Information hygiene and the human levers of disinformation

By Azim Khodjibaev and Ryan Pentney
# Information hygiene and the human levers of disinformation

## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>The social media ecosystem</td>
<td>3</td>
</tr>
<tr>
<td>The filter bubble, recommendations algorithms and microtargeting</td>
<td>4</td>
</tr>
<tr>
<td>Bots</td>
<td>5</td>
</tr>
<tr>
<td>Fake personas</td>
<td>6</td>
</tr>
<tr>
<td>Propagation tactics</td>
<td>6</td>
</tr>
<tr>
<td>The levers of human behavior</td>
<td>6</td>
</tr>
<tr>
<td>Continued influence effect</td>
<td>8</td>
</tr>
<tr>
<td>Affect of disinformation on information consumption</td>
<td>8</td>
</tr>
<tr>
<td>Divisiveness as instruments of foreign disinformation</td>
<td>9</td>
</tr>
<tr>
<td>Polarization and piggybacking</td>
<td>9</td>
</tr>
<tr>
<td>The role of memes</td>
<td>9</td>
</tr>
<tr>
<td>Disinformation and micro-targeting</td>
<td>10</td>
</tr>
<tr>
<td>Using targeted ads to drive followers to a page</td>
<td>11</td>
</tr>
<tr>
<td>Practicing information hygiene</td>
<td>13</td>
</tr>
<tr>
<td>How to identify bots</td>
<td>15</td>
</tr>
<tr>
<td>Speaking with friends and family</td>
<td>16</td>
</tr>
<tr>
<td>Mitigation and policy</td>
<td>16</td>
</tr>
<tr>
<td>Talos and the Fake News Challenge</td>
<td>17</td>
</tr>
<tr>
<td>Conclusion</td>
<td>17</td>
</tr>
</tbody>
</table>
INTRODUCTION
As the 2020 U.S. presidential election draws nearer, it recalls still-active memories of reports of foreign interference in the 2016 election. Crucial to this interference was an unprecedented and sophisticated series of online disinformation campaigns that exacerbated partisan entrenchment, influenced voters and fostered doubt in the American election system — and Western-style democracy as a whole. As the adversary’s methods centered around manipulating sentiment rather than vote tallies, we set out to more deeply understand the ways in which the human mind itself can emerge as an attractive attack surface.

In Talos’ “What to expect when you’re electing” report, we shared what we learned about U.S. election systems’ architecture by working with and learning from the local, state, and federal officials who oversee them. In Talos’ follow-up report, “The building blocks of political disinformation campaigns,” we shared what we learned about the technical and logistical infrastructure actors behind these campaigns leveraged. Now, we’ll turn to the human mechanisms these operations rely on, why they’re effective, the long-term impact on our relationship with online information, and what steps each of us can take to become more responsible consumers of information.

One thing to note as we move forward is the difference here between our usage of the terms “disinformation” and “misinformation.” Disinformation is commonly defined as the intentional spreading of false information with the intent to deceive, while misinformation is the unintentional spreading of false information that is not meant to cause harm. While the definition of disinformation seems straightforward, it is often difficult to determine how to apply the label, as the information in question often includes portions of truth mixed with fabrications. In fact, information that is first employed deceptively may, in turn, find itself spreading through unsuspecting participants with the sincerest of intentions.

THE SOCIAL MEDIA ECOSYSTEM
While disinformation campaigns depend on technical tools and infrastructure to be successful, they are ultimately concerned with manipulating the feelings and behavior of people by driving false narratives that serve particular goals. Depending on the actor behind a campaign, the goals and level of sophistication can vary widely, but online disinformation tends to find maximum traction by leveraging social media platforms. To that end, it's helpful to take a step back and examine the factors endemic to our relationship with these platforms that contribute to shaping the environment into fertile disinformation ground.

Social media platforms have unique features that allow them to stand out among their competition and attract users. Disinformation campaigns often focus on these unique features and weaponize them to spread their desired messaging. For example, one of the main features of Facebook's monetization model is targeted ads and posts. Users can manipulate their ads for their groups, pages, or products by using internal demographics tools. As we saw in 2016, Facebook’s ad targeting features were weaponized to spread targeted disinformation about sensitive political topics to voters in key swing states. Similarly, very popular platforms such as Instagram and Twitter have ad targeting systems that function similarly.
These ad targeting mechanisms tie into a greater monetization model that is incentivized to maximize users’ interactions. For example, many users post daily updates about themselves. These posts gain engagements that carry emotional responses such as “Likes” or other reactions. Part of the weaponization of social media through disinformation strongly relies on this emotional attachment. A user may find themselves interacting with a disinformation post simply because it makes them feel good, validated and/or vindicated. Following this model, posts that are emotionally charged, regardless of the veracity of the content, tend to be successful in soliciting engagements. This, along with ease-of-use and instant accessibility, make social media an attractive tool. A prominent example of this is the Heart of Texas social media profile that was directly linked to trolls from the Russian propaganda firm Internet Research Agency. Heart of Texas attempted to create provocative memes that tied the results of the 2016 U.S. Presidential election to an existential threat. In the example linked above, the meme attempts to tie violent criminal behavior to immigration policy and religion. The social media platform’s counter shows that it solicited thousands of emotional responses that ultimately helped spread the meme — and by proxy — the message.

Finally, the backbone of social media experiences is the connections between users (e.g. friends, followers, etc.). These are the accounts that users interact with a significant amount, whose content they read, and with whom information is shared. For many people, a large portion of this group includes people with whom they have a high degree of familiarity, such as friends, family, and coworkers, or larger accounts from social media influencers that can reach a sizable audience. The established social (or sometimes parasocial) relationships with these users confer a certain amount of authenticity and trust in the information coming from those accounts.

**THE FILTER BUBBLE, RECOMMENDATIONS ALGORITHMS AND MICROTARGETING**

It is not uncommon for human beings to want to associate with groups that affirm preexisting points of view. For example, many individuals select their acquaintances based on shared experiences and characteristics. This unconscious selection process can often filter out those with differing perspectives and result in interactions with only a homogenized world view. This dynamic not only plays out in real life, but also on social media, and is interwoven in the way we intake and make sense of information. Online “echo chambers” can come about when interactions and sources of information become so tailored that our exposure to competing voices is either shut out or filtered and misrepresented through the lens of those we already agree with.

At a time when people are increasingly getting their news through social media, one of the most pernicious contributors to polarization is the ability social media platforms give their users to personalize and curate what voices they’re exposed to. Users can filter various content and opt out of undesirable sources of information. This highly customized experience is sometimes referred to as the “filter bubble” for the way in which it can insularize users away from competing voices and perspectives they might find confusing or conflict with. However, conscious actions users can take to shape their experiences only account for part of this. The algorithms on social media platforms that govern what additional content to recommend to their users play a powerful role in narrowing what users see to more of what they’re likely to already want to see. And while platforms such as Facebook have attempted in recent years to take steps to shake up the designer realities platforms help create for their users, such as removing the ability to personalize the trending topics page, social media platforms ultimately thrive (both mechanically and financially) on engagements. As such, without significant changes to how they operate, there will always remain an incentive to continue to show platform users content they are likely to engage with.

One example of this is the socially driven news aggregation platform Reddit. It is a very popular platform for information and is driven by human interaction. Popular topics are voted up or down, where topics and comments with more upvotes are driven to the top of the page where they are more visible. This limits the user to an environment where information is presorted by consensus and the subjective filters users create may inadvertently validate unverified pieces of information. Reddit users, aka “Redditors,” can filter their interests based on certain “subreddits.” These subreddits are moderated groups that focus on particular subjects and interests. Inevitably, this process allows Redditors to create bubbles where their intuition for information becomes validated by the filters they created. The customization also allows Reddit’s algorithm to match
users with new interests to drive more user interaction. These algorithms are often based on the behavior of Redditors with similar demographics, further enhancing the customized echo chamber.

Another problem with so-called filter bubbles — beyond just their tendency to limit our exposure to information we don’t like — is the continual feedback they give to platforms about what individual users’ preferences, affinities, and ideologies lean toward. A user will get more similar content based on what they “like” and who they interact with. This allows for more precise information targeting, a system which has been weaponized by disinformation campaigns to finely tune their messaging for disparate target groups. For example, in 2016, internet trolls based in Russia created fake personas, groups and content on several social media platforms. They then used the platform’s features, including ad targeting to spread false or emotionally volatile information. The IRA used campaigns that leveraged politically inflammatory imagery and language to target gender, LTGBQ, and race relations among others (Figure 1). The aim of the race relations campaign, for example, was to sow distrust among the African American community. This affords disinformation operations a delicate degree of control with which to maximize the success of narratives with different audiences.

**BOTS**

Bots are automated user accounts that amplify a disinformation campaign’s messaging on platforms and outlets. The activity can often look like targeted and amplified harassment of an individual or group, also known as “brigading.” The tactic is used with the intent of attacking an undesirable message. In one notable campaign, bots were even used to manipulate online polls and other interactions.

Bots can also be programmed to hijack trending topics. If a topic is gaining momentum and hits a certain threshold, bots may pick up on that popularity and take advantage of it by sharing a similar message that is curated to their own ulterior messaging.

Bots may also be employed as amplifiers to increase the aura of authenticity, validity and popularity of a given piece of information. They can also sow doubt both in legitimate information and the sources from which they would usually receive that information.

Cisco Duo [conducted research](#) into bots and categorized them as such:

- **Content-generating bots** actively generate new content, such as spam or links to malicious content.
- **Amplification bots** like and retweet content to artificially inflate the tweet’s popularity
- **Fake followers** are bots that artificially inflate the number of followers for an account to make that account seem more popular — and thus credible — than it actually is.
Information hygiene and the human levers of disinformation

FAKE PERSONAS

We’ve previously looked at the account creation process and the roles different accounts play in either establishing, magnifying or reintegrating disinformation. We saw that accounts with an established history tend to be more valuable because of their relative perceived legitimacy over newly created accounts. Another aspect of account legitimacy worth examining is the persona created for the account in question. While the personas themselves may be fictitious, our assumptions about who the person behind the account is can increase faith in the legitimacy of the information propagated through it.

BBC published a list of the different types of people they observed spreading disinformation during the COVID-19 pandemic. These categories appear to be modeled after real people behind legitimate accounts, but some of these archetypes manifest as personas constructed by the actors behind them. Moreover, the sometimes-high difficulty involved in verifying the authenticity of individual accounts can make these especially attractive when employed intentionally as playbooks for operating fake personas.

The insider persona, for instance, is an account claiming to be an authority on a given matter, either by being an expert in the field or someone who has access to inside information. Likewise, conspiracy theorist accounts may serve a similar function for those with an already established and deep distrust of conventional media. An example of both persona types is the 4chan persona “Q” which contributed to the ignition of the QAnon conspiracy. Q claimed to be a high-level government insider in possession of classified information about an evil worldwide conspiracy that only U.S. President Donald Trump could put an end to.

PROPAGATION TACTICS

While employing a myriad of signal-boosting bots can help flood social media with a message and ensure it gets seen, that alone is no guarantee that the message itself will be absorbed and propagated by its intended recipients. Rather, the message itself needs to be constructed in such a way that maximizes the likelihood of it resonating and influencing the target group. This is where rhetorical techniques come in.

Rhetorical techniques can be employed not just to influence or strengthen a viewpoint, but also to confuse observations, conflate ideas, sow cynicism and ignite emotions. By carefully crafting spurious claims and fallacious pivots of reasoning, disinformation actors can achieve aims such as circulating false facts, discrediting competing messaging, and seizing control of ideological narratives.

While these techniques are often observed in conjunction with foreign influence campaigns, it’s important to note that they didn’t originate with them. Part of the reason these techniques are employed is because they’re pre-existing artifacts of human discourse that have a proven track record of being effective, particularly in an age when information can circulate faster than the speed of rigor.

The Center for European Policy Analysis (CEPA) published a list of the most common techniques leveraged by Russia disinformation operations to spread false stories and disinformation. The list was recently moved, but a cached version is still available via the Wayback Machine. Learning to spot these techniques is useful to identify attempts to actively and purposely misinform and scrutinize the methods we as individuals use to pass along information.

THE LEVERS OF HUMAN BEHAVIOR

Issues of identity and the communities we belong to tend to play a large role in shaping how we experience the world as
Information hygiene and the human levers of disinformation

Here are a few common techniques worth highlighting commonly used by disinformation actors:

**Whataboutism:**
This is an evasive tactic you’re likely to see just as often in responses by governments to foreign criticism as partisan political arguments between individuals. It boils down to “how can you criticize A but not B.” For example, the Chinese government’s criticism of the U.S. authorities’ response to the Black Lives Matter protests that occurred in mid-2020.

**Drowning facts with emotion:**
When the emotional presentation of a piece of information supersedes the importance of verifying the claim. During the 2016 U.S. presidential campaign, a widely circulated image of a woman said to have been beaten at a rally for supporting then-candidate Donald Trump turned out to be a miscaptioned photograph of actress Samara Weaving in on-set makeup. The apparent disturbing nature of the image was meant to scare users into sharing the claim without scrutiny.

**Card Stacking:**
When a false claim is packaged with facts that are otherwise verifiable and true. This may increase faith in the original claim by association and reduce the likelihood of it raising flags. For example, birther movement disinformation asserted that former President Barack Obama was not born in the U.S. Facts such as Obama’s biological father hailing from Kenya were manipulated to legitimize the false claim of his birth.

**Narrative Laundering:**
This is when a narrative or talking point is repackaged and resourced so that the actual origin can be reassigned. This is often achieved by having an expert of questionable integrity or background stand in and advance a false claim. Once this happens, the origin of the narrative can cite the newly assigned source as justification.

**Changing the quotation, source, or narrative:**
Similar to card stacking, but this technique relies on smaller, more subtle alterations of an existing piece of news to change its original context. An example of this might be altering a word in a quote, removing it from its context, editing a video or audio clip or falsely summarizing an article to attribute a new message to its source. Another highly effective but more subtle example of this is when a news story is published with an unrelated image next to it. The misleading connection between the article and the image may grant significance to the image for the reader that helps to validate an otherwise dubious story.

Well as our framework for making sense of new information. The advent of the internet and social media has brought with them the ability to freely carve out and associate with like-minded groups independent of geographical separation. Naturally, this has an impact on our disposition towards information that plays on divides between the identities we associate with and those we might be in conflict with. Information that paints ideological narratives in a flattering light may be more easily accepted and rebroadcast than information that does not, and becoming absorbed in an “us vs. them” conflict can similarly short-circuit our willingness to interpret information critically.

The human tendency to form strong (and sometimes arbitrary) loyalties on the basis of ideology, identity, and community is often referred to as “tribalism.” The need to belong underscores the human vulnerability we feel as we attempt to learn and interact with the world. Disinformation campaigns take advantage of tribalism to increase trust in a targeted message and motivate the sharing of it by...
information hygiene and the human levers of disinformation

incorporating identity and sense of belonging into the process. This may work by coupling certain beliefs/positions with identity to the degree where failure to conform may threaten a person's sense of belonging. This need to reaffirm membership to a group may facilitate the acceptance of a piece of information without any scrutiny. In one analysis of tribalism, we see that these types of disinformation campaigns use clearly seen divisive lines. These lines may look like “us vs. them” campaigns. Often, these campaigns are accompanied with some sort of existential threat that compels urgency, such as if we don’t rally around our candidate, then life as we know it is over. As mentioned in the analysis, this type of rallying among a group of people goes further than populism, where a more targeted approach is necessary.

We found the research by Adam Waytz, associate professor of Northwestern University’s Kellogg School, particularly enlightening when looking at the psychological dynamics for information processing. He breaks down information processing into the following mechanisms:

**Naive realism:** The mistaken belief that we see the world through an objective lens and not through the lens of assumptions, values, and experiences that are unique to us. Naive realists may feel, for instance, that those who disagree with them must be poorly informed, or worse, concealing an insidious agenda.

**Motivated reasoning:** An inherent preference for a particular conclusion that affects our ability to accept contravening evidence resulting in an uneven weighting of evidence that challenges an established belief.

**Social consensus fallacy (wisdom of the crowd):** Ascribing truth/falsehood based on what popular consensus (e.g. our social media networks) are telling us, resulting in fast and shallow evaluation of information.

CONTINUED INFLUENCE EFFECT

Even when disinformation is identified, rooting it out entirely can be an uphill battle. The continued influence effect is a phenomenon wherein people demonstrate a tendency to hold onto the information that arrived first, even after that understanding is then updated. This effect plays out when individuals are confronted with new information that would rescind or correct an initial claim.

We see this in the common false idea that vaccinations cause autism. There have been many studies by different groups and health institutions that have not found any evidence of this claim. One flawed study, however, published in 1998 by Andrew Wakefield suggested a possible link between the MMR (measles, mumps and rubella) vaccine and autism. Although Wakefield’s research has since been debunked, his work is still being used to propagate misinformation about vaccinations. As a result, many people who were initially convinced by the anti-vaccine falsehood continue to believe that vaccinations cause autism, even after presented with clear evidence that is not the case. Once the falsehood has been established, it is difficult to undo the impact simply through fact-checking.

EFFECT OF DISINFORMATION ON INFORMATION CONSUMPTION

One of the most disturbing side-effects of our increased exposure to and awareness of the potential for social media to amplify and propagate disinformation is how it has shaped our relationship with our consumption of online information.

Naturally, people have adapted to this exposure to disinformation campaigns by developing a greater skepticism for the information they are ingesting. While this may sound like a constructive adaptation, a Pew Research Center study suggests that this scrutiny may be reserved for only the information that comes into direct conflict with existing beliefs. This can manifest as suspicion of every piece of news, dismissing real news as fake, and making us less likely to consume or accept information. The net effect has been to increase entrenchment rather than promote an overall stronger, more critical and conscious evaluation of the information we ingest.

The sheer overwhelming volume of information social media users are regularly subjected to presents yet another challenge in discerning truth from disinformation. Under this level of cognitive strain, it becomes increasingly unfeasible to critically evaluate and litmus check every piece of new information. Instead, we begin to develop broader filters, more hastily discarding what we see as irrelevant or undesirable information. This environment is ideal for fostering the success of disinformation campaigns, which are ultimately designed to overwhelm us and train us to disengage.
DIVISIVENESS AS INSTRUMENTS OF FOREIGN DISINFORMATION

Some of the long-term effects of polarization are widely known but others are less obvious. Studies suggest that high degrees of societal polarization in western democracies decrease resilience to online disinformation. The further the divide between basic observations of reality, the more difficult it is for citizens to tell the difference between true and false information.

But aside from heightening the overall credibility of disinformation, highly polarized environments also increase a disinformation actor’s available attack surface. A Harvard Kennedy School review found that “Russian social media trolls exploited racial and political identities to infiltrate distinct groups of authentic users, playing on their group identities.” This suggests that rather than doing the work of constructing a narrative from scratch, actors can latch onto existing ones with already established baggage and utilize them to circulate messaging. This way, social media users already caught up in the throes of clashes between group identities may become unwilling carriers of disinformation.

POLARIZATION AND PIGGYBACKING

Disinformation campaigns exploit polarization by piggybacking their messaging on a divisive issue. This causes an already polarized sensibility to a subject to fall under a larger group of beliefs and ideas. Sometimes, this new messaging can replace a group’s understanding of a movement entirely. We can see an example of this in recent media coverage of the Black Lives Matter movement. In August 2020, Fox News highlighted President Trump’s referral to the Black Lives Matter movement as “Marxist,” a topic that is anathema to conservatives. Marxism is not at the core of the movement, but rather racial injustice, police reform and systemic bias.

THE ROLE OF MEMES

As mentioned, one effective way for disinformation messaging to find footing is by targeting heightened emotions. Our emotions have a strong effect on how we perceive, process and store information. One method for achieving the emotional effect in an efficient and targeted manner is by using memes. Memes are not just pictures of a culture reference with words on them but can be anything on the internet that fits into a particular trend, whether it be text, images or a video. While memes may seem like a recent phenomenon, they have always been around. We have seen them in political cartoons, art, media and even share them verbally as anecdotes among ourselves. Given all of these emotional and cultural interactions, disinformation campaigns have historically leveraged memes as a utility to convey their message. Therefore, memes will often use recent cultural references that most people are familiar with. The cultural reference establishes the mood, theme and emotional plane instantly. At that moment, the emotional state of

The function of a meme isn’t to prove or establish a truth, but to establish an idea the perceived truth may be flawed.
Information hygiene and the human levers of disinformation

Mind has been manipulated and the target then begins to interact with the accompanying message. This particular interaction has a major influence on how we store that piece of information. At that point, correcting that information to truth will already require some energy. One of the best examples of disinformation campaigns and memes was seen during the 2016 U.S presidential election. Memes were used to convey certain political positions and eventually a phenomenon occurred on social media that was dubbed “Meme Wars.” Warring sides would create and spread provocative memes to convey or counteract the other side’s memes. During this time, a state-sponsored actor (Russia) interfered with these meme wars by creating memes of their own and paying hundreds of thousands of dollars to spread them across social media.

DISINFORMATION AND MICRO-TARGETING

Leveraging ads within the social media framework is easy to do with several options available to an adversary to allow them to spread a false message.

Anyone who uses social media, especially Facebook, has likely been subjected to ads like the following. Ads are served to users based on the information Facebook collects about them, which includes information about the user’s interests, actions, and connections, as well as other parameters such as previous clicks and post comments. The more history the user has, the more data they provide to the platform, allowing Facebook to show the user content matching their interests.

Provocative ads are designed to grab the attention of users interacting with a platform. They can be flashy and/or use a lot of color and enhance their imagery. The messaging is often simple so that it can be easily absorbed by its target audience, sometimes posing a provocative question to the viewer. Figure 2 shows an example from a Facebook ad created during an election in the United Kingdom incorporates several of these features, combining color and selected imagery to trigger emotional responses, and attaches a simple and clear message built on existential fear.

Ads intended to draw followers and likes to pages/groups funnel traffic to specific pages or groups. They take the form of posts, memes, provocative statements, or anything that is likely to generate an emotional reaction and compel the viewer to share it or comment on it. Those interactions,
Information hygiene and the human levers of disinformation

in turn, help facilitate further propagation by increasing the popularity of the ad. Simply put, the more interaction, the more people will see it. Figure 3 shows an infamous Facebook ad that has been linked to Russian election interference, provocative imagery is used that ties the results of the election to religion. The text in blue at the bottom makes clear its intent to use interactions to help further promote the message.

Facebook does have policies governing ad content. Anything deemed inappropriate by their community standards will be removed. This includes content such as sexual or violent imagery. Additionally, ads regarding social issues, politics, or elections targeting or originating from designated regions must be authorized by Facebook and require more detailed information about the advertiser. We have found, however, that you can bypass that method simply by focusing on the kinds of ads that drive traffic and likes to your page or group.

For example, a Facebook page with an insidious anti-pineapple-on-pizza agenda may first attempt to gain followers by simply attracting pizza enthusiasts, using an engagement-soliciting ad of the form “like and share if you think pineapple on pizza is bad.” Once it reaches a desired number of enthusiasts, it can then create posts that are not promoted and leverage the organic interactions of followers to facilitate propagation. Conceivably, an adversary can create a page and use their resources to garner thousands of followers and then create and post disinformation in hopes that the artificial echo-chamber will spread the message through interactions such as likes and shares. This can be achieved without the actor having to reveal detailed information about themself.

It’s worth noting Facebook has implemented control policies for disinformation when it comes to major events that have implications of life and death during their policy update on the COVID-19 pandemic. The policy update aims to leverage fact-checkers and their internal mechanisms to quickly remove any false information about the pandemic.

**USING TARGETED ADS TO DRIVE FOLLOWERS TO A PAGE**

The user information Facebook collects is used to create a digital profile that enhances the ad creator’s ability to conduct very precise ad targeting. The process to create and promote ads within Facebook is pretty simple (Figure 4). Let’s run through the basics to see how simple it is.

The ad creator can be anyone with a Facebook account. After they have established a themed entity (for example: a group for pizza lovers) they can then create a group or a page that provides a centralized area for their messaging. Once created, the group or page will need to establish an aura of popularity by increasing page likes (and users who like a page automatically follow it). Due to the perception that popularity is a good litmus for legitimacy, this enhances a group’s or page’s sense of authenticity for the unsuspecting user that may be targeted. Additionally, this helps Facebook
Information hygiene and the human levers of disinformation

reach a wider variety of users by introducing new data about demographics, friends on the platform and content interaction history.

Facebook’s built-in ad-creation features to allow you to do exactly that. The ad can be created with a simple message and image (Figure 5). Facebook provides their estimated daily results and the price it would cost to reach that audience. You can even increase your budget for more exposure.

The Edit Audience section (Figure 6) is where the precision of the targeting is tuned. The ad can be based on attributes such as age, geographic location (down to a street), key interest words, gender, those who like your page (if there are any) and their friends. There are thousands of keywords to choose from, providing a method to even more narrowly refine the users who will be targeted.

The last step of the ad creation process is to choose how long the ad will run for. The creator can choose a date range or choose it to run continuously (Figure 7).

In the interest of research, we ran an ad on a page we created. The ad cost us about $15, reached approximately 1,400 people and garnered 66 engagements. In this particular instance, the cost of a view ended up being approximately 10 cents per person. The implications of this are far-reaching. For several thousand dollars and from anywhere in the world, ads can be used to drive traffic to an actor-run regionally-themed page after which

Figure 5: Facebook’s ad creation interface.

Figure 6: Facebook’s user-targeting interface.

Figure 7: Facebook’s ad budgeting interface.
content can be served on that page designed to influence an election in a targeted region (perhaps one usually decided by a couple of hundred votes).

Another popular platform for information sharing is Twitter. It also has a precise ad targeting system that leverages similar data as Facebook, such a geography, followers, and interests (Figure 8). Twitter’s and Facebook’s incentives for promoting social interactions overlap and their approaches to ads are similar. As you can see, Twitter’s ad creation platform allows for precise targeting on par with Facebook’s. In fact, both platforms’ interfaces for ad creation and monitoring bear quite a bit of resemblance. In an example pointed out in our previous paper on disinformation infrastructure, in a possible effort to gather data on users for the targeting phase of their interference campaign, UReputation posted conflicting content for multiple sides of the 2019 Nigerian election.

Social media has revolutionized the way we create and ingest ads. Today, people can create powerful, persuasive and targeted ads at a low cost. The time and resources it takes to set up an entity and start running disinformation campaigns on platforms such as Facebook and Twitter are small and takes very little planning and effort. Because of this, adversaries developing and carrying out disinformation campaigns will continue to rely on these types of platforms as a part of their disinformation operations strategy.

**PRACTICING INFORMATION HYGIENE**

Acknowledging that each of us is vulnerable to disinformation tactics means accepting a certain amount of responsibility when it comes to ingesting and amplifying the information we consume. That disinformation depends in part on the aided propagation of everyday people implies that we do have some influence over the success of campaigns, even if small. To that end, we can think of information hygiene as a set of practices that help to ensure we consume and share information responsibly.

It’s natural to want to share/like/retweet a headline when it resonates with us or makes us feel good, but this resonance may also make us less likely to critically evaluate the information we’re reading. A study finds that roughly 6 in 10 Americans only read headlines. In one Vanderbilt study, users were less likely to share a headline when they were asked to pause to consider why they believed a headline was true or false. This suggests that some of the habits responsible for helping disinformation to spread...
Information hygiene and the human levers of disinformation

may come down to our initial emotional response to the concise, but context-free bites a headline provides. And while the convenience of social media sharing functions like retweets make it easy to amplify information, it does not incentivize users to first spend time considering it. In service of this necessary survival skill, we’ve created a checklist to help users navigate the information vetting process before they decide to pass on what they read.

Images and videos may themselves not be inauthentic but may be included in posts unrelated to what the media in question is showing or accompanied by text that offers an inauthentic explanation. This is sometimes called “miscaptioned” media. A useful mechanism for spotting miscaptioned media is to use Google’s reverse image search on the image or some frames of a video to see if the source of the media matches the content. It’s also good to be aware of and avoid studies published in predatory journals with a reputation of publishing papers with little-to-no peer review.

In situations where the source can’t be traced, or the claim is apocryphal, fact-checking bodies can be a useful resource against which to test the veracity of circulated information. Sites such as factcheck.org and snopes.com tackle common public claims and rumors in an effort to help sort out fact from fiction. Twitter has introduced new misleading content labeling that fulfills a similar function. Getting ahead of disinformation campaign messaging by standing up early factual resources can be the key to helping to mitigate the spread of misinformation, and helping familiarize friends or loved ones with these resources early may help prevent disinformation from later coming in and filling the void. Disinformation actors, however, are aware of these resources, the threat they pose to their operations, and will go to great lengths to sow public distrust in

Follow this checklist before clicking "share"

We’ve outlined some steps below to consider before you choose to share a piece of information, link, news story, meme or anything else online.

1. Acknowledge we’re human and be wary of our own bad habits
2. Before sharing, ask yourself whether the information makes sense or just provokes an emotional reaction
3. Can I trust the source?
   a. Is the source traceable?
   b. Is the original information represented fairly, with proper context?
   c. What is the journalistic track record of the source?
4. Does the information reference a study?
   a. Are the conclusions of the study accurately represented?
   b. Do the visualizations originate from that study and have they been altered meaningfully?
   c. Was the study published in a predatory journal?
   d. Was the study commissioned by an organization with a known agenda?
5. Does the information include an image or video?
   a. Can the original context of the media be confirmed?
   b. Does it show up in a Google image reverse search?
6. Is the rhetorical technique identifiable and does this increase doubt about the intentions of the post?
7. Is there a claim in the information that can or has already been fact-checked?
8. Is there an open-source tool that can help?
9. What do I know about the account?
   a. Does the account have a history of altering information in any of the ways previously mentioned?
   b. Does the account have a history of pulling from questionable sources?
   c. Does the account seem focused on any particular messaging?
   d. Is there any account activity consistent with bot behavior?
Information hygiene and the human levers of disinformation

them. One example of this is the undermining of the research company Bellingcat. It is based in the United Kingdom and staffed by a mix of workers and volunteers that use open source intelligence to investigate major global events. After doing research into the assassination of former Russian military intelligence officer Sergei Skripal in London and the downing of flight MH17 over Ukraine, online trolls and social media personas accused Bellingcat of being a front for the CIA and/or MI6 (Figure 9), which may reduce the perception of research done there to those who distrust intelligence agencies.

There are a variety of published applications that can help users distinguish disinformation online, from web extension fact-checker to educational tools. Rand.org has published a list of many of those available online.

With the U.S. election on the horizon, disinformation campaigns aimed at the American voting population are in full swing and likely to continue to intensify. As it was in 2016, these campaigns are hoping to affect the integrity of the U.S. election system and Western democracy. With this in mind, politicians and the media have a shared responsibility to ensure no further damage is done to trust in the processes by calling out misinformation and refusing to participate in false narratives, regardless of the temporary political gain or increased viewership it may yield.

Voters, too, will need to be cautious, maintain a heightened awareness of, and be more critical than ever of the information they are consuming. This may require voters to more closely scrutinize news from online sources and the courage to challenge pre-existing intuition about what information appears to pass a preliminary truth test. Furthermore, voters should be careful about online attempts at voter suppression in the form of disinformation about the time, place, and manner of voting, which have become increasingly common in recent years.

HOW TO IDENTIFY BOTS

Another invaluable survival skill is the ability to tell human-operated accounts from bots. As outlined by our previous piece on disinformation infrastructure, bots are instrumental in amplifying the false news, facts, and narratives the disinformation actor is attempting to propagate.

Bots can be tricky to identify. It’s common for bot accounts to have generic names like “Gene Johnson” or “John

Figure 9: Online thread accusing Bellingcat of being a CIA front
Information hygiene and the human levers of disinformation

Smith." They may even have a convincing profile picture, either real ones taken from photos online or deep fake generator sites like thispersondoesnotexist.com.

Here are some features we found useful in distinguishing bots:

Account history: While having an established account history doesn’t guarantee an account isn’t a bot, because of their relative scarcity, it’s much more common to see bot-associated accounts that have been created much more recently, with fewer followers, and a shorter history of posts.

Time to respond: Humans can only respond to posts so quickly. Bots can consistently and immediately post replies without requiring the need to process and think over the information in real time before replying.

Ways in which platform is being accessed: It’s highly common for a human-run account to be accessed through multiple platforms: mobile, desktop, browser, client, etc. Scripted accounts may be connecting through an API or some unchanging interface.

SPEAKING WITH FRIENDS AND FAMILY

It’s often not empowering to hear that the best solution to a distributed problem hinges on the accountability of individuals, especially when you have friends or family who may or may not struggle with telling true and false information apart. However, when it comes to correction, one study finds that relying only on facts to correct erroneous beliefs isn’t always effective at helping people to update their understanding of a given matter. Rather, there is some evidence that people are better at correcting for their biases when they’re made aware of them.

As demonstrated by the continued influence effect, there may be situations where a loved one will continue to hold onto a belief seeded by misinformation even after being shown new information that debunks it. The goal in these situations should be to develop open-mindedness, which may require an approach guided by empathy.

False beliefs, such as those rooted in conspiracy theories, often develop as a response to a lack of control during uncertain and worrying times, and are especially likely to develop during disasters and tragedies. Experts suggest avoiding ridiculing and affirming critical thinking, redirecting it instead to the critical analysis of the conspiratorial belief.

Furthermore, there is evidence to suggest the more people can legitimize an opposing point, the more likely they are to process it with fairness. In service of this, one approach that may be worth considering is eschewing the language of the narratives and talking points that are prepackaged for circulation on social media in favor of using your own. These talking points often come with emotional baggage and trigger pretrained reactions that can preempt any meaningful dialogue. Rather, attempt to rephrase potentially controversial viewpoints in your own words, and attempt to keep the discussion sincere and compassionate. Finally, it’s crucial in our engagements with others that we hold ourselves open to being wrong, generous in our patience for differing voices, and critical of the ideas we agree with when they are represented in bad faith.

Unfortunately, not everyone will be interested in challenging their commonly held beliefs, checking facts, and corroborating information with additional sources. For those who are unlikely to make use of any of the advice laid out here, institutional changes, such as laws and policy changes, may be required to help limit the potential for damage on a macroscopic level.

MITIGATION AND POLICY

Fully countering disinformation at scale may seem unfeasible, but there have been discussions at the government level worldwide to attempt to adapt and respond to the problem. While this is no replacement for the action, everyone should take to increase their media literacy and information processing abilities, it may help to dissuade larger campaigns of foreign influence.

One thing to note as well as we discuss government intervention, is the intended or unintended consequences of individual content distribution. While those in the U.S. are protected by the First Amendment governing freedom of speech, platforms are increasingly taking more active roles in moderating content spread on their sites. Individual repercussions could include account removal, banishment from social media sites, and if legislation gets passed, potential passing of fines or litigation from social media platforms for individual users’ violation of policies on their websites.

Governments are combatting disinformation in several ways. The European Parliament and European Commission have set up pages and resources to fact-check false information narratives surrounding the COVID-19 crisis. Finland has partnered with social media influencers who help share official information on the pandemic, and the UK...
Information hygiene and the human levers of disinformation

and Poland have launched various apps or WhatsApp chatbots to provide legitimate information around COVID-19. Despite these efforts, which appear to have been successful in identifying legitimate content and checking facts, there are few proactive measures put in place to help curb disinformation narratives before they gain a foothold in mainstream media.

TALOS AND THE FAKE NEWS CHALLENGE

In 2017, Talos participated in the Fake News Challenge, a competition set up by academia and industry to foster the development of tools to help people identify disinformation in the news. Talos won the event, beating 80 teams worldwide. During the competition, Talos tested how various machine-learning techniques performed in identifying fake news on the internet. After successfully implementing several different models, the team found that their results were best when combining multiple models in an ensemble. The team made its final submission, a model based on 50/50 weighted average between decision trees and deep learning predictions, available on GitHub.

The deep learning approach applies several different neural networks to an article’s headline and body text. The researchers selected the architecture of this model because of its easy implementation and fast computation, but they noted that the model is limited in that it only gets to observe the text once. In the decision tree approach, some text-based features of an article’s body and headline are fed into “gradient-boosted trees” to predict the relation between the headline and the body. After exploring the dataset, Talos found several features that were suggestive of headline/body relationships, such as the number of overlapping words between the headline and body text and similarities measured between the word counts. The researchers selected this model because it is easy to scale and use with large data sets. While more research needs to be done, Talos’s research on stance detection is an important first step toward tackling the problem of fake news and disinformation in the 21st century.

CONCLUSION

At their core, disinformation campaigns take advantage of and rely on fundamental properties of human behavior to efficiently and effectively assimilate and retransmit campaign messaging. Furthermore, the coping strategies and filter mechanisms that emerge as people become increasingly inundated by the volume of information, they’re being exposed to daily make the process of distinguishing trustworthy, accurate information all the more challenging.

Let’s take the information hygiene practices discussed previously and apply them to the post in Figure 10. It is loaded with words designed to elicit an emotional response. This can fall into the drowning facts with emotion rhetorical technique. Looking closer, the post uses miscaptioned media, exaggeration and over-generalization to grab the reader’s attention. The meme also has a misleading title/message. The claim that “veterans are

Figure 10: An example of a phony post on social media that uses a mislabeled photo, misinformation and unconfirmed facts.
Information hygiene and the human levers of disinformation

still fighting for a job” does not appear to be sourced. Moreover, the text accompanying the image suggests unemployment figures that would actually undermine this meme’s message but the poster’s anecdotal experience supersedes it. Referring to factual data on the subject, it goes on to specifically state that “In my opinion all these numbers might be manipulated.” This can fall under the denying facts technique. Notice the number of users that interacted with the post. Do any of them indicate bot activity? Also, the post leverages an impressive number of hashtags that the platform uses as a part of its algorithm to spread the post to other participants. While the hashtags have an overarching military theme, some of them go beyond the subject of veterans. This could be another indicator that the intended meme has ulterior motive.

If we conduct a basic fact check by looking into the agency that tracks this information, we can in fact see that in 2019, the unemployment rate was 3.1 percent among veterans in the U.S., which was the lowest in 19 years. In fact, at the time the meme was made in 2015, the unemployment rate for veterans was about the same as it was for civilians.

The tactics leveraged in these campaigns play off human cognitive biases, manipulating pre-existing, deep-seated rifts about identity and political beliefs against us to further the goals and narratives of malicious actors who have little interest in seeing resolution to these conflicts. Furthermore, the engagement-increasing mechanisms around which social media platforms themselves are developed play a part in exacerbating polarization, isolating each other from different voices, and much more highly weighting those perspectives validated by social consensus.

The ways in which we’ve adapted to awareness of these tactics are likewise not always constructive. We’ve seen a response with an uneven suspicion for new sources of information in a way that biases our own preconceptions, shutting ourselves down to access to crucial legitimate sources of information that may help us root out and correct misinformation that may have slipped through. It’s this mistrust of traditional information sources that helps remove potential barriers to misinformation’s ability to spread, by turning the public away from valuable common reference points for discerning truth from fiction.

Thankfully, there are measures we can take to improve our information hygiene practices and help to mitigate our part in the spread of misinformation (and therefore be less complicit when disinformation campaigns strike). Approaching information consumption with little more intention and deliberateness can mean all the difference when it comes to how successfully a campaign propagates. Our relationship with our own subjectivity, and how successful we are at acknowledging and correcting for it, can have an important impact on how we integrate information that contradicts our understanding of the world and people around us.
Information hygiene and the human levers of disinformation

About Talos

The Talos Threat Intelligence Group is Cisco Security’s threat intelligence organization, an elite group of security experts devoted to providing superior protection for our customers, products, and services – as well as a vast collection of open source security products and tools. Talos is among the largest threat research teams in the world, encompassing seven key areas: Community & Open Source, Detection Research, Engineering & Development, Incident Response, Intelligence and Interdiction, Outreach, and Vulnerability Research & Discovery.

Talos detects and correlates threats as they happen, pushing coverage to customers globally within minutes to protect against known and emerging cyber security threats. With great visibility comes great responsibility – Talos also supports open-source security and often undertakes interdiction efforts to mitigate threats in the wild that pose significant risk to the internet at-large.

For more information, visit www.talosintelligence.com.