Using Outlier Analysis for the Detection of Propagandists in Epistemic Networks

Dylan Small Anderson

Defended on April 1st, 2021

Honors Committee:

Professors Jeanne Clelland (MATH), Brian Talbot (PHIL), and Iskra Fileva (PHIL)

Honors Advisor:

Professor Brian Talbot (PHIL)

Honors Council Representative:

Professor Iskra Fileva (PHIL)

In the middle of 1983, an article was written in a small English-speaking newspaper in India. The article reported that the United States had manufactured the AIDS virus at Fort Detrick in Maryland. It was yet another link in a long chain of stories being circulated at the time that the United States was routinely violating treaties and international law regarding the creation of biological weapons. However, this wasn't simply a story about the US defense community conducting research or experiments meant to create or study potential biological weapons which could be used in warfare; the charges leveled something even more sinister.[1]

The article was published just as public attention was being captivated by the spread of a terrifying new virus. As millions of people across the globe were coming to understand the dangers posed by HIV, this story implied that the populations being disproportionately affected by the virus were at best selected for testing the effects of the manufactured virus. At worst, they were the target of its creation. The story bounced around between relatively small and innocuous publications across the world for a few years, until on March 30th, 1987, Dan Rather informed millions of unwitting Americans who tuned into CBS for their nightly dose of information that the epidemic affecting the world may have originated from their own defense community.[2]

The story was, of course, false. In fact, it had been manufactured by the Soviet propaganda arm, and carefully planted so as to distance the story from its original source. The dissemination of the story to the general public in America caused a stir in the US Intelligence Community. The Active

[1] United States. Department of State. Soviet Influence Activities: A Report on Active Measures and Propaganda, 1986-87. Washington, D.C.: U.S. Dept. of State, 1987.

[2] "AIDS: A GLOBAL ASSESSMENT : Soviets Suggest Experiment Leaks in U.S. Created the AIDS Epidemic." Los Angeles Times. Last modified August 9, 1987. https://www.latimes.com/archives/la-xpm-1987-08-09-ss-592-story.html.

Measures Working Group, an interagency group chaired by the Secretary of State, had already been working on combating the Disinformation, known internally within the Soviet Union as "Active Measures", with which our most notable foreign adversary was saturating the information economy. This story seemed more powerful than the run of the mill narratives about the US harvesting children's organs in South America though; it seemed as if it was the first time a Disinformation campaign had gained enough steam to break down the defenses put up by good journalists like Dan Rather, and manage to find its way into the living rooms of average Americans.

In 2021, however, we are all familiar with stories like this. The AIDS story was outright and demonstrably false - but we tend to think of these false stories as being something that comes with the territory of freedom of speech. We've all probably heard about stories like "Pizzagate"[3] about Hillary Clinton running a pedophelia ring out of the basement of a DC pizza parlor. We've seen how a President actively participating in the spread of false narratives surrounding important elections can pose a frighteningly immediate threat to the functioning of our democracy. The AIDs story may have taken years to spread from a small newspaper in India into Grandma's living room, but these other stories had much shorter incubation periods. What exactly are the epistemic implications of this type of information? How have they changed since the advent of social media, and the subsequent flooding of our information diets with problematic stories?

Individuals are in increasingly difficult and uncharted waters when it comes to navigating the new information economy. In the course of forming our belief systems we no longer simply rely on

[3] **Mueller, Robert S., Special Counsel's Office U.S. Department of Justice, and Alan Dershowitz. New York: 2019.**

Walter Cronkite or Dan Rather to spoon feed us relatively reliable and seemingly unbiased information about the outside world while operating under the safe assumption that our neighbors are using the same or very similar sources in the formation of their own belief systems. We are now faced with a 24 hour news cycle, social media, and a level of political punditry which all pose significant epistemic burdens for the individual seeking to form beliefs.

In contemporary discussions about the epistemic effects of Misinformation and Disinformation on the individuals operating in this environment, the distinction between different types of bad information is often overlooked. While both categories almost certainly pose threat to individuals, the information that seems to be causing us problems seems to come in a variety of different forms. For example, we may want to evaluate someone who is honestly advocating for a belief which they subscribe to differently than someone who is spreading something they know to be problematic in order to seek some sort of gain from the spreading of the belief. It seems plausible that many would find there to be an intuitive moral and epistemic difference between these two acts.

A real world example of someone spreading a belief in order to seek some sort of benefit is the tobacco industry's effective manipulation of research showing the cancer-causing nature of their product.[4] In this case, the individuals in the tobacco industry were well aware of the negative health effects of their product, but were actively pursuing the denial of those health effects in consumers in order to preserve the profits of their companies. They utilized a well-structured public relations campaign to prevent the individuals consuming their product from coming to the conclusion that

[4] O'Connor, Cailin, and James O. Weatherall. The Misinformation Age: How False Beliefs Spread. New Haven: Yale University Press, 2019.

their product was causing them harm. However, once this campaign had affected an individual, and that individual had come to the genuine conclusion that smoking was not harming them, then the action of spreading that information to their friends, family, and neighbors has cause for a different type of evaluation. They were not hoping that the people to whom they were advocating their belief would come to some harmful end, or that they would somehow get rich off of successfully persuading other individuals to join them in subscribing to a false belief.

This distinction between good and bad faith speech is how, for the purposes of our discussion, we will distinguish between the two largest categories of problematic information. We will refer to Disinformation as the information which, in our example, came from the tobacco industry. This type of information is spread in bad faith - the person or entity propagating the information has some sort of ulterior motive. In regards to the Tobacco Strategy, the content of the information was part of a narrative which was directly related to the end which was being pursued, which was to convince people to continue buying cigarettes.

Good faith "bad" information, like the advocacy of someone who genuinely believes that cigarettes are harmless, will be referred to as Misinformation. This information is propagated by people who, like every human being does at some time, have unwittingly come to believe something that is false or problematic. While in the case of the Tobacco Strategy we have an example of this information being outright false, we will also include views which are inflammatory, hyperbolic, or misleading as belonging to this category. The latter addition is made to include types of information which may be generally true, but may be lacking in completeness in some way. It is worth noting the importance of the observation that this information does not necessarily need to be false. A statement such as "I heard that cigarettes aren't bad for you" may technically be true, but the overall effect of the speech itself is misleading.

While the distinction between good and bad faith information helps us establish a clear line between Misinformation and Disinformation, respectively, a further distinction within the category of Disinformation is necessary. The narrative propagated by the Tobacco Strategy was narrative dependent. In other words, it seems unlikely that the tobacco industry's goal of maintaining profits would have been obtained by spreading a completely unrelated idea such as platypi being colorblind. The goal being sought was directly tied to the false ideas being pushed - in order for people to continue buying cigarettes, the tobacco industry needed to persuade them that cigarettes were not as harmful as they were being reported to be. In order to persuade individuals of that false belief the Disinformer (the person spreading the Disinformation - in this case the tobacco industry) needed to find research and funnel money into advocacy campaigns if they were to have a shot at achieving their end. We will refer to this "classical lying" as NBD (Narrative-Based Disinformation).

The type of Disinformation which we will be chiefly concerned with throughout this paper needs to be categorized differently. It is the type of Disinformation which was often propagated by the Soviet Union during the cold war. Internally referred to as "Active Measures", this type of Disinformation seeks to attain a larger goal which is not directly related to the content of the information or rhetoric being spread. Instead of pushing a specific narrative to achieve an end, it seeks to spread various types of information to decrease the stability of the Information Economy. This decrease in stability leads to an erosion of the individual's ability to sort out good information from the bad. It plays on psychological phenomena to which we are vulnerable in order to shift the picture of reality in a specific direction based on confirming our currently held beliefs, our fears in an ever-changing world, or some other rational process upon which we rely on to form beliefs. We will refer to this type of "chaos seeking" information as PD (Polarizing Disinformation).

It's worth noting how subtle the distinction between NBD and PD can be. After all, if we define NBD as being disinformation spread seeking a specific gain, and PD as that which is spread to erode the ability of an individual to fulfill their epistemic commitments, when an entity creates a PD campaign seeking to utilize that erosion to shift the balance of global power then it seems like PD can sometimes effectively be NBD. Take the aforementioned story of the US manufacturing AIDS for example. While the effort to convince the world that the US had manufactured the virus was underway, the Soviet health community was actively spreading the counter-narrative that AIDS was natural in its origin and was not able to be manufactured[5]. What was being sought by the Soviets in that case was the controversy that resulted from two categorically opposed viewpoints on a very specific idea needing to fight it out in the marketplace of ideas. This is opposed to what was sought by Donald Trump's spread of NBD regarding the 2020 election; he didn't want the controversy, he wanted enough people to agree with him so that the election could be overturned.

What makes PD like the AIDS story so damaging and effective is just how different every individual in our society's epistemic world is from one another. Someone who was raised on a reservation with a grandparent who attended a mandatory Indian boarding school[6] might not find it

[5] United States. Department of State. Soviet Influence Activities: A Report on Active Measures and Propaganda, 1986-87. Washington, D.C.: U.S. Dept. of State, 1987.

[6] Treuer, David. The Heartbeat of Wounded Knee: Native America from 1890 to the Present. London: Penguin, 2019.

surprising in the slightest that the US Government would intentionally infect prison populations to test the effects of a newly created biological weapon. Similarly, a homosexual who justifiably feels unwanted, rejected, and targeted by their community may feel that the new information about the origins of the virus aligns perfectly with the policies and actions taken by their government which they are affected by every day.

If either of these individuals discussed the new story with someone who, through no fault of their own, was mostly familiar with the flagship ideas of equality and justice within the US, it's easy to imagine how the epistemic can quickly become personal. Instead of arguing over something relatively objective like which math proof should be preferred to instruct new students, key components in these individuals' perspectives on the story have to do with their own experiences, and by extension their ability as epistemic agents. The person who can't imagine how someone could believe such a story immediately and passionately informing the other two of their wrongness is sending a message far more implicative then simply discounting a truth value.

What PD seeks to accomplish is to find the cracks already present in our society and exploit people's rational and justified beliefs about their government, neighbors, and society to erode the social fabric which enables us to function as a democracy. This goal has only been made easier to accomplish with the advent of social media, as illustrated by the 2016 US presidential election. As documented in what is colloquially referred to as the Mueller report,[7] the Russian government utilized Soviet tactics to spread discord with the intent of helping Donald Trump get elected to the Oval Office. While the

[7] **Mueller, Robert S., Special Counsel's Office U.S. Department of Justice, and Alan Dershowitz. New York: 2019.**

exact degree of how effective those tactics were at achieving that goal are unclear, it is sufficient to say that the potential for PD having such a large impact on our citizens has caused questioning of previously taken-for-granted appreciation for our First Amendment right to freedom of speech.

For the remainder of our discussion, we will be focusing specifically on PD, the challenges it presents, and how to combat it. The problem which we seem to be facing is one of understanding how to go about preventing PD from affecting our political reality while preserving the fundamental aspects of freedom of speech that we find valuable. It seems that a reasonable person may want someone who is spreading Misinformation to be allowed to have their speech go uncensored. After all, if the good-faith racist cannot air their views, then I am unable to challenge them, and that individual will go on subscribing to views which are problematic. In the next section, I will explain the possible approaches to solving this problem and provide an argument for why finding a way to censor PD would not violate the First Amendment, while simultaneously providing the best way of repairing the information economy.

Combating Disinformation:

The first assertion that must be established when approaching the problem of preserving freedom of speech while preventing Disinformation, and PD more specifically, from affecting our political landscape is that it is an imperative that an action must be taken. This must be established because it seems possible that a reasonable person who values freedom of speech may shy away from any action taken by an entity, specifically by the government as we will eventually establish, to limit the allowable types of discourse present within the Information Economy. There are good reasons for this conclusion - many of which would appeal to the ideas present within John Stuart Mills' argument for freedom of speech.[8]

This argument generally appeals to the idea that, as with the free market, limiting individuals' ability to freely exchange epistemic goods in the marketplace of ideas also restricts the plurality of ideas present in the marketplace, the diversity of entities with the power to shape discourse, and the ability of the individual consuming information to be a fully realized epistemic agent. In the free market economy, we generally view there to be some benefit to a low barrier of entry into buying and selling goods within the marketplace. This low barrier of entry allows for innovation to occur, better and cheaper manufacturing methods to prevail over less advantageous ones, and for everyone utilizing the marketplace to have the best chance of getting beneficial goods and services. This would be opposed to the idea of a heavily regulated market where producers and sellers lack the incentive to seek innovation due to the lack of competition guaranteed to them by the regulation. The competition resulting from the free market incentivizes businesses to continuously seek out better products and methods in order to attain a profit, and relies on the individual consumer to evaluate what goods are best for them.

Drawing the parallel to the information economy, Mill argued that a similar process occurs when individuals are exchanging ideas. It may be the case that a specific idea is true, false, or somewhere in between, but the presence of all ideas within the marketplace of ideas benefits the individual consumer of those ideas by allowing healthy reexamination of ideas which otherwise may go unchallenged. Censoring an idea because of a government's evaluation of that idea being harmful or false implies an assumed infallibility of that idea, and overall prevents the individuals utilizing the

[8] Mill, John S. On Liberty. 1863.

marketplace from being exposed to ideas which may actually be superior to those they've already been exposed to. Because this process of continuous reexamination of ideas is so crucial in the pursuit of truth, a plurality of both ideas and entities sharing those ideas are viewed as healthy goals of a healthy information economy.[9]

These goals set forth in Mill's argument are ones that I agree wholeheartedly with - but the argument itself seems to make ambitious assumptions about how history will play out in regards to the exchange of ideas and information within something akin to the marketplace. It seems intuitive upon first examination that allowing completely free exchange of information would lend itself best to every individual having the power to engage with and consume the best types of information, but the aforementioned example of the Tobacco Strategy seems to hint at a different idea.

What examples like the AIDS PD campaign and the Tobacco Strategy seem to suggest is that the marketplace of ideas might benefit from some regulation in the same way the free market does. In the free market a consumer looking to buy a product should be able to evaluate the available products of that type, evaluate them on their respective merits, and come to an informed decision about which company to purchase from. This doesn't always occur if larger companies are able to influence the market in some way to make it artificially more difficult for the consumer to either be exposed to or purchase a competitor's product. Regulation to prevent those with an advantage in the market from steamrolling new companies helps secure the benefits of having the free market.

[9] Mill, John S. On Liberty. 1863. And "John Stuart Mill (Stanford Encyclopedia of Philosophy)." Stanford Encyclopedia of Philosophy. Accessed March 25, 2021. https://plato.stanford.edu/entries/mill/.

Keeping in mind the types of regulation which we generally find acceptable in the marketplace of goods, and the specific idea that regulation helps us attain the benefits of the free market which we find valuable, we can make the parallel to the marketplace of ideas. In order for the marketplace of ideas to successfully provide the epistemic benefits which we find valuable, similar types of regulation may be necessary. This argument itself could warrant a more in-depth and explicit argument for the imperative of action, but this illustrates a reasonable view of why many may be inclined to believe that restricting speech in certain cases may be crucial to preserving the benefits we enjoy from having freedom of speech.

However, even if one accepts the imperative of action in combating Disinformation, the question of how to go about doing so is the primary question many have when contemplating recent social problems. As previously mentioned, many people would be inclined to say that spreading Misinformation should be protected - the reason being that it's easy for us to imagine any individual, including ourselves, simply believing something problematic and advocating for that idea within the marketplace. Restricting something based on whether or not it's a "good" type of information not only seems to punish us for a crime to which we are all susceptible to committing unwittingly, but also hints at an Orwellian entity deciding what should pass as an acceptable type of information. This type of Orwellian entity is exactly what one might have in mind when objecting to the idea of regulation of speech. So how should we go about taking an action which preserves our rights and the benefits of freedom of speech while also protecting ourselves from the harms of Disinformation?

There are two broad categories which possible solutions to our problem may fall within: engagement or censorship. A solution will count as engagement if it is focused on the agents within the network of individuals utilizing the marketplace of ideas. This category contains responses aimed at educating individuals on how to better identify bad sources of information, or how best to think critically about information which they are presented with. It would also include methods such as public affairs campaigns to attempt to convince the agents that a specific piece of information is problematic or wrong. This type of solution seeks to engage with the problematic information and defeat it, and seems to at least be a somewhat effective response to Misinformation and NBD in some cases. In the case of the Tobacco Strategy, the scientific community was eventually successful in convincing the general public about the health hazards associated with smoking. The engagement solution may have obtained the end goal of public consensus on the correct idea less rapidly in this case than ideal, but it nevertheless was effective at combating the consensus arriving at the incorrect or bad conclusion.

The second category of censorship is characterized by a removal of the information from the marketplace of ideas. Censorship aims to attain the goals being pursued by freedom of speech by removing ideas from the marketplace as opposed to engaging directly with those ideas. Censorship solutions would include holding people financially or legally accountable for the damages their speech caused, as well as actions such as deleting a Facebook or Twitter account or post. These solutions are in effect in different capacities in a variety of areas within our society - from restrictions on speech intended to incite riots, slander or libel, and Facebook policies of removing posts which threaten violence.

What I will argue for is a censorship solution specifically aimed at PD. This does not mean to imply that in general censorship solutions should be treated as the preferable type of solution.

Engagement solutions certainly have their place in public schooling, the scientific community, and other areas aimed at arming individuals with epistemic defenses against bad information, but for PD specifically, engagement with the information is an explicit goal of the individual spreading the Disinformation.

To illustrate this point, it is helpful to once again reference the Soviet-style Active Measures aimed at damaging the institutions of adversarial nations. This PD is most effective when the entity combating it is overwhelmed by the amount of information that they are so busy engaging in debunking, fact-checking, or mitigating the spread of the information that they are unable to allot resources to pursuing more productive goals, such as governing. In the aforementioned case of the HIV/AIDS epidemic being manufactured by the US government, the Intelligence community formed a team of individuals who wrote a report on how the story was created and its subsequent progress throughout the world until it was regurgitated by Dan Rather. This report eventually elicited a public apology from Mikhail Gorbachev, but not until well after the damage was done. Despite this public apology, Kanye West has put the story in a song many years after its debunking[10], and many redditors reported the story as a conspiracy theory they believe to be true, indicating that the short period of time that elapsed between the amplification of the story and the publication of the report debunking it was enough time for the PD to become an at least semi-permanent part of many Americans' belief systems.

This becomes even more troubling when coupled with the knowledge that PD is often manufactured to fit a range of different narratives, tailored to targeting individuals who subscribe to a

[10] West, Kanye, and Jon Brion. "Heard 'Em Say." Song. November 2005.

wide variety of beliefs, and that the individual spreading the PD is not particularly invested in any single piece of the PD they produce. Where NBD is lying to attain a certain end, combating it can oftentimes be accomplished by an equally powerful counter-narrative. PD instead seeks to throw as many lies at the wall as possible and see which ones cause chaos. Commissioning a report for every PD campaign which gains traction in the Information Economy allows the individual spreading the PD to continue spreading the divisive stories. The polarization which results then weakens the ability of institutions in charge of responding to the PD to effectively combat it, due to the extreme variety of narratives flooding the marketplace. Continuing to attempt to combat the PD by engaging it on its merits only plays the Disinformer's game with the losing strategy - the engagement of each successive piece of PD is less effective, and we now have even more PD in the queue to be countered then we did before engaging the first piece.

For these reasons, it seems reasonable that one might come to the conclusion that censorship solutions may be crucial for responding to the levels of PD that we have seen in recent years. The next question that needs to be answered is how do we avoid censoring the Misinformed in our pursuit of removing PD from the marketplace of ideas. This is perhaps the most important question to be asked when regarding any type of censorship - the difficulty of proving that someone is speaking in bad faith is what presents a challenge when pursuing prosecution for types of speech that are already restricted.

Take for example speech that is intended to incite riots. This is an example of speech which is restricted for generally uncontroversial reasons; screaming fire in a crowded theater just to watch the chaos that ensues may result in damage to people's bodies or property which provides sufficient justification to view the act as one with criminal intent. However, one can imagine that if someone genuinely believed that there was a fire in the theater and their intent was to allow those in the theater a chance to escape to safety then the speech would lack the sufficient basis for legal accountability. Similarly, the intent behind a piece of speech in slander/libel cases is what makes conviction in those cases particularly difficult to obtain. It is not illegal to tell the New York Times that you heard that the CEO of a particular company was responsible for the large number of penguins recently stolen from the Central Park Zoo. However, if it can be proven to a jury of your peers that you reported the rumor to the New York Times because you were seeking to damage the profits of a competing company then you can be held responsible for the damages that were caused.

In both of these types of speech the difficulty in responding to them with some form of censorship lies in being able to prove that the speech itself was intended to cause the harm. Notice that in both of these cases, the content of the information itself is not what makes the speech prohibited. You are legally permitted to scream “Fire!”, but when that act is coupled with the intent to incite panic then it moves into the realm of being a criminal act. If we were to seek to censor PD in any way, the bulk of the difficulty would be similar - saying “Hillary Clinton runs a pedophilia ring out of the basement of a pizza parlor” is prima facie permissible. The intent to spread that story with the explicit purpose of weakening the ability of the American democracy to function overrides the permissibility of the speech.

When seeking out a conviction for speech-related offenses, a lawyer does not necessarily rely directly on the content of what the accused individual says. If no appeal is made to the mental state of the accused at the time of the act, the circumstances surrounding the speech, or the ability of the speech to actually cause damage, then the lawyer has not made a particularly strong case for conviction.

Similarly, we can not simply look at a Facebook or Twitter post and use the blatant falsity of the information presented therein to justify censorship of the speech. This is where many attempts to combat PD may have gone astray - it is not simply enough to identify falsehoods and make an argument for censorship; the falsehoods need to be spread to attain a specific end. Detecting and proving that intent requires some methods outside the realm of identifying true information from false. In the next section I will describe how formal methods currently in use in the area of Social Epistemology may be particularly useful in identifying when that information is being spread with bad intent.

Formal Methods:

A formal framework for identifying the pieces of information which are subject to censorship lends itself well to a less controversial approach towards identifying accounts or information that should be targeted. If we can somehow provide a type of analysis which identifies PD to a sufficiently high degree, then people who have agreed thus far should have little to complain about when any specific item or entity gets censored. Should we be able to provide this type of framework, it would accomplish the difficult task of proving the intent in the aforementioned uncontroversial cases of restricted speech. It is worth noting that this framework must deviate from the framework we use to establish intent in those cases - even if a lawyer presents an astoundingly good argument for the intent behind a piece of speech, that argument is still susceptible to some level of individual interpretation. The lawyer for the individual being accused would present their own argument for why the speech should not be subject to some form of censorship, and individuals on the jury would need to perform some type of epistemic evaluation of both arguments and come to a conclusion about which one proved the relevant criteria more effectively. This would be considered an informal method of intent identification - it is subject to a high degree of human interpretation, and consequently, error.

The task we set out to accomplish by establishing a formal framework to identify the information is to prove analytically a statement of the type "95 percent of the items in the set of information which possesses these traits are PD." This specific statement is obviously up to a high degree of argument in and of itself - it's possible that 95 percent is too high or low of an error bound to be uncontroversially applied to the marketplace of ideas. I will not get into the weeds trying to provide a definite statement of the type which should be used here, but the idea I will be arguing for is that there exists some metric by which we can model or analyze speech which should help us identify exactly what criteria makes PD as effective as it is, and therefore how to best identify items of that type and target them for censorship.

An analogy here can be made to the Electromagnetic Spectrum (EMS). The EMS is used to transmit information through space on specific frequencies. When we hop in our car and tune our radio to NPR, we hear our reliable publicly funded journalists enthusiastically fulfilling their epistemic obligation to the public. If you were to look on a spectrum analyzer at the range of frequencies allotted for commercial radio use, you would detect "spikes" at all of the frequencies that you can tune to and hear meaningful sounds. You can imagine that if someone were to point some sort of Radio Frequency jamming device onto those frequencies, you would detect some other type of spike and a corresponding decrease in quality of audio. If you could program your receiver to identify the radio waves coming from the jammer and simply ignore them, then your radio would once again become clear - allowing you to clearly hear and understand the aforementioned journalists.

What we seek to do using a formal framework is very similar. The dangers PD presents epistemically is primarily one of noise - the good and bad information which is perfectly acceptable is still out there, but it's incredibly difficult to identify anything other than a tangled rat's nest of narratives. The noise prevents us from being able to reliably identify what information is actually useful, and leads us to believe that no source itself is particularly reliable or, even worse, that all sources are equally unreliable. Formal methods can help us turn down the noise while preserving the things which behave similarly enough to permissible information to arguably not pose the same epistemic threat. This last point is a subtle one, and warrants a bit of re-emphasis. By arguing for this formal framework, we are not attempting to unilaterally be able to identify all PD attempting to spread. We are instead creating boundaries which information within the marketplace must stay within. A disinformer may wish to continue spreading PD, but theoretically we should have set up the system in such a way that the only PD that gets through does not have the negative effects which we are seeking to prevent.

Recalling from the previous section, it also seems to be crucial for the boundaries to be drawn on the basis of some content-neutral criteria. If we set up the framework to detect and target based on the content of the message, then we open ourselves up not only to accidentally removing speech that we want, but also to reasonable accusations of becoming the Orwellian truth-defining entity which rational beings generally fear. There are a few measurable criteria which one may attempt to use to create this boundary.

One criteria which has been considered is evaluating the structure of the language used in the information.[11] While this isn't totally content independent, it achieves a result that might be acceptable to those worried about the Orwellian truth-defining entity. This type of analysis doesn't rely on someone ascertaining what is true or not true, but instead seeks to identify natural patterns of language used by the disinformers. A basic example of what this kind of analysis looks for are things like native Russian speakers saying "the virus flu" instead of "the flu virus". While this type of analysis has been looked at by some, a reasonable concern about its applicability might be that due to the significant variety of natural language rules used by individuals speaking the same language, we may not necessarily want to rely on speech patterns to attempt to detect the disinformers. After all, not only may an English speaker from Florida speak very differently than one from Maine, but there are people who may share speech patterns with the disinformers who are completely disconnected from the PD.

The type of analysis I favored is one of analyzing the belief within the network. Using models in use in Formal/Social Epistemology, we can create a network of agents whose behavior and interactions with one another are defined in such a way to simulate people within an epistemic network. The general approach that my analysis takes is to use a model which simulates on some basic level how people come to form beliefs and identify whether or not there is some way of detecting whether beliefs within the model are changing differently when there is a propagandist present.

[11] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." (2018).

Building off of ideas present in the works of Formal Epistemologists such as Cailin O'Connor, James Owen Weatherall, and Justin Bruner, I seek to run simulations enough times to establish a baseline representation of how beliefs are formed and shaped in a healthy network. After a baseline is established, I add an agent to the model who behaves similarly to how propagandists behave in the real world. Then, I compare how the beliefs are shaped when the propagandist is present with the baseline to see if they constitute being labelled as an outlier by some metric. If the beliefs represent an outlier, then I predict that a propagandist is present. In the model I used based on a model by Rosenthal, Bruner, and O'Connor[12], we are seeking to model a network of agents presented with a two-arm bandit problem.

Two-arm bandit model:

In the two-arm bandit problem, the agents are basically presented with two options: option A and option B. These options can represent two different levers on a slot machine, or some other experiment where the agent is presented with two avenues for accomplishing the same task. One of the options is better than the other option - lets say option A. What the agents in the model do is try and figure out which option has the higher probability of success. To answer this question, the agents conduct experiments using each option and then use the experiment's results to update their belief about which option is better. In the model, the agents know for certain that option B has a .5 chance of reward,

[12] Rosenstock, Sarita, Justin Bruner, and Cailin O'Connor. "In epistemic networks, is less really more?." Philosophy of Science 84, no. 2 (2017): 234-252.

while the option A can either have a probability of (.5 + ε) or (.5 - ε), where ε is representative of how much better option A is than option B.[13]

At the end of the experiment, we hope that the network will have achieved one of three states: correct convergence, incorrect convergence, or polarization. Correct and incorrect convergence simply means that the beliefs of all the agents in the network converged on the correct or incorrect option as being superior. Polarized signifies when there has been a type of "grouping" within the network - where different parts of the network no longer listen to one another even when they are connected with one another. The simulation consists of 10,000 trials, which run for 10,000 rounds or until the network achieves one of these three states.

The first part of a round within a trial is experimental. In order to ascertain whether option A is better or worse than option B, the agents conduct an experiment by pulling the lever a certain number of times each round. Then, the agents randomly generate their "friends list", which is the set of people that they're connected to in the network. This is meant to simulate the fact that users on real social media are exposed to information from pages which they are not in total control of. In the real world, what individuals are exposed to is based on an algorithm which is meant to maximize engagement with the platform[14], but for this initial simple simulation random generation was sufficient. After conducting their experiments and generating their friends list, all agents in the network update their belief with the results from their experiment using Bayes' rule.[15]

[13] If P(B) = .6, then epsilon is equal to .1

[14] The Social Dilemma. Directed by Jeff Orlowski. 2020. Netflix, 2020. Film.

[15] O'Connor, Cailin, and James O. Weatherall. The Misinformation Age: How False Beliefs Spread. New Haven: Yale University Press, 2019.

The second part of a round within a trial is social. After updating their belief according to the new information generated by their own experiment, they receive information from the agents within their friends list. After receiving this information, they weight how seriously to take the information by assigning a score[16] to their friend's information based on how close their beliefs are to one another. This is meant to simulate how epistemic agents behave in the real world. Drawing from an intuitive idea presented by Regina Rini, we tend to place a higher credence in evidence provided to us by those which seem to possess similar world views to our own[17].

These assumptions within the model are meant to roughly represent the process by which we take information from others on Social Media. The degree to which a specific piece of information influences the belief of another agent is a function not only of the content of the information, but also the evaluation of how reliable the source of that information is. In the model this is represented by weighting the degree to which an agent's belief is altered based on how closely aligned the beliefs of the agents are, but in the real world, there are a lot of other factors which may influence how much a new piece of information influences our beliefs; how reliable we evaluate the source to be, humor, and our own skill at identifying problematic information all factor into this process (among others). Complexity may be added to the model later, but for this initial proof of concept, the simplest interactions between agents is useful.

[16] Brier, Glenn W. "Verification of forecasts expressed in terms of probability." Monthly weather review 78, no. 1 (1950): 1-3.

[17] Rini, Regina. "Fake news and partisan epistemology." Kennedy Institute of Ethics Journal 27, no. 2 (2017): E-43.

In order to simulate Disinformation within the network, an agent which consistently spreads experimental results which favor the bad option was added to the model. This agent, known as a propagandist, poses as an agent whose belief is the average of those agents who are in their friends list, and then shares experimental results which will lead the other agents to belief in option B. This is meant to represent a disinformer targeting segments of a social network, and presenting themselves as someone who would generally be viewed as a reliable source to the group, before intentionally trying to sway the beliefs within the network in a specific direction.

The basic idea of how this model will help us understand how PD spreads is to identify patterns in how the beliefs in the network are shaped by the presence of a disinformer. In this early stage, all I am trying to show is that we can identify with some reasonable degree of accuracy that there is simply a propagandist present. The metric I am using to detect when a propagandist is present is the distance between the average belief of agents that believe in both options. This is essentially a metric which should measure some degree of polarization within the network. If agents in the network are becoming polarized faster than they are in a healthy network, then we can theoretically identify in real-time when there is a propagandist present.

The reason for choosing this metric is that the examples mentioned earlier in the paper hint that many times Disinformation is not necessarily targeted at trying to get everyone to believe something that is false or untrue to some degree, but instead to try and push groups of people subscribing to similar beliefs as far as they can apart from one another so healthy discourse seems impossible between those of different groups. By measuring how the beliefs of groups that believe in each option are being manipulated by the propagandist we may be able to identify whether beliefs being shaped in a certain way would be considered an outlier for what would normally be the case if everyone in the network was spreading information in good faith.

To put in less convoluted terms, we may be able to identify how a healthy network of agents forms beliefs. If we can analytically see that beliefs are being formed in a way that would be incredibly weird for a healthy network, then we may be able to safely conclude that there is a propagandist present. In the real world this would be much more complicated - as already mentioned, belief formation in the real world isn't so clean and easy to understand, but the basic idea appeals to what we mentioned earlier in the paper. In order for PD to be effective, it must be influencing belief formation in some way which would deviate from a healthy network sharing information and forming beliefs. If we can identify a metric which effectively measures the change which the disinformer is trying to affect, then we may have a method of identifying and censoring that PD without violating the First Amendment.

I ran a total of 120,000 trials, 10,000 for each box in the right-hand column of the following table. The size of the network refers to how many agents are in the network, P-difference represents the difference in probability of reward for the two levers, and the propagandist column indicates whether a propagandist was present within the network.

Size	**P-difference**	**Propagandist**
5	.01	No
		Yes
	.005	No
		Yes

7	.01	No
		Yes
	.005	No
		Yes
11	.01	No
		Yes
	.005	No
		Yes

The results from the simulations I ran were underwhelming. There were a few definite trends that were observable by adding the propagandist that were not totally surprising: propagandists increased the likelihood that a network would converge incorrectly, but this was only particularly observable in networks with five agents. This makes sense for a few reasons. As the number of agents in the network increased, there was still only one propagandist. When the agents generated their friends list, they were connected to one third of the network. This meant that as more agents were added to the network, there was more reliable information to overpower the Disinformation being spread by the propagandist.

The reason the propagandist was more effective at achieving incorrect convergence was due to the tactics that I implemented in their behavior. The propagandist represented their own belief by looking at the agents in their friends list and broadcasting their belief as the average belief of those agents. This might be somewhat effective at having the propagandist affect the group which they are connected to, but it might not be very effective at causing the groups to polarize. More discussion of this will come in the Future Work section.

While these observations probably stemmed from the propagandist being too weak in some form, the main problem I had with the simulations was planning and analytical. I was able to easily keep track of how many trials resulted in correct or incorrect convergence, but what I needed to actually keep track of was the beliefs at each round in the simulation. Essentially, I was trying to do the analysis of the model in real time as the simulation was running. It will be immensely easier for me to do my analysis if I instead run the simulation, create the data, and then do analysis on the data.

Even though these results were not exactly what I was looking for, it informed how best to continue doing this research in the future. One of the most significant benefits these simulations provided me was an awareness of the gaps I will need to fill as I continue, such as coding experience, knowledge of probability, and awareness of literature within Formal Epistemology. I am glad to have conducted the experiments in order to best set myself up for continuing with the project in future works.

Future Work

My goal for this project is to continue conducting research over the summer and have something which may be accepted for publication by August of this year. These initial results were meant to simply show a proof of concept that there may be some content neutral criteria by which we could detect the spread of PD. For future work, I plan on restructuring the code I used for the model in order to make the propagandist a little bit easier to implement, as well as allowing for easier manipulation of the propagandist's tactics. The purpose for this restructuring will largely be to incorporate a more diverse set of propagandists within the model. One of the most reasonable criticisms of how the model is currently implemented would be that it seems like the propagandist functions more like an agent spreading NBD than one spreading PD. There are a lot of questions that need to be answered before an effective modeling of a disinformer spreading PD may be successful that might not necessarily be obvious to the general audience.

For one thing, in the current model there is a single propagandist that simply poses as someone who would appeal to the average of the people they are connected with for that round. While this might be effective at illustrating what the disinformer is aiming to do in some cases, one can easily imagine how that tactic wouldn't necessarily work in the real world. It may be helpful to simulate having multiple propagandist nodes that are controlled by the same agent. To illustrate why this might be helpful, and why it may pose a more significant analytical challenge in real social media accounts, we can refer to the case of Russian PD accounts targeting two separate groups of social media users in the lead up to the 2016 presidential election.[18] The end result was two very agitated groups of protestors screaming at one another across police barriers, but the cause was two separate Facebook groups created and run by the FSB.

This is a good example of the type of PD which I am generally targeting with my analysis - the people who were exposed to the group's rhetoric were unaware that they were reacting to foreign influence, and the passion that the rhetoric incited made national headlines; reinforcing the idea that

[18] Mueller, Robert S., Special Counsel's Office U.S. Department of Justice, and Alan Dershowitz. New York: 2019.

Americans are simply at each other's throats and unable to make productive change happen in their country. My model currently is aimed at simplicity, but this kind of PD might not benefit from analysis aimed at detecting abnormal behavior from single accounts. In this case, detecting the PD would rely on our ability to detect the PD coming from two different nodes on the network, targeting two separate groups, and behaving in a coordinated manner to weaponize the passions of other nodes in the network. The question that seems to be on the horizon for this line of research is not only whether or not we can detect individual accounts that are attempting to spread PD, but also being able to identify when accounts identified as spreading PD are coordinating together. This ability would allow us to have an informed opinion about what cracks the disinformers are targeting in order to cause chaos. Having an awareness of where the disinformer is targeting may allow the relevant entities to preemptively spread awareness about the campaign. This tactic may be an alternative to censorship, though it may also be helpful to use it in tandem with the general methods outlined in this paper.

The question of how to detect coordinated group propagandist behavior within the network is primarily one of mathematical analysis. The relevant philosophical questions that I will eventually need to answer are voluminous. The first, and the one which all others will generally stem from, was outlined in the first part of this paper - the question surrounding freedom of speech in a democratic society when that freedom of speech itself seems to pose an existential threat to democracy. The answer of exactly whose responsibility it is to control the PD affecting the democracy is one that warrants an extensive argument; while social media companies will inevitably be the ones who implement whatever analysis can be created to identify PD, many people are uncertain whether the government should hold those companies accountable for speech that is present on their platforms. For some types of speech, such as violent or extreme examples of hate speech, social media companies already censor largely out of a desire to avoid government regulation of their platforms. The question of whether or not the political effects of this PD provide sufficient justification for an argument to be made that they must handle it or be subject to government regulation is one that still needs answering.

In this same vein, an additional question that many would have about the type of speech present on social media platforms is whether or not censorship should be utilized to mitigate the spread of the much larger category of hate speech on these platforms. While this argument may not rely so heavily on analysis to identify hate speech, there is a question of whether or not hate speech itself should be protected by the First Amendment and whether a similar argument from earlier in the paper could be used to justify its censorship. Reaching back to the example about screaming fire in a crowded theater, an interesting question is whether or not we should allow someone who is operating under impaired mental faculties to continuously incite riots using their speech. Imagine if someone in a psychotic state repeatedly entered theaters, and through no fault of their own, imagined themselves to have seen fires when there definitely was never actually a fire. After some type of validation of the person's impaired mental state, many might say that while it may be unjust to arrest or otherwise legally hold the individual accountable for that speech, it may be a moral imperative for the government to ensure the individual no longer enters a theater.

The line of questioning which may be the most pressing is what justification PD confers in terms of Just War Theory. If the Russian Navy was off the coast of California and was about to send in landing craft the US response would be swift and decisive. Evacuations would take place, reserve troops would be activated, and munitions prepared - it would be the type of act of war which we are familiar with and understand how to react to. In this case, the enemy would be at the gates, the entities entrusted with our defense would position themselves between the citizenry and the threat, and the citizens would (at least in theory) be protected from the bullets and bombs aimed at ripping them apart.

What implications does it have on National Security when the opposing forces are not attacking with traditional munitions, but instead with divisive memes and lies that spread like wildfire through a population with an already fraught relationship with their government? Instead of the government, or whatever entity you wish to assign the responsibility of protecting the citizens of a nation, being able to get in between the threat and its citizens, these PD campaigns seek to weaponize the citizens themselves against the government. These are the same citizens which would need to support the response aimed at protecting them from the threat which they are facing, but when the point of the attack is essentially to weaken a governing body's ability to function that responsibility is likely to remain unfulfilled.

If a foreign entity can influence another nation's elections such that they democratically elect the political equivalent of a drunk driver, those actions can have at least as damaging an effect as an attack which would be uncontroversially seen as an act of war. However, the philosophical basis for responding to PD campaigns in a similar way as we would to traditional acts of war is still an open question. It seems reasonable that many would agree that the government should do something at the very least, the implications on a sovereign nation's foreign policy could be extremely extensive or relatively narrow. These all seem like questions which Philosophical argument is uniquely equipped to provide reasonable solutions to, and I am hoping to be a part of the generation of philosophers which confronts them.

Works Cited

"AIDS: A GLOBAL ASSESSMENT : Soviets Suggest Experiment Leaks in U.S. Created the AIDS Epidemic." Los Angeles Times. Last modified August 9, 1987. https://www.latimes.com/archives/la-xpm-1987-08-09-ss-592-story.html.

Brier, Glenn W. "Verification of forecasts expressed in terms of probability." Monthly weather review 78, no. 1 (1950): 1-3.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." (2018).

"John Stuart Mill (Stanford Encyclopedia of Philosophy)." Stanford Encyclopedia of Philosophy. Accessed March 25, 2021. https://plato.stanford.edu/entries/mill/.

Mill, John S. On Liberty. 1863.

Mueller, Robert S., Special Counsel's Office U.S. Department of Justice, and Alan Dershowitz. New York: 2019.

O'Connor, Cailin, and James O. Weatherall. The Misinformation Age: How False Beliefs Spread. New Haven: Yale University Press, 2019.

Rini, Regina. "Fake news and partisan epistemology." Kennedy Institute of Ethics Journal 27, no. 2 (2017): E-43.

Rosenstock, Sarita, Justin Bruner, and Cailin O'Connor. "In epistemic networks, is less really more?." Philosophy of Science 84, no. 2 (2017): 234-252.

The Social Dilemma. Directed by Jeff Orlowski. 2020. Netflix, 2020. Film.

Treuer, David. The Heartbeat of Wounded Knee: Native America from 1890 to the Present. London: Penguin, 2019.

United States. Department of State. Soviet Influence Activities: A Report on Active Measures and Propaganda, 1986-87. Washington, D.C.: U.S. Dept. of State, 1987.

West, Kanye, and Jon Brion. "Heard 'Em Say." Song. November 2005.