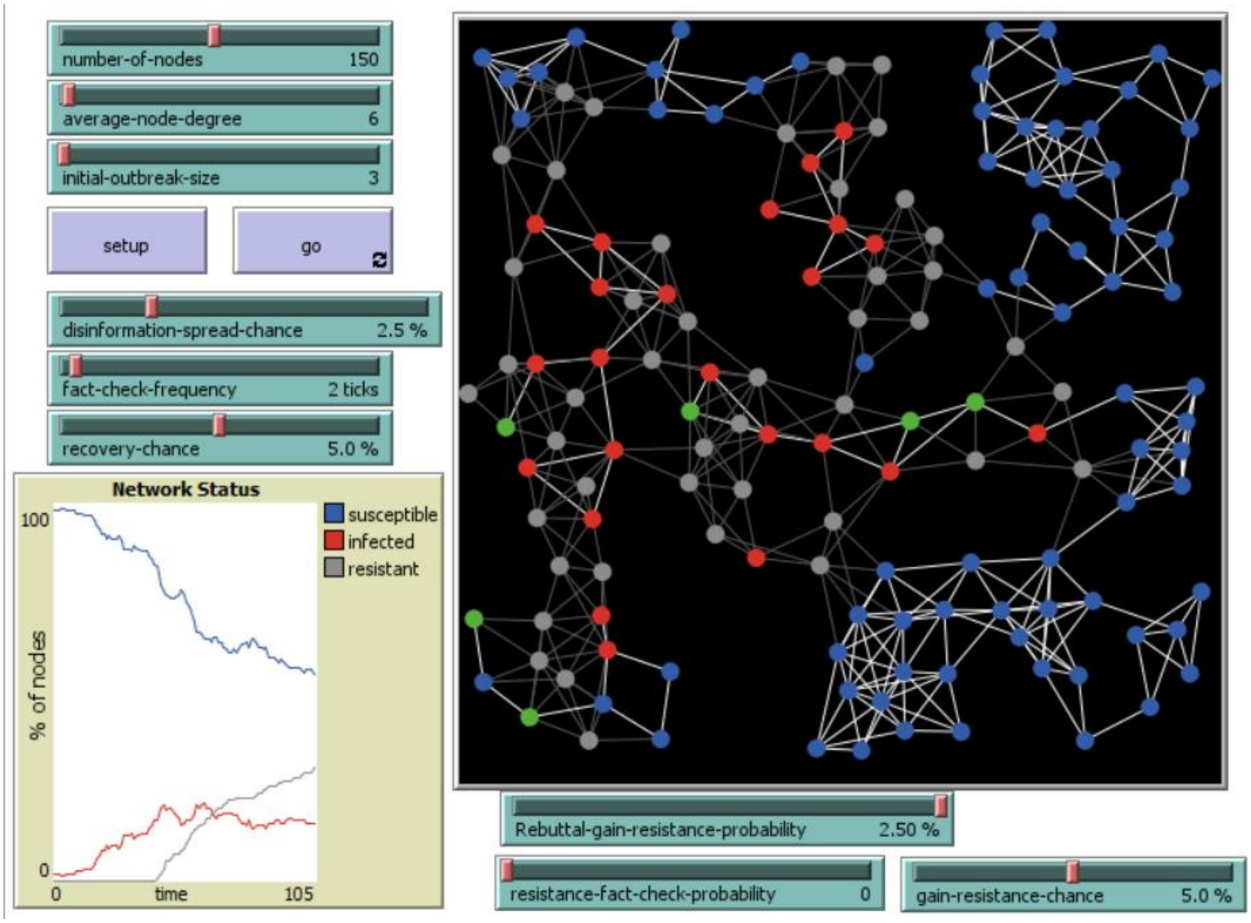


Figure 7 shows the maximum misinformation infection size as a function of rebuttal persuasiveness. As with Figure 6, there is a clear pattern of very rapid reduction that flattens towards the higher values of rebuttal persuasiveness, though the flattening occurs a little bit later. In terms of both maximum spread of the disinformation infection (at a given point in time) and steps to its elimination, the basic lesson seems to be that even a relatively unpersuasive rebuttal (one with less than 50 percent of the effectiveness of the disinformation) can none the less make a tremendous difference in terms of the average scope of the disinformation infection.

Why is this the case? One possibility is that the rebuttal works at the frontier of the spread of the disinformation. As a result it builds 'walls' of resistant individuals who thwart the spread of the disinformation to other parts of the network. Figure 8 shows a screenshot of a fairly typical run with rebuttal persuasiveness of 1. Nodes are colored four ways. Blue nodes are nodes which have never been infected and have never recovered or become resistant. Green nodes have recovered but have failed to become resistant – they may become reinfected. Grey nodes have become resistant. In this run, here paused at just over 100 tics, there are still a substantial number of actively infected nodes, but they have become almost entirely surrounded by grey nodes, almost completely cutting off their ability to spread to the not yet infected regions.

Figure 8: Typical Model Run with Rebuttal Persuasiveness of 1



number-of-nodes 150

average-node-degree 6

initial-outbreak-size 3

disinformation-spread-chance 2.5 %

fact-check-frequency 2 ticks

recovery-chance 5.0 %

Rebuttal-gain-resistance-probability 2.50 %

resistance-fact-check-probability 0

gain-resistance-chance 5.0 %

Table 2: Time to Elimination of Misinformation and Maximum Misinformation Infection by Model Condition

Condition (each was run 1000 times).	Mean tics to elimination of infection (standard error of mean)	Mean maximum percentage infected (standard error of mean)
Baseline 1 Model	1722.6 (12.6)	61.2 (0.4)
Baseline 1 plus Tagged Rebuttal Model	864.7 (5.5)	52.5 (0.3)
Difference	857.9*	8.7*
Baseline 2 Model	1578.9 (11.8)	59.9 (0.4)
Baseline 2 plus Tagged Rebuttal Model	478.2 (3.7)	38.3 (0.3)
Difference	1100.7*	21.6*

* $p < 0.01$ based on z-statistic for difference.

Table 2 examines the outcomes from a set of runs with rebuttal persuasiveness of just 0.1. It is worth noting just how weak the simulated tagged rebuttal is relative to the disinformation infection in the Table 2 runs (in which it none the less has a substantial impact.) In the settings used, once a rebuttal exists, when an agent is threatened with infection it has a 2.5 percent chance of becoming infected, but only a 0.25 percent chance of developing resistance if it does not get infected. That is, these are runs in which the tagged rebuttal is substantially less persuasive than the initial disinformation it is paired with. When the tagged rebuttal is more effective, obviously it has an even greater impact. Even under these weak conditions, however, the tagged rebuttal brings about a very dramatic decline in the spread and duration of the disinformation.

With the ability of the truth to begin hitching a ride on the disinformation, the duration of the disinformation infection (and extent to which the disinformation controls a significant portion of the population) drops dramatically. Versus the baseline 2 condition, for instance, time to elimination of the disinformation is cut by more than 2/3, and the maximum percentage of agents infected drops by more than 1/3.

These simulation runs demonstrate the potential potency of a rebuttal-based system for containing the spread of disinformation in a network. Once at least one agent has developed resistance, the growth of the disinformation is curtailed because it becomes difficult for the disinformation to spread without being accompanied by a rebuttal to it. The rebuttal triggers fact checking on the part of those threatened with the spread of disinformation, and thereby curtails the spread of disinformation.

Concerns and Limitations

Widespread application of a rebuttal tagging tool would potentially give partisans of particular (often limited) viewpoints a socially productive activity to pursue – making sure that those they disagree with get appropriately rebutted across media spaces. But such partisans might also attempt to use the rebuttal more as a form of propaganda and intellectual graffiti. Consider this scenario. Proponents of disinformation or misinformation might post their rebuttals to many legitimate and accurate sources as a way of spreading the disinformation and misinformation ideas. Thereby, potentially, the disinformation could hitch a ride on the truth instead of the other way round.

It is possible that this concern is overrated. After all the evidence seems to indicate that falsehood is more likely to go viral than truth (Vosoughi et. al. 2018). Hence, while falsehood might hitch a ride on truth, this is potentially a less potent ride.

Several aspects of the design of a rebuttal system could also limit this problem. First, there are several ways in which rebuttals could be moderated, as discussed above, including either through active moderation, through crowd-sourced user ratings, or through a process of asking those who shared the original post to rate the credibility of the proposed rebuttal to it. Such filters would make it more difficult for low quality rebuttals to win approval. Second, if approved, such postings might well win low ratings, leading them to be eventually demoted.

Ultimately, however, how one views this problem depends upon whether one has faith in the capacity of the average reader to exercise critical judgment. In his work “On Liberty” John Stewart Mill famously put the case for hearing all arguments from all sides.

> “He who knows only his own side of the case knows little of that. His reasons may be good, and no one may have been able to refute them. But if he is equally unable to refute the reasons on the opposite side, if he does not so much as know what they are, he has no ground for preferring either opinion... Nor is it enough that he should hear the opinions of adversaries from his own teachers, presented as they state them, and accompanied by what they offer as refutations. He must be able to hear them from persons who actually believe them...he must know them in their most plausible and persuasive form.”

The process of tagging rebuttals will improve the likelihood that everyone has an opportunity to have exactly this experience.

One needn’t be particularly optimistic about human nature to hope that tagging rebuttals might help the current situation. Obviously, confirmation bias will have its innings no matter what. However, the rebuttals idea aims to make confirmation bias a bit less powerful. The foundation of this idea is the notion that if truth can closely follow falsehood, thoughtful readers will have a better chance, even if only slightly better, to able to distinguish between arguments that are more sound and arguments that are less sound. An additional key advantage is that tagged rebuttals can accompany exposure to false views, thereby undermining their opportunity to gain adherence. The goal of the rebuttal system is to put truth on a more equal footing with viral falsehood. The falsehood is already viral. Hopefully by giving truth a tow-rope, the falsehood will be less damaging. All that is necessary to reduce the damage is for the rebuttals idea to improve, even slightly, the ratio between the spread of truth and falsehood as compared to the present.

Another concern is that users might abuse the system by posting as “rebuttals” arguments which are in fact mostly supporting the original post or link. One possibility would be to create an additional category for “support” and/or allow users to flag rebuttals that actually belong in a list of supporting links.

Conclusion – three scenarios

The basic conclusion is that having a rebuttal option whereby counterarguments could consistently follow arguments around cyberspace could be a substantial benefit when it comes to reducing the tendency of people to believe things that are not true. To wrap this up in a somewhat more fanciful way, let’s imagine three scenarios: call these social-networks-before-censorship, social-networks-after-censorship, and social-networks-with-rebuttals.

Social-networks-before-censorship: Jones writes a post making a series of questionable claims. Smith writes an article criticizing the logical fallacies and lack of evidence in Jones’ argument. Jones and friends share links to Jones, and mostly ignore Smith. Smith and friends mostly share links to Smith’s article. Everyone ends up sure they are right. And they start to hate each other.

Social-networks-after-censorship: Jones writes a post making a series of questionable claims. Smith writes a post criticizing the logical fallacies and lack of evidence in Jones' argument. Jones and friends share links to Jones until they are de-platformed. Smith and friends mostly share links to Smith's article. Everyone ends up sure they are right. And they start to hate each other. Jones and friends start to hate the Social-Network too and start looking for new platforms to share their questionable claims.

Social-networks-with-rebuttals: Jones writes a post making a series of questionable claims. Smith writes a post criticizing the logical fallacies and lack of evidence in Jones' argument. Smith tags Jones' article with Smith's rejoinder. Smith and Jones' friends all see both articles because every time Jones' article gets shared, the network automatically attaches a link to the leading rejoinder (by Smith). Jones writes a rebuttal to Smith's rejoinder and tags it as a rebuttal to Smith's article. Smith writes a rejoinder to Jones' rebuttal. Their friends still see everything. Some people conclude that Jones is mostly right (in fairness though many recognize that Smith has a few good points.) Some people conclude that Smith is mostly right. Others conclude that since neither side had very good arguments they probably had better wait for more information or better arguments. Smith and Jones eventually become rich from advertising revenue generated by their controversy but are also eventually convinced that the other did have a few good points. They become friends and go have coffee or a beer together.

Which world would you rather live in?

Works Cited

Allen, J., A. A. Arechar, G. Pennycook, D. G. Rand, Scaling up fact-checking using the wisdom of crowds. Sci. Adv. 7, eabf4393 (2021).

Barrett, Paul M., Justin Hendrix, and J. Grant Sims. 2021. "Fueling the Fire: How Social Media Intensifies U.S. Political Polarization – And What Can Be Done About It" NYU Stern Center for Business and Human Rights. https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/613a4d4cc86b9d3810eb35aa/1631210832122/NYU+CBHR+Fueling+The+Fire_FINAL+ONLINE+REVISED+Sep7.pdf

Bauckhage, C., Kersting, K., & Rastegarpanah, B. (2014). Collective Attention to Social Media Evolves According to Diffusion Models. WWW'14 Companion, April 7–11, 2014, Seoul, Korea. ACM ACM 978-1-4503-2745-9/14/04. http://dx.doi.org/10.1145/2567948.2577298

Bauckhage, C., Hadiji, F., & Kersting, K. (2015). How Viral Are Viral Videos? https://www.researchgate.net/profile/Christian_Bauckhage/publication/276282493_How_Viral_Are_Viral_Videos/links/55605eb708ae86c06b641ad5/How-Viral-Are-Viral-Videos.pdf

Beskow, D. M, Kumar, S., & Carley, K. M. (2020). The Evolution of Political Memes: Detect-ing and Characterizing Internet Memes with Multi-Modal Deep Learning. Information Pro-cessing and Management, 57(2). https://doi.org/10.1016/j.ipm.2019.102170

Bhuiyan, Md Momen. Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, CSCW2, Article 93 (October 2020). ACM, New York, NY. 26 pages. https://doi.org/10.1145/3415164

Brady, William J., Julian A Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. "Emotion shapes the diffusion of moralized content in social networks." *PNAS: Psychological and Cognitive Sciences*. 114| (28) 7313–7318. https://www.pnas.org/doi/full/10.1073/pnas.1618923114

Daley, D. J. & Kendall, D. G. (1964). Epidemics and Rumours. Nature, 204, 1118–1118.

Goffman, W. & Newell, V. A. (1964). Generalization of Epidemic Theory: An Application to the Transmission of Ideas. Nature, 204, 225-8.

Haidt, Jonathan. 2022. "Why the past 10 years of American life have been uniquely stupid: It's not just a phase." *The Atlantic.* https://www.theatlantic.com/magazine/archive/2022/05/social-media-democracy-trust-babel/629369/

Jerit, Jennifer, and Yangzi Zhao. 2020. "Political Misinformation" Annual Review of Political Science 2020 23:1, 77-94

Ji, S., Lü, L., Yeung, C. H., & Hu, Y. (2017). Effective Spreading from Multiple Leaders Identi-fied by Percolation in the Susceptible-Infected-Recovered (SIR) Model. New Journal of Physics, 19, 073020.

Kermack, W. O., & McKendrick, A. G. (1927). A Contribution to the Mathematical Theory of Epidemics. Proceediings of the Royal Society A: Mathematical, Physical, and Engineering Sciences, 115, 700-721. https://doi.org/10.1098/rspa.1927.0118

Kuklinski JH, Quirk PJ, Jerit J, Schwieder D, Rich RF. 2000. Misinformation and the currency of democratic citizenship. *Journal of Politics* 62:790–816

Lorenz-Spreen, Philipp, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. 2022. "Digital Media and Democracy: A Systematic Review of Causal and Correlational Studies." SocArXIV Papers Doi: 10.31235/osf.io/p3z9v

Nyhan B, Reifler J. 2010. When corrections fail: the persistence of political misperceptions. *Political Behavior* 32:303–30

Quote Investigator. 2014. A Lie Can Travel Halfway Around the World While the Truth Is Putting On Its Shoes. Downloaded April 6, 2022 from https://quoteinvestigator.com/2014/07/13/truth/

Rathje, Steve, Jay J. Van Bavel, and Sander van der Linden. 2021. "Out group animosity drives engagement on social media." *PNAS: Psychological and Cognitive Sciences*. 118(26) e2024292118 . https://doi.org/10.1073/pnas.2024292118

Resnick, Paul, Aljohara Alfayez, Jane Im, Eric Gilbert. 2022. Informed Crowds Can Effectively Identify Misinformation. https://doi.org/10.48550/arXiv.2108.07898

Richman, Jesse, Lora Pitman, and Girish S. Nandakumar. 2022. A Gamefied Synthetic Environment for Evaluation of Counter-Disinformation Solutions. *Journal of Simulation Engineering*, Volume 3 (2022/2023). Article URL: https://articles.jsime.org/3/1

Richman, Jesse and David Richman. 2021. "Media Fact Checks, Polarization, and Trust in Contemporary US Politics" Southern Political Science Association Annual Conference. January 2021.

Schiller, R. J. (2019). Narrative Economics: How Stories Go Viral and Drive Major Economic Events. Princeton University Press. https://doi.org/10.2307/j.ctvdf0jm5

Weeks BE. 2018. Media and political misperceptions. In Misinformation and Mass Audiences, ed. BG Southwell, EA Thorson, L Sheble, pp. 140–56. Austin: Univ. Texas Press

Wood, T., Porter, E. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. Polit Behav 41, 135–163 (2019). https://doi.org/10.1007/s11109-018-9443-y

Stonedahl, F. and Wilensky, U. (2008). NetLogo Virus on a Network model. http://ccl.northwestern.edu/netlogo/models/VirusonaNetwork. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among Memes in a World with Limited Attention. *Scientific Reports*, *2*(1), 335.

Weng, L., Menczer, F., & Ahn, Y. (2013). Virality Prediction and Community Structure in Social Networks. *Scientific Reports*, *3*(1), 2522.

Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*. 359:6380, pp. 1146-1151. DOI: 10.1126/science.aap9559

Zhao, L., Cui, H., Qiu, X., Wang, X., & Wang, J. (2013). SIR Rumor Spreading Model in the New Media Age. *Physica A: Statistical Mechanics and its Applications*, *392*(4), 995–1003.