

Samantha Richter

THE INFORMATION VACCINE

Using Inoculation Techniques to Combat Digital Misinformation



Photo by [Jakob Owens](#) (Unsplash)

Abstract

The spread of online misinformation is a growing and pervasive problem, with particular potency in the realm of politics. Despite the many attempts to address the issue by technologists and policymakers alike, success has been limited. The debunking strategies most often used to combat false information, in which the user is presented with evidence that the information they are seeing is inaccurate, do not effectively curb the exponential

sharing of misinformation on the internet. This presents a significant threat to the American democratic process, as the public becomes more vulnerable to misleading information and more immune to interventions.

New psychological research indicates that a different technique, prebunking, which ‘inoculates’ the public against misinformation, may be a more promising and effective strategy. The inoculation process works much like a vaccination, exposing people to small doses of the thought processes, strategies, or examples of disinformation that threat actors disseminate in order to help the public identify and reject misinformation when they encounter it in the future. A prebunking campaign can be implemented in several different ways, including using graphics and images or posting educational statements about disinformation on users’ social media feeds, but very few companies or organizations have utilized this evidence-based strategy. Some challenges remain regarding how to sustain the effectiveness of prebunking campaigns over long periods of time. However, there are very encouraging and largely unexplored applications of this approach that hold great potential to prevent and even reverse the harmful effects of digital misinformation.

Background

Political, financial, and social incentives have given rise to an increasingly alarming spread of digital disinformation, which is deployed with greater speed and agility than ever before. Disinformation is the targeted and purposeful attack on an objective truth by a “threat actor” with the goal of misleading, rallying, sowing discord, or building a following. Disinformation and its perpetrators predate digital media, but its emergence and growth on digital social media platforms has made it harder to detect than ever before, posing an even greater threat to democracy and political stability. On the internet, false news spreads [six times faster](#) than accurate information, employing sensationalism to reach the largest audience possible.

We have witnessed the impacts in real time. During the course of the 2016 and 2020

presidential elections, fear over “fake news” and foreign interference in elections took the

2

spotlight, relegating other substantive issues to the sidelines. During the Covid-19 pandemic, the dangerous and harmful effects of disinformation reached new heights, with many families choosing to remain unvaccinated based on ill-founded claims of a link between vaccinations and autism or infertility, leaving the United States in a vulnerable position with a significant portion of the population remaining unprotected.

Successful disinformation campaigns often utilize social frustrations over [emotional](#) issues like religion, race, gender, and the economy. Drawing from sensitivities about topics like the elevation of certain social groups over others causes people to ignore indications that hateful information may be untrue. When an individual’s emotions are manipulated by false information, it is nearly impossible to change their opinions through debunking, meaning they will never be completely disabused of their belief. Because social media sites like Facebook, Instagram, and YouTube are designed to rely on clickbait, rabbit holes, recommended user algorithms, and [echo chambers](#) to generate clicks, the platforms tend to keep users surrounded by those with identical beliefs, rather than expose them to new perspectives. In some disinformation campaigns, these social media echo chambers are exploited through political strategies like [strategic extremism](#)—using disinformation to create a cult-like following while isolating those users to increase loyalty—which heightens support for radical groups while reducing trust in all news sources and institutions, and degrades perceptions of governmental stability.

Current attempts to combat the spread of disinformation include using [algorithms](#) to track and delete misinformation, investing in media literacy education, and passing laws that set restrictions on spreading fake news. However, the rapidly changing nature of the internet has rendered all of these strategies inadequate. Disinformation strategies evolve faster than attempts at educating the public about it, and detection mechanisms often border on censorship or are simply outdated. The U.S. [Department of Homeland Security \(DHS\)](#) recommends the establishment of a third-party trusted database created by political

organizations across the spectrum, research institutions, and members of the private sector to share information about websites' accuracy. However, the utility of such a database is questionable because regardless of the quality of the contributions, the

¹ Disinformation is defined as false information that is spread with the intention of misleading while misinformation is defined as false information without regard to its origins.

3

members of the public most vulnerable to disinformation are unlikely to utilize this information successfully or will doubt its authenticity. DHS has also proposed media literacy programs, framed as patriotic initiatives, to attract groups that otherwise would not seek out training or inoculation. Finally, the Department supports legislative changes, including amending section 230 of the Communications Decency Act to make platforms civilly liable for content that their users post (a proposition that is unlikely to pass, due to pushback from social media sites and difficulty of enforcement). It also advocates for the passage of the Honest Ads Act, which would change the definition of "electioneering communication" to include internet based political ads, making social media sites accountable for the same disclosure requirements as TV, radio, and print.

While these interventions face significant challenges, there are other strategies that show signs of promise. Similarly to how a vaccine works, experts have found that [prebunking](#) disinformation campaigns is far more effective than attempts to erase and counteract false messages. The predominant approach to dealing with this issue in the past has been to debunk, or disprove, untrue theories or statements on social media following circulation. ² But this approach has proven to be largely ineffective and performative. In contrast, [inoculation theory](#) suggests that preemptive exposure to smaller-dose versions of persuasive disinformation campaigns can trigger a cognitive-motivational process (similar to how antibodies are formed following vaccinations) that makes someone less susceptible to persuasion by threat actors in the future. Prebunking utilizes this inoculation theory to build psychological immunity against online misinformation.

Inoculation works by exposing people to the thought process, strategies, and examples of threat actors in order to make them question misinformation when they come across it in

the future. Games like [“GoViral,”](#) created to share the motives of Covid-19 misinformation spreaders, and [“Bad News,”](#) designed to test people’s awareness of clickbait, both showed promising results of increasing people’s skepticism towards disinformation tactics.

However, the effectiveness of the inoculation [faded](#) after one to two months unless refreshed. Researchers have proven the [success](#) of using positive and motivational messaging in prebunking while also sharing the exploitative nature of disinformation campaigns. For example, when a user correctly identifies disinformation in a prebunking

² Debunking is defined as exposing the falseness or lack of evidence behind an idea, belief, or claim.

4

exercise, messages of positive reinforcement will incentivise them to continue searching for false information. Therefore, the challenge has become elongating the results of inoculation and reaching people that do not view themselves as vulnerable.

In order to protect against the threats of disinformation and combat existing disinformation campaigns, there needs to be cooperation between social media companies, policy makers, and academics to advance the approach with the strongest evidence base: inoculation.

Psychological Research on Inoculation Techniques

There is robust psychological literature that provides support for inoculation techniques, as well as examples of specific strategies of inoculation that have proven successful in trials or research studies. Additionally, studies have explored the failures of debunking misinformation after it is spread and the potential for success using inoculation strategies.

Evidence for Inoculation

The effectiveness of explaining “why”: When misinformation is consumed by a population that benefits from the false messaging, that population tends to spread and act on it even after it has been corrected. The emotions triggered by threat actors and people’s resistance towards feeling “tricked” make post hoc debunking strategies ineffective. Inoculation can overcome this phenomenon because when people understand the underlying motives, the

“why”, of threat actors, they tend to take a more objective perspective on misinformation instead of feeling embarrassed or belittled by fact checks that prove them wrong. [Studies](#) have found that prebunking attempts are more successful when the public is presented with examples of the reasoning and bias behind disinformation posts instead of just examples of disinformation posts without context of why that false post may have been created by a threat actor. Social media sites enforcing transparency of who created a post/article/etc., when it was created, and its affiliation with other sites or organizations will encourage people to ask why it was created - therefore allowing the effects of prebunking to materialize.

Criteria for Acceptance: [Five Criteria](#) that help people determine the validity of a post or message are: 1) its compatibility with other relevant information, 2) the reputability of the source, 3) other people’s response to it, 4) internal consistency, and 5) details of supporting evidence. A successful prebunking or inoculation campaign effectively touches on all of these topics in both a specific way that is relevant to a current issue (e.g., how a QAnon theory gained traction) and also a more generalized way (e.g., how the tactics QAnon used can be observed among all threat actors). Using real life examples combined with psychological reasoning to explain why disinformation campaigns are successful increases the likelihood of someone questioning misinformation in the future. Frequent exposure to multiple disinformation scenarios in which the five criteria are laid out and explained will build people’s resilience towards believing falsehoods in the future.

Strategies of Inoculation

Empowerment: People responded most strongly to inoculation and intervention when the [campaign objective](#) of a threat actor was revealed, and when it became evident to the targets that the threat actor’s agenda was the primary focus of the misinformation. This inoculation was further enhanced by exposing their methods of building controlling infrastructure and seeding misleading content. When the intervention was focused on how

threat actors reached vulnerable populations, the effects of their goals and how they used unwitting agents, people were less receptive. That is, there was a more welcoming response to inoculation when the deceitful motives of the threat actor were explained rather than attributing it to the gullibility of the public. As one might expect, putting the public in a position of power and intellectual superiority over threat actors typically made them more interested in and willing to discount disinformation.

Visual power: Social media is now largely composed of images, memes, and short videos (such as those shared on TikTok). Unsurprisingly, these visual agents tend to be the most contagious way to spread misinformation. Since images and other visuals can cross language, age, and cultural barriers in a way that text often cannot, they have become the most pervasive drivers of disinformation campaigns. It is easier to use [subversive messaging](#) through visuals as they often awaken a subconscious emotive response from the viewer, and technology like “deep fakes” and clipped videos continue to complicate this

area even further. Prebunking attempts in the future must include preparation for disinformation spread through GIFs, memes, videos, and images. There is a gap in the current body of research on disinformation, as it does not address this growing issue and the need to educate people on visual literacy. Just as visual disinformation tactics are effective for threat actors, a similar approach could be taken by prebunking campaigns to engage the public and inspire [regulatory obligations](#) for social media companies.

Techniques rather than topics: Inoculation attempts that focus briefly on real life examples but spend more time on reasons *why* disinformation spreads and *how* to recognize the warning signs of a false post are more successful. People retain a healthy skepticism and reduced susceptibility to disinformation campaigns much more effectively when they are taught about the techniques of threat actors rather than when the focus is on how disinformation appears in specific instances.

For further detail about the studies that support these inoculation strategies and theories see

Current Applications

Disinformation prevention initiatives currently direct most of their attention to “behind the scenes” ways of reaching the public. This includes creating digital credibility scoring tools that classify other websites as biased or accurate, bot/spam detection systems that can identify automated social media accounts, and whitelists that create lists of trusted web addresses or accounts. Although these efforts are helpful, they often function more as debunking techniques and only serve those who seek them out (i.e., likely those least prone to the adverse effects of disinformation attempts). There are also countless courses and games that educate about disinformation, but these also only help those already aware that the problem exists and interested in addressing the problem.

Almost all startups and companies that are combating disinformation do it in a “fact-checking” or debunking way, and almost none of them disclose the extent to which their techniques are based on psychological research or evidence. There is a significant gap between the research evidence on the efficacy of prebunking and the application of this

7

strategy within the digital world. Most companies that are currently combating misinformation or disinformation on social media are not transparent about the psychological support for their strategies and the use of prebunking strategies is rare or nonexistent.

For further research on the failure of current debunking strategies and suggestions into how debunking could support inoculation initiatives see Appendix B.

Analysis

Inoculation has already proven successful especially in regards to disinformation in [politics](#), so the main goal for those studying and applying the technique ought to shift to ensuring

long-term effectiveness and reaching a larger audience.

Current challenges remaining in the prebunking mission to create lasting success:

1. ***Western focused:*** Almost all of the recent research done on this topic has been based in western nations and cultures. There may be unexplored discrepancies in the success of prebunking due to different cultural beliefs and norms, as well as different types of media in other parts of the world. In order for prebunking start ups and other initiatives to be successful, there must be research into how other types of government, religion, culture, and social dynamics might influence the efficacy of current inoculation strategies.
2. ***Tackling the divide between social media companies and disinformation researchers:*** Social media companies do not currently have adequate incentives to collaborate with researchers on this topic. Social media platforms tend to benefit from misinformation, clickbait, and rumors, so they have often been [unwilling](#) to share data with researchers. In addition, due to ethical and legal concerns over this type of data, it remains uncommon for social media companies to release information to researchers about algorithmic models and success rates of taking down posts with false information.

The most logical way to implement prebunking techniques will be through social media platforms, underscoring the urgent need to incentivize platforms like Facebook and Twitter to cooperate. The solution to this issue may lie in transparency with users. [Mechanisms](#) could be put in place to allow social media users to consent to research studies that analyze their everyday use of social media platforms. It could be beneficial in the long term to allow the public to participate in these types of studies and then have access to the findings as it would increase awareness about the issue of disinformation while giving researchers a much more accurate and larger data pool than what they're currently working with.

For examples of companies that are using inoculation strategies successfully see Appendix C.

3. **Creating lasting success:** The most important question that remains is how to ensure that the effects of inoculation create long-term immunity to disinformation.

[Research](#) has suggested that monthly inoculation training is successful in maintaining results (observed through testing of the “Bad News” game), but it would be difficult to engage the public in [calculated repeated inoculation](#).

4. **Implied truth effect:** [Labeling](#) information as “false,” as is often the case in debunking strategies, can have the unintended effect of increasing people’s confidence in all unlabeled material, decreasing the likelihood that they will think analytically about its accuracy. When using inoculation, it’s important to avoid this same mistake.

Unexplored ventures and possibilities for innovation:

1. **How current debunking strategies can be used to support prebunking:** There is little to no research on how current attempts to curb disinformation can work in tandem with the growing inoculation movement. Future research should explore how the two strategies can work in tandem; for example, whether inoculation can be supported by debunking tactics or whether debunking strategies can help extend the short-term success rates of inoculation. Although debunking has proven to be largely ineffective on its own, it may support inoculation attempts if it was used as an additional mitigation strategy to combat disinformation. For example, if there

was a prebunking campaign to inoculate the public against anti vaccination hoax stories it may also be helpful to have a “warning: false Covid-19 information” tag on false social media posts. This type of debunking warning currently exists on many social media sites including Instagram and Twitter, but its effectiveness may increase if it’s paired with prebunking efforts because it will serve as a reminder of what the public has already been exposed to instead of a first attempt to disprove

misinformation.

2. **A “think before you share” feature:** To overcome the challenge of how to provide frequent and consistent inoculation, social media sites should be encouraged to make a feature that encourages users to reevaluate what they’re sharing. For example, when someone presses the share button on a flagged instagram post with political or controversial hashtags, there can be a three to five second graphic that says “do you trust the source that created this post?,” “Make sure you have verified the information that you are about to share with others,” or “Please think twice about sharing this - reposting can lead to more harm than good.”

3. **Herd immunity - the societal effects of inoculation:** Most studies conducted on this topic focus on the effects of inoculation on the individual but don’t observe how this new form of skepticism may affect the larger public. Similar to a vaccine, if a larger group of people have been inoculated against disinformation it will be much harder to spread, resulting in a weakening of the influence of threat actors. It is now important to research how much of the population needs to be inoculated in order to see a significant decrease in the spread of disinformation. Prebunking initiatives should also explore how to demonstrate the benefits of inoculation to the public.

4. **Inoculation with visuals:** Using images, memes, videos, and other graphics to combat disinformation campaigns has been proven successful by studies but is barely used in real life attempts to inoculate.

5. **Reaching older and younger audiences:** Success rates of inoculation have not yet been researched or tied to age in any way. An initiative to teach prebunking strategies to kids at a younger age through school programs or games may be

successful. There is also a lack of inoculation resources for the elder population, who may be eager to learn through media literacy education programs.

The Promise of Prebunking to Combat Disinformation

The threat to democracy wrought by the digital transformation of politics paired with the pervasive spread of false information makes finding new solutions to online misinformation a matter of growing urgency. The stakes are particularly high to find a new way to counter both the purposeful and accidental spreading of disinformation, which is why it is crucial to turn to the psychologically backed theory of inoculation.

Recent research has overwhelmingly shown that using a prebunking approach to protect the public against the harms of online false information can be achieved through social media messaging, instructional videos or lessons, games, and more. All of these avenues show promising results; however, the research on this topic has been underutilized by social media companies who hold responsibility for combatting much of online misinformation. Social media companies prefer debunking tactics because it is easier to eliminate existing misinformation on their platform than to help users become discerning through inoculation tactics. Since social media companies lack the incentive to invest in new prebunking strategies and because inoculation research struggles to keep up with the new ways in which disinformation spreads, combatting this issue has proven a daunting task.

To overcome these challenges and begin ‘incolutating’ the public against misinformation on a larger scale, it’s important to show social media companies, government agencies, and the education system that prebunking now will save energy and time that is wasted with current, less effective debunking strategies. Just as a vaccine temporarily and harmlessly flares up the immune system to prevent far worse harm in the future, inoculation against disinformation prepares the individual to be more judicious with the online content they digest and lowers the long-term need for debunking corrections to every item of misinformation.

Just as many people needed financial incentives to get the COVID vaccine, it is likely true that many social media companies will need incentives as well. This may look like an actual tax on the spread of disinformation, similar to a 'carbon tax' placed on manufacturing companies, but more likely, it will be a higher standard for quality control set by the government and the public that social media companies will conform to. Future prebunking strategies must also adapt and respond to new and emerging forms of misinformation. There are gaps in the current body of research on inoculation, including how to prebunk against videos or images that contain misinformation and how to ensure positive long-term effects. While further research will answer some of these questions, prebunking will only succeed if supported by government policy and the public's willingness to engage. Whereas establishing the evidence base for prebunking is critical, it will only be effective if implemented in a manner that appeals to the public. In this 'post-truth' era, there is an urgent need for prebunking initiatives, since the speed at which disinformation is spreading threatens to overrun the thin lines of defense that guard our fragile democracy. 12

Appendix A

Evidence in Favor of Inoculation Strategy

1. ***Inoculation's success against extremist propaganda:*** In a [study](#) by Braddock in 2019 where groups were either exposed to [attitudinal inoculation](#) (a strategy to reduce susceptibility to persuasion by exposing people to small arguments against their stance) or a no-inoculation control message before reading politically extremist propaganda, inoculation successfully reduced participants' trust in the extremist group's credibility and predicted their psychological reactions to it. For example, when the inoculation group faced propaganda from political extremist groups on both ends of the spectrum they were able to identify the manipulation tactics of the advertisements and recognize how their emotional instincts were susceptible to exploitation. This study showed the effectiveness of inoculation regardless of political leaning and how it can be used even in situations of extremism.
2. ***Inoculation success in a study about climate change:*** In another study by Linden and others, multiple groups were exposed to varying ratios of true and false information

about [climate change](#). These statements included, “97% of climate scientists have concluded that human-caused global warming is happening” (true) and “over 31,000 scientists have signed a petition that there is no scientific evidence for human-caused global warming” (false). In the groups exposed to different inoculation attempts, participants were presented with mild statements like, “politically motivated groups use misleading tactics to try to convince the public that there is a lot of disagreement between scientists” and more contextualizing facts, such as, “although 31,000 may sound big, it only comprises 0.3% of US science graduates,” and “it is impossible that Charles Darwin could have been a real signatory.” All levels of [inoculation](#) proved effective in decreasing the persuasiveness of the false statements and, importantly, the inoculated participants’ reactions to the falsehoods were nearly identical regardless of their prior knowledge and beliefs about climate change.

3. *Prebunking effectiveness in regards to vaccination:* In a study by Jolley and Douglas in 2017, two groups were exposed to [anti vaccination conspiracies](#). The group that went through an inoculation that foreshadowed the arguments that are often made against vaccinations and *why* they are made was significantly less likely to believe or repeat the conspiracies in the long term than the other group, which was subjected instead to after-the-fact debunking attempts.

4. *The disconnect between detecting misinformation and choosing not to engage:* A [study](#) by Pennycook and others showed that when participants were shown article titles, they were able to identify accurate headlines as correct 80% of the time and only identified false titles as “true” 10% of the time. However, when asked which titles they would share with others, their previous classifications of the true vs. untrue article headlines seemed irrelevant to their decision about whether to circulate the article. The [desire to share online content](#) and engage with as many people as possible often clouds people’s judgment. This experiment was then

performed again with one modification: right before the participants were asked which article titles they would want to share with others, they were reminded of their previous commitment to accuracy. Under this condition, their tendency to share misinformation decreased significantly. Remarkably, a one sentence reminder of the importance of accuracy made people three times more likely to refrain from sharing the false information. Repeated inoculation that occurs before and during the potential sharing of misinformation is most effective.

Appendix B

Evidence Against Debunking Strategy

1. ***Inadequacy of debunking:*** In a [study](#) by Ecker and others, two groups were presented with a story about a house fire. The control group was presented with the exact same scenario twice and both times were informed that the kitchen cabinet was empty. The group exposed to misinformation was first told that the kitchen

14

cabinet had flammable materials like oil. In the second presentation of the scenario, this information was explicitly corrected, indicating that the cabinet was empty. Despite this correction, when questioned, most members of the disinformation group still mentioned the oil as a reason for why the fire got out of control. Countless studies have demonstrated that when misinformation appears relevant or supports an explanation that a person subconsciously hopes is true, debunking is unable to counter the [“continued influence effect.”](#)

2. ***Fabricated memories show the pervasiveness of disinformation:*** With politics specifically, people tend to recall and fabricate events that never happened. In a [study](#) by Murphy and others in 2019, when participants were presented with true and false news stories about a referendum on abortion in Ireland, people correctly

categorized the stories 56% of the time. Alarming, they also recalled a distinct memory for one of the false stories 37% of the time. Other studies with [US participants](#) have found a concerning pattern of belief in misinformation even after the false information had been retracted. People refused to acknowledge that they had believed a falsehood even when told they were purposely lied to and continued to center their reasons for believing it around conspiracies, religion, or distrust in the government.

3. ***The harm in misinformation retractions:*** If a statement is [retracted](#) or stated to be untrue the amount of evidence or emphasis with which it is disproved is irrelevant. Two retractions of a piece of misinformation were found to have the same effect on a tested population as one retraction. They believed the misinformation to the same extent regardless of how heavily it was debunked. Ironically, literature on metacognition has shown that because misinformation is often repeated during the process of retraction or disproving, its reach and impact can even be heightened during a debunking attempt.

Appendix C

Current Research-Based Inoculation Attempts by Companies

1. [Jigsaw's literacy tip experience](#): In Jigsaw's research project, they showed users quick graphics with one-line tips about identifying and classifying digital misinformation. Their strategy of inoculation simply shared basic manipulation tactics of disinformation threat actors to make the public more consciously aware of malicious strategies. Jigsaw found that it increased truth discernment by up to 50%. While this work was conducted for research purposes, it shows how companies may

use prebunking techniques successfully through social media. Example below:



2. **Games:** ["GoViral"](#), ["Bad News"](#), ["Harmony Square"](#)
3. **Facebook funded media literacy initiative:** [Fátima](#) (\$1.1 M in funding, Seed round funding) - Created by the Brazilian fact checking company [Aos Fatos](#), Fátima is designed as a user-friendly bot that gives social media users prompts and tips to help identify misinformation. The goal of Fátima is to allow users to be their own fact checking machines while scrolling through their social media feed.
4. [Userfeeds](#) (\$800K funding, seed round funding) - This Warsaw based startup was created by co-founders Maciej Olpinski and Greg Kapkowski. It uses [blockchain technology](#) to develop a content algorithm that relies on [ranking](#) algorithms (backed by

the exchange of digital tokens) to signal quality of information instead of the up-voting-like comment system that drives the algorithms for most social media

sites. This introduces the economic incentives around content integrity while using prebunking strategies to challenge users to look out for misinformation.

