IDENTIFYING DIS/MISINFORMATION ON SOCIAL MEDIA

Purdue University Diplomacy Lab: Strategies for Identifying Mis/Disinformation

December 8, 2022

Acknowledgements

The Purdue University Diplomacy Lab: Identifying Dis/Misinformation project, with cohorts in Spring 2022 and Fall 2022, was a partnership between the U.S.Department of State and Purdue University.

We thank Professors **Bethany McGowan** and **Matthew Hannah** from the Purdue University Libraries and School of Information Studies, who guided this student-led project. We would also like to thank our State Department liaison, **Theresa Dixon**, Innovation Facilitation Officer at the State Department Operations Center, and the Watch Officers in the Operation Center for their gracious help and wealth of information.

This project was generously funded by the Joanne J. Troutner Innovative Educators Program.

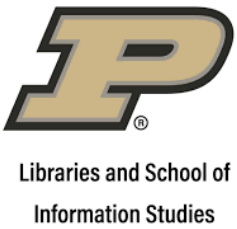
Four questions were central to our investigation:

1. What are the differences between misinformation and disinformation, and do these differences help identify one or the other?
2. What are the hallmarks of misinformation and disinformation?
3. Does the State Department have a proposed process for evaluating information found on open-source media?
4. What current and future tools could be used to support this proposed process?

To address these questions, we created:

1. Information literacy video tutorials that instruct on the differences and links between mis/disinformation;
2. A social listening dashboard that uses machine learning to identify the hallmarks of mis/disinformation;
3. Flowcharts for quickly determining if a social media post is mis/disinformation;
4. Checklists to evaluate information on open-source media;
5. A policy report that highlights the current and future tools, and primary mechanisms of mis/disinformation.

These deliverables are available at https://diplomacy-lab.lib.purdue.edu/.



Affiliated Students

Members of the **policy report drafting team** were:

- Sofia Babcock, *Political Science* (Spring 2022 cohort)
- Kate Biggs, *Biomedical Health Sciences* (Fall 2022 cohort)
- Lara Chuppe, *Computer Science* (Fall 2022 cohort)
- Christina Galiatsatos, *Political Science* (Spring 2022 cohort)
- Jannine Huby, *Political Science & Global Studies* (Fall 2022 cohort)
- Michael Kuczajda, *Psychological Sciences & Global Studies* (Fall 2022 cohort)
- Bennet Miller, *Law & Society* (Spring 2022 cohort)
- Stephanie Perun, *Political Science* (Spring 2022 cohort)
- Amanda Shie, *Political Science* (Spring 2022 cohort)
- Alicia Stevance, *Political Science* (Spring 2022 cohort)
- Andrew Yason, *Russian Language & Culture* (Spring 2022 cohort)
- Charlotte Yeung, *Political Science* (Fall 2022 cohort)

Members of the **information literacy video tutorial development team** were:

- Adam Munshi, *Health and Disease* (Fall 2022 cohort)
- Addie Powell, *Law and Society* (Fall 2022 cohort)
- Anushka Sharma, *Aerospace Engineering* (Fall 2022 cohort)
- Vinnie Vuskalns, *Professional Flight Technology & Aeronautical Engineering Technology* (Fall 2022 cohort)

Members of the **social listening dashboard development team** were:

- Supriya Dixit, *Data Science and Computer Science* (Spring 2022 cohort)
- Reece Fleck, *Computer Science and Political Science* (Spring 2022 cohort)
- Nicholas Gorki, *Computer Science* (Spring 2022 cohort)
- Keshav Iyengar, *Computer Science* (Spring 2022 cohort)
- Eric Liu, *Computer Science & Data Science* (Spring 2022 and Fall 2022 cohort)
- Gabe Mason, *Computer Science & Philosophy* (Fall 2022 cohort)
- Jeremy Rifkin, *Computer Science* (Spring 2022 cohort)
- Shelly Schwartz, *Data Science* (Spring 2022 cohort)
- Yaseen Shady, *Computer Science* (Spring 2022 cohort)

Members of the **flowchart and checklist design team** were:

- Olivia Anderson, *Cybersecurity* (Fall 2022 cohort)
- Becca Counen, *History & Political Science* (Fall 2022 cohort)
- Alison Hannon, *Nutrition Science* (Fall 2022 cohort)
- Ksenia Lewyckyj, *Economics* (Fall 2022 cohort)

Members of the **website design team** were:

- Phoenix Dimagiba, *Cybersecurity & Network Engineering Technology* (Fall 2022 cohort)
- Jake Valdez, *Aerospace Engineering* (Fall 2022 cohort)
- Sean Zak, *Chemical Engineering* (Fall 2022 cohort)

Identifying Dis/misinformation on Social Media

Dis/misinformation was a major concern in the 2016 US presidential election and has only grown worse in recent years. Even though dis/misinformation is often spread by domestic actors, actors abroad can use it to spread confusion and push their agenda to the detriment of American citizens. Even though this report focuses on actors outside the United States, the methods that they use are universal and can be adapted to work against domestic agents as well. A solid understanding of these methods is the first step in combating foreign dis/misinformation campaigns and creating a new information literacy paradigm.

This report highlights primary mechanisms of dis/misinformation: multimedia manipulation, bots, astroturfing, and trolling. These forms of dis/misinformation were selected after thorough research about common pathways dis/misinformation are spread online. Multimedia manipulation details image, video, and audio dis/misinformation in the form of deepfakes, memes, and out-of-context images. Bots are automated social media accounts that are not managed by humans and often contribute to dis/misinformation campaigns. Astroturfing and trolls use deception to sway media users to join false grassroots campaigns and utilize emotionally charged posts to provoke a response from users.

This policy report also specifically defines case studies of disinformation seen in China, Russia, and Iran, outlining common patterns of dis/misinformation specific to these countries. These patterns will allow for more accurate and quick identification of dis/misinformation from the outlined countries by State Department Watch Officers. Recommendations have also been provided for each type of disinformation and include a list of what individuals should look for and how to make sure that the information they receive is accurate and from a reputable source. The addendum at the end of the paper lists all of the recommendations in one place so that individuals do not have to search the paper for the recommendation they are looking for.

The intention of this report is to aid State Department Watch Officers as they work to accurately identify foreign developments, but researchers may also find this information useful in anticipating future developments in foreign dis/misinformation campaigns.

Contents

Introduction

The advent of the internet allowed for the sharing of information on a mass scale. Its impact has become more apparent with the growing use of social media as a legitimate source of news, global updates, and intelligence. While initially there were hopes that the internet would serve as a democratizing agent, events in recent years have highlighted the many challenges that accompany the benefits of the internet. The United States became acutely aware of these issues of online dis/misinformation following a growing conversation around the role of bots, trolls, and other online manipulation strategies during the 2016 presidential election.

In light of these recent events, it is more necessary than ever to recognize the signs of and discern information challenges in the media as state and non-state actors continue to misrepresent and obscure truth to divide public opinion, encourage chaos, and promote their own agendas. The introduction will differentiate between misinformation, disinformation, and malinformation. Additionally, it will discuss biases and factors of decision-making that impact individuals' abilities to accurately identify incorrect information on social media.

Information Challenges Defined

There are three key terms: misinformation, disinformation, and malinformation. Misinformation refers to false information presented as fact that was not shared with the intent to harm. For example, if someone retweets a news story not realizing the information is incorrect, this would be considered misinformation. In contrast, disinformation, is untrue, or semi-truthful content presented as fact, and spread with malicious intent. This could include spreading false information to manipulate people to support a certain viewpoint, or to incite chaos and discord. Finally, malinformation describes truthful content shared out of context with the intent of misleading or manipulating the audience (CISA, n.d.).

Disinformation campaigns are not a new concept, however. During the Cold War, the Soviet Union's premier intelligence agency, the KGB, launched a campaign known as Operation "Denver" (Selvage, 2020). They pushed the hypothesis that AIDS resulted from failed experiments by US government biological weapons labs, dismissing the correct claim that the HIV virus originated in Africa. In coordination with the intelligence agencies of other Soviet countries, such as the East German Ministry for State Security, the campaign shared forged documents and testimony of purported experts with the intent of undermining the United States' reputation on the world stage. For a complex set of political and social reasons, some world leaders and members of the public quickly believed in this conspiracy theory, further aiding in its spread across the globe (Selvage, 2020).

This campaign was so effective that a 2005 study published in the *Journal of Acquired Immune Deficiency Syndrome* found that 26.6% of the African Americans they surveyed supported the claim that AIDS was created in US government laboratories, and 53.4% of the respondents agreed with the statement that a cure exists for AIDS, but it is being with withheld from poor communities (Bogart & Thorburn, 2005). The distrust of government institutions that leads people to accept these conspiracy theories reflects the community trauma of decades of mistreatment of black Americans by the US government. Thus, when examining the public's support of suspect content and conspiracy theories, it is important to be aware of the social and political history and tensions that influence the public's orientations on these issues.

More recently, the US House of Representatives Permanent Select Committee on Intelligence identified over 36,000 Twitter bot accounts, and 450 Facebook pages linked to the Russian government that were involved in spreading information about the 2016 US presidential election (US House). This coordinated campaign attempted to portray Donald Trump favorably while casting Hillary Clinton in a negative light (Linvill et al., 2019). Russia also purchased over 3,000 Facebook ads, exploiting Facebook's advertising algorithms to target US users (Linvill et al., 2019). Disinformation in elections is not limited to the United States, though. Attempts to influence elections via social media have also been reported in South America, Europe, Africa, and Asia (Downing & Ahmed, 2019; Ferreira, 2022; Ndlela & Mano, 2020; Neyazi, 2020; Uyheng & Carley, et al., 2020).

Many states have also used online tools to suppress dissent. In the People's Republic of China, the government uses its control of digital media to suppress ideas in opposition to their agenda and to influence countries and individuals' stances on Taiwanese and Hong Kong's independence. For instance, during the Hong Kong protests against China's controversial Extradition Bill, government-sponsored media used Twitter to discredit protesters and increase support for police forces (Wood et al., 2019).

If successful, deception can erode the public's faith in the media and trust in the government, divide the public, and rewrite history. As the political landscape becomes more polarized and social issues worsen, it becomes increasingly easy for foreign governments to manipulate social media users.

News Outlets' Role in Deceit

When assessing the potential role of news outlets in sharing dis/mis/malinformation it is important to distinguish between state-controlled media and privately-owned media. State-controlled media companies do not have independent editorial control over stories they release and often share information portraying the state in a favorable light and the opposition negatively, regardless of its accuracy. China Central Television is an example of state-controlled media (Bleck & Michelitch, 2017). An important distinction is that just because a media company is state-funded does not mean that it is controlled by its respective government. For example, the BBC is funded by the British government but has independent control over its editorial guidelines and so, therefore, is less prone to pushing deceitful information (*BBC*, n.d.-a, n.d.-b). Sometimes it can be difficult to distinguish state-sponsored publications from private ones, however, as they can appear privately owned to an untrained eye. In many case studies, state-run media is the main source of information in a given country (Walker & Orttung, 2014). This may create situations where citizens accidentally propagate false information given to them by the government, online.

Even when an outlet is privately owned, it is difficult to fully banish misinformation. In a Harvard study of the 2016 election, researchers found that traditional news sources were a greater source of dis/misinformation than Russian-operated/owned ones. Furthermore, they noted that Russian disinformation campaigns would not be successful without their being reported on and amplified by general news sources (Farris et al., 267-268). News outlets often must rely on sensationalist stories to attract viewers and profit: Donald Trump's Tweets were a prime example of this issue. Very few Americans learned of Trump's tweets through his Twitter feed. Instead, news outlets were most people's first exposure to his Tweets, which was an issue because nearly two-thirds of the news outlets did not include a disclaimer that his claims were false (Patterson, 2020). Even routine journalism practices can accidentally deceive audiences, allowing misinformation to spread.

Decision-making: Biases & Rationality

Heuristics and biases can impact humans' ability to make rational decisions. Fortunately, decision-making is a learned skill and can be improved with experience and practice. In the subsections that follow, rational decision-making is defined and several examples of processes that interfere with making reasoned judgments are presented. It is important to understand the influences these biases and heuristics have as it allows for better decision-making, including when assessing the accuracy of sources of information.

Rational Decision-making

Rational decision-making is when one makes decisions based off reason or logic. There are seven steps involved in making reasoned decisions (Uzonwanne, 2016).

1. Identify the problem – This involves clearly defining the context and scope of the problem as it is difficult to address a problem if it is not well-understood. For Watch Officers, the challenge is already fairly well understood: identifying dis/misinformation on social media. Any gaps in the understanding of the problem can be addressed through training or other education strategies.

2. Identify what a solution looks like – It is important to know when a solution is reached by clearly defining the characteristics of a solution. For Watch Officers, the solution is generally a decision on whether a post should be trusted and amplified to higher departments.

3. Conduct gap analysis – This involves identifying the space between the problem and a solution. This gap is often similar across many different posts Watch Officers might be analyzing. Generally, when first looking at a post, a reader is missing information about the profile of the user, their activity on the platform, the activity surrounding the topic on the platform, and potentially background information on the event or conflict.

4. Gather facts, options, and alternatives – It is necessary to conduct background research on the topic to be informed when making decisions, and to understand the different stakeholders in the decision. In a context requiring rapid decision-making, online tools can expedite the process of assessing tweets' validity. Further, developing a working knowledge of major conflicts and political tensions worldwide can make it simpler to gauge what dynamics might be at play in a post or claim of a new event.

5. Analyze option outcomes – It is essential to consider the costs and benefits of each possible outcome because that factors into choosing the best decision. When deciding whether posts online are correct, this step involves considering the outcomes of amplifying or choosing not to amplify the message. It is important to consider negative effects of both options, as well as the magnitude of those harms.

6. Select best possible options – After reviewing possible outcomes, select the outcome that minimizes costs and optimizes benefits.

7. Implement decision and evaluate - Once making the decision, it is important to evaluate whether it was the correct decision and reflect on how that could impact similar future decisions (Uzonwanne, 2016). Watch Officers could achieve this through creating a classified set of data of past decisions and whether they were accurate. This could be helpful for identifying what misled

Officers in past incorrect decisions as well as for curating a dataset to train machine learning algorithms to automate some of these decisions.

Rational decision-making may seem like a long process, but individuals usually run through these steps in their heads quickly. The importance of this model is that it demonstrates how decision-making is a learned skill and can be improved. Watch Officers cannot make intuitive decisions, which are influenced by irrational influences. Some common irrational decision-making influences to be aware of are habits, conformity, and cultural bias (Tsohou et al., 2015). Habits can be harmful because the fact that a habit feels familiar does not always indicate that it is right. Conformity is when one chooses what they think other people may choose. One may have overheard a co-worker or someone in a higher position than oneself speak on a topic, and when they subsequently see an article, they may be more apt to believe it (Tsohou et al., 2015).

Lastly, cultural bias and religious preferences come into play when individuals ignore their influence over mental processes. When a decision-maker is choosing an article specifically because they have an unrecognized bias, it is likely that the article is deceptive.

Heuristics and Biases

Biases allow individuals to make inferences about underlying processes, and heuristics serve as simple judgment rules that expedite or shorten the decision-making process. When applied inappropriately or excessively, however, it can lead to errors in the resulting judgments (Adame, 2016). Recognizing that there are several common biases and heuristics that influence decision-making allows for more logical decision-making less affected by bias. Below are listed several common heuristics in the judgment and decision-making process with definitions and potential strategies for avoiding using them when making crucial decisions.

The Anchoring Effect

The anchoring effect occurs when an individual biases future estimates of a numerical value to fall closer than random to a previous value or estimate they had read or been told, even if that value was incorrect (Adame, 2016). This numerical value or 'anchor' serves as a starting point. These anchors can be entirely arbitrary and can come from knowledge and experience, attitudes and preferences, other people, or from inferences based on certain proximity cues (Doherty & Carroll, 2020). An example of a negative use of this heuristic would be if social media users are exposed to false statistics on a controversial topic or from a public opinion poll online. If asked to make their own estimates of these statistics, they will likely anchor those estimates on recollections of the false statistics they have seen online, even if subconsciously. Researchers have identified that encouraging individuals to create self-generated reasons anchoring values could be false or presenting them with a list of reasons to choose from is effective in helping to mitigate the anchoring effect (Adame, 2016).

The Availability Heuristic

The availability heuristic is the technique used when individuals base their judgment of what is most likely based on how easily they can recall an example. More frequent events are easier to recall and imagine than infrequent ones. Media coverage and reporting can distort the perceived frequency of different events. For example, people's fear of a particular type of catastrophe, especially when disproportionate to the actual risk, likely reflects the media's level of coverage of that type of event (Doherty & Carroll, 2020; Sunstein, 2006). This can in turn impact individuals' ability to assess the

accuracy of new information, as they base those judgments off their ability to recall examples of similar cases in the past.

Confirmation Bias

Confirmation bias refers to people's general receptiveness to new information that supports their beliefs and past judgment versus their reluctance toward information that contradicts those convictions. While often discussed in the context of ideology, confirmation bias can also impact judgments on information related to beliefs held about the character or typical behavior patterns of another state or world leader. Furthermore, a recent study indicates that in conjunction with reluctance about accepting contradictory information, the confirmation bias tends to lead people to discount opposing opinions even if the strength of the information backing the opinion emphasizes its credibility (Kappes et al., 2020). This highlights the importance of individuals' objectively analyzing the credibility of sources and the value of identifying their opinions about social media posts and other media to better understand how confirmation bias could appear in their judgments.

State Actors

Different states use different methods to spread disinformation, and they benefit politically, socially, and financially from sowing dis/misinformation in the international sphere. This section will focus on China, Russia, and Iran, and their slightly different goals and tactics. China, like Iran, is focused on improving its image internationally and manipulating public opinion on domestic political conflicts whereas Russia aims to flood the internet with dis/misinformation, focusing on eroding truth to create distrust on a global scale. China primarily employs astroturfing and trolling, although current efforts point towards the use of nationalist hackers in the future (Harold et al., 2021). In contrast, Russia focuses on flooding social media with accounts, catfishing, and paying hackers to do their work (Treyger et al., 2022). Iran uses proxy sites and a vast network of Facebook and Instagram pages to disseminate state propaganda as well as catfish western journalists (Brooking & Kianpour, 2020). All three are capable and have communicated dis/misinformation in multiple languages. They generally focus more on open platforms such as Facebook and Twitter than closed messaging platforms like Telegram and WhatsApp. Further, China and Russia have both noted the US intelligence's reliance on social media to get information (Harold et al., 2021; Treyger et al., 2022).

China's Approach to Social Media

From Chairman Mao's slogans to Deng Xiaoping's posters, Chinese governments have a long history of using propaganda to promote political agendas at home and abroad. The use of Twitter and other social media platforms is merely a new form of this historic practice. The Chinese Communist Party (CCP) focuses on monitoring its domestic audience as well as the international audience. Though China has banned Twitter, Meta, WhatsApp, Google, and other commonly used social media platforms, it does use these platforms to spread misinformation to users abroad. NPR reported that there were several bids by Chinese government agencies to purchase followers on Twitter and Facebook (Wood et al., 2019). For example, Chinese state-run outlet China News offered 1.25 million yuan ($176,900) to acquire more Twitter followers ("zhōng guó xīn wén shè zhōng xīn shè [China News Agency]", 2019).

Social Media Tactics

China uses a combination of astroturfing, discrediting enemy leadership, and trolling on social media (Harold et al., 2021). In the past, China has not used many bots, unlike Russia. However, in the last

couple of years, that has started to change. The CCP also generally prefers open platforms such as Facebook over closed platforms like WhatsApp because the open, public platforms make it easier to obtain data regarding the effectiveness of disinformation campaigns. Additionally, disinformation can be disseminated much faster on these platforms than it can on closed social media. Finally, China spreads disinformation in multiple languages (Wood et al., 2019). For example, Twitter and Facebook both announced that Chinese-backed social media accounts released posts written in English targeting the Hong Kong protests.

Concerningly, Chinese military authors have recognized the United States' reliance on open-source information for intelligence operations, such as the State Department's Operation Center's monitoring of social media. An article by the Rand Corporation identified catfishing, "the use of fake identities designed to lure people into the mistaken belief that they have developed an online relationship (romantic or professional)", as a mechanism the Chinese government hopes to use to coax those with access into divulging classified material (Harold et al., 2021). Chinese military authors also attempt to amplify the voices of people who echo their beliefs, including retweeting celebrities and hacking their accounts (Harold et al., 2021). China is also open to outsourcing social media messaging and disinformation. For instance, after Tsai Ing-wen was elected president of Taiwan, thousands of Chinese users accessed Facebook to complain on Ing-Wen's Facebook page. Referred to as the 'Diba Expedition', Chinese military authors pointed to this as being tacitly approved by the CCP indicating that this behavior will become normalized over time (Harold et al., 2021).

Recommendations

- Raise awareness of China's dis/misinformation tactics.
- American agencies should establish a trusted online presence on domestic media and possibly even create one on Chinese platforms.
- Reach out and build trust with Chinese Americans, Taiwanese Americans, and Hong Kong Americans.
- Build a database of examples of Chinese dis/misinformation on social media to reference when making judgements on social media posts.

Russia's Approach to Social Media

Russia has viewed social media as a tool of the West that they can use to spread their beliefs and wreak havoc (Treyger et al., 2022). They were particularly interested in the power of social media after the color revolutions of former Soviet states and the social media-coordinated Arab Spring uprisings. Rand corporation suggests that Russia has embraced social media due to its many benefits: it is low-cost, can reach large audiences, and makes it challenging to trace online behavior back to a state-sponsored campaign. On social media, Russia focuses on stoking divisions within Western states, promoting Russian foreign policy narratives, supporting military action involving active warfare, and targeting elections and political/social circumstances in certain countries.

Social Media Tactics

The Rand Corporation calls Russia's disinformation model the "Firehose of Falsehood" because of the "high numbers of channels and messages and a shameless willingness to disseminate partial truths or outright fictions" (Paul & Matthews, 2016). Messages are continuous, repetitive, and inconsistent, with

accounts sometimes contradicting each other. Russia's disinformation outside of the former Soviet Union states increased after 2014, the year of Russia's annexation of Crimea. Despite this cited activity, some assert that Russia's disinformation campaigns are poorly organized and funded, and that impacts they have on Western countries remains uncertain.

These information efforts are contributed to by a mixture of state, state-affiliated, and non-state actors, ranging from YouTube channels like Russia Today (RT) to hackers supporting the Russian government (Treyger et al., 2022). This is advantageous for the Russian government as it allows Vladimir Putin to blame online activity on Russians acting to support the government, rather than blaming it on the online activity of the Kremlin itself.

State-affiliated actors, like the YouTube channels RT and Sputnik, are protected under the First Amendment of the US Constitution, as they are news organizations. Thus, they cannot be treated as hackers. Troll farms are the other kind of state actor. Like China, it seems that Russian leadership pays bloggers and others for posting pro-Kremlin pieces. Russia has also begun to outsource their work to freelance hackers, which is relatively cost effective. As an example, a group of researchers once hired Russian cyberhackers to attack a fake website. In two weeks, they created 730 posts from 25 different Twitter accounts and generated 100 posts on different forums and blogs, and the researchers only had to pay $250 (Greenberg, 2019).

Russia employs the vilification of enemy leadership, multiple platforms, astroturfing, troll farms, and catfishing specifically to learn personal details of a target and runs the campaigns in multiple languages (Treyger et al., 2022). Russia will also target and harass critics and people who expose misdeeds. This can include targeting officials and celebrities and leaking personal information, sometimes in altered form. Russia uses a variety of platforms and blogs, including Facebook, Twitter, YouTube, Instagram, 4chan, Pinterest, Reddit, Tumblr, LiveJournal, and 9GAG. They have also looked into encrypted platforms such as WhatsApp and Telegram. They often set up fake accounts with stock images for profile photos that sometimes have extensive histories. A DFR lab report found that Russian operatives were behind dozens of fake accounts on 30 platforms in 9 languages (Nimmo et al., 2019). They also target specific users using Facebook ads.

Russia will amplify native content on their news sites as well as try to organize rallies. Regarding rallies, they will try to pit opposing sides against each other such as having suspected Russian actors pit anti-fascist demonstrators against Germany's far right movement in Berlin during the 2019 European Parliament elections (Apuzzo & Satariano, 2019).

Recommendations

- Highlight the ways in which Russian actors try to manipulate people rather than individually refute everything they create (that would be exhausting and perhaps impossible when they create so much conflicting information).
- Fine or sanction Russian news sources that spread Kremlin ideals such as RT.
- Look at the objectives of a specific campaign and, rather than refute the propaganda, find other ways to spoil their objective.
- Make the public and government officials media literate in the complex tools these actors use on social media (ex: astroturfing and catfishing).

Iran's Approach to Social Media

Iran embraced the internet early on. Its universities had an online presence by 1993, and it was the second Middle Eastern country to embrace the internet (behind Israel). Iran differs from Russia and China in that it has very consistent disinformation campaigns that have gone on for years. According to the Atlantic Council, it was as early as 2009 that Ayatollah Khamenei stated that "content promotion" was "the most effective international weapon" against foreign adversaries (Brooking & Kianpour, 2020). It seems that the IRIB, the office of the supreme leader, the intelligence services, the IRGC and associated militias, and the regular Iranian military each employ their own Internet operatives though it is unclear if they collaborate. The Atlantic Council believes that due to US sanctions and increasing scrutiny of its public entities, Iran will continue to lean into clandestine digital activities.

Social Media Tactics

Iran uses proxy sites, fake accounts, astroturfing, troll farms, a network of content-focused Facebook and Instagram pages, and catfishing (Brooking & Kianpour, 2020). The proxy sites regurgitate state media, with many proxy sites reporting on each other's news. These sites can be identified because they have no ads, unlike commercial information mills. Some versions try to make it seem like they are the official sites of foreign media outlets. Sometimes, this is done through something as straightforward as a misleading domain name (ex. "tel-avivtimes.com") and other times it involves plagiarism or clever misspellings. Fake social media accounts are interested in promoting and leading normal social media users to their proxy sites; however, generally these fake accounts do not appear particularly convincing. Content-focused pages have the similar goals as fake accounts but also aim to promote Iranian propaganda. The Atlantic Council emphasized Iran's reach with these accounts, noting that "through Facebook especially, Iran has built hundreds of region-specific pages that have reached millions of users in every corner of the world". Some more sophisticated groups will even target specific Western journalists via catfishing (Brooking & Kianpour, 2020).

Solutions

- Invest resources in identifying and neutralizing (but not sensationalizing) Iran's digital influence networks.
- Use automated search and text-matching of Iranian state propaganda products. If the State Department chooses a machine learning solution, they must create a declassified data set for American developers to train AI, but they must not disclose this data to the foreign adversaries behind the disinformation operations.
- Address regular press briefings to the Iranian people. The attacks by Iran work best when they go unanswered.
- Work with social media companies to find and shut down foreign influence operations because a lot of the work has currently fallen on private enterprises.

Types of Disinformation

Multimedia Manipulation (Deepfakes, Memes, and Out-of-Context Images)

With recent advances in artificial intelligence, a new wave of multimedia manipulation has begun in the form of deepfakes, memes, and out-of-context images. Deepfakes include algorithm-edited photos, videos, and audio, that generally depict a person saying or doing something that did not actually happen. They have the "potential to rapidly spread false words and actions to a global audience and can be extremely difficult to distinguish from real content" (Lam, 2021). While deepfakes can be challenging to identify without extensive knowledge of artificial intelligence because they appear so realistic, the general population has been able to recognize some deepfakes produced by Russia, China, and Iran. Deepfakes often promote emotions and beliefs already held by the intended audience, reaffirming those convictions. Despite the potential effectiveness of these methods, the impact they have on spreading dis/misinformation is unclear, as researchers claim "[deepfakes have not] yet shaped major world events" (Yankoski et al., 2021).

Memes are another type of manipulated media known as "shallow fakes". A shallow fake meme is an image (or occasionally a video) that is repurposed from its original use to make a humorous statement. Memes typically contain a short amount of text to alter the actual meaning of the image and usually reference current events. While deepfakes are generated by artificial intelligence technology and aim to look realistic, shallow fakes can be manually manipulated, and the quality can greatly vary (Yankoski et al., 2021). The minimal skills required to create memes make them accessible to most, meaning in contrast to deepfakes, memes are produced at a quicker rate (Yankoski et al., 2021).

Older photos used out of context are a form of malinformation that are used to deceive viewers and promote a false narrative. In contrast with memes, out-of-context images do not have humorous intentions and are not attempting to make a statement about a social issue. Instead, these photos from past events are stolen and reassigned with incorrect captions to mislead audiences into believing a new event has occurred. These images are harmful because people are more likely to trust the veracity of a news story or social media post if the false claim is paired with photos confirming the alleged event, even if, unknown to the audience, the images are from a different, past event (Fazio, 2020).

There are rising concerns about all forms of multimedia manipulation as technology advances since "they challenge real footage and undermine the credibility of civic media and frontline witnessing" (Gregory, 2022). Furthermore, when constantly confronted with manipulated content, social media users could start to question all information sources they encounter.

Case Study: China

The majority of deepfakes observed in Chinese media are imperfectly fabricated videos, but nonetheless, have similar effects to well-created deepfakes. This is because the messages of these videos still aid deceptive media users in spreading false information in support of their agendas. For example, in 2018, Chinese deepfakes covered the internet as the development of new apps for creating AI-manipulated videos provided increased accessibility to users who have minimal AI knowledge (Seta, 2021). This caused deepfake videos to surface on social media sites such as WeChat that implied the occurrence of events that never happened. This included a video of President Donald Trump and Secretary of State Mike Pompeo singing 'Wo Ai Ni Zhongguo' [我爱你中国 'I Love You, China'], a patriotic Chinese song (Seta, 2021). While this video was widely accepted as fake news, there are still reasons for concern.

Additional stories or media posts based on the video could lead to an even wider spread of this manufactured content where the original deepfake video is not linked, preventing users from assessing the quality and accuracy of the video. Also, while the public generally accepted this video as fake, this does not mean future, higher quality deepfake videos could not more effectively trick or polarize online users.

Case Study: Russia

Russia has also used deepfakes to spread disinformation despite the often amateur quality and flaws apparent to those even untrained in artificial intelligence. In March 2022, during the Russian-Ukraine War, Russia posted a deepfake portraying Volodymyr Zelensky telling Ukrainians to "put down their weapons" on a hacked news website (Broinowski, 2022). While the deepfake was easy to categorize as disinformation due to its poor quality, "digital forensics expert Hany Farid described it as 'the tip of the iceberg,' in a global information war increasingly characterized by the use of deepfakes to spread military propaganda" (Broinowski, 2022). Further, in 2021, UK and EU officials conducted several video calls with Leonid Volkov, Alexei Navalny's chief-of-staff, only to realize later that they had actually been discussing the sensitive information with a deepfake of the man engineered by the Russian using AI simulation (Broinowski, 2022).

The use of memes in Russia peaked during the COVID-19 pandemic and often used images and concepts related to the Soviet Union in the 1990s (Borenstein, 2022). Some memes took 1990s movie quotes and altered them to fit situations related to the pandemic (Borenstein, 2022). Other memes were created from fake news stories, further amplifying that disinformation (Borenstein, 2022). For example, fake news about wildlife returning due to lock-ins resulted in the production of memes quoting, "30 days of quarantine in Italy: the dolphins are back. 30 days of quarantine in Wales: the wild goats are back. 30 days in quarantine in Russia: the 90s are back" (Borenstein, 2022). While the memes intended to be humorous, they also have the capability to amplify fake news narratives. Fake news about the return of wildlife might be a more innocuous example, but these patterns point to future, more serious possibilities of disinformation coming from Russia.

Case Study: Iran

In July of 2018, an Iranian Facebook account posted a doctored image of Tom Hanks wearing a shirt with phrases including "Science is Real", "Black Lives Matter", and "No Human is Illegal" in order to gain followers on a propaganda account called "No racism no war" (Madrigal, 2018). According to Facebook, this account posted about "politically charged topics such as race relations, opposition to the President, and immigration" which ultimately led to its removal from the platform, but not before it was able to acquire 400,000 likes (Madrigal, 2018).

Case Study: Other Examples to Consider

The use of manipulated media is not limited to China, Russia, and Iran, however. The Flemish Socialist Party and an alt-right Israeli group have utilized notable deepfake campaigns to polarize discussions on political issues (Broinowski, 2022). For example, in 2018, the Flemish Socialist Party promoted a deepfake of Trump encouraging Belgians to no longer side with the Paris Climate agreement. Some members of the party believed the video, despite its technical flaws and the Flemish Socialist Party's use of the video to incite debate and further political tensions (Broinowski, 2022). Additionally, an alt-right Israeli group produced deepfake videos utilizing "'sock puppets' (synthetic humans generated from deepfake photographs)" depicting previous supporters of the political left expressing their change of heart on political discord and sharing that they now supported Prime Minister Benjamin Netanyahu of the

political right (Broinowski, 2022). These videos were posted to Facebook on a heavily conservative account called Zionist Spring (Broinowski, 2022).

Recommendations

- Learn how to detect and practice (by clicking the "PLAY" button) identifying false and edited images with resources such as https://www.whichfaceisreal.com/learn.html.
- Use reverse image searches to check for out-of-context images or memes using tools such as https://tineye.com/.
- Consider that sometimes the most harmful deepfakes and memes are not the highest quality, but are simply the most popular.
- Pay attention to the quality of videos (large pixels, poor audio, etc.), if the actions or words said seem out of place, and if there are any inconsistencies.
- Consider the account posting the deepfake or meme to understand if it is part of a political media campaign. Consider user following, follower interactions, what kinds of accounts are posting similar information.

Bots

Social media accounts classified as bots are fully automated, and do not require human intervention (Nimmo, 2018). This is in contrast with cyborgs which involve a combination of pre-programmed functionality and human activity (Martini et al., 2021). Bots can have a variety of innocuous functionalities, including posting weather alerts and news stories; however, the focus of this discussion is bots with malicious intent (Nimmo, 2018). Furthermore, many large-scale disinformation campaigns use bot swarms or botnets, large groups of bot accounts that all perform similar tasks (Jones, 2019); (Nimmo, 2018). Individuals or groups responsible for disinformation campaigns can design these bots themselves, or they can purchase "commercial" bots for rent to gain retweets, likes, or follows (Nimmo, 2018).

The rising presence of bots present many challenges regarding the spread of information on social media platforms. One of the most prevalent functions of malicious social bots is the retweeting of posts based on hashtags or other metrics. This functionality is much easier to program than other abilities that more closely mimic human behavior (Martini et al., 2021). Even small percentages of bots are sufficient to cause shifts in opinion on social media. For example, only 0.5% of Nicolás Maduro's followers were bots, but when Twitter banned them from their platform, retweets of Maduro's posts decreased by 81% (Martini et al., 2021). This ability of bots to artificially inflate the popularity of hashtags on Twitter can give the false impression of public support for certain positions or ideas (Jones, 2019; Stieglitz et al., 2017). Because it is difficult to design programs capable of replicating human writing patterns, bots often do not generate their own tweets, so a more common mechanism employed by bots is making posts with many hashtags and/or links to other sources created by a human (Jones, 2019). This makes it less obvious that a bot made the post, and also, it serves as another mechanism to create fake trends on Twitter not actually reflective of trends in public opinion. Not only do these two techniques create artificial trends, these trends and masses of support potentially obscure actual trends on Twitter, or they can create such spam associated with a hashtag that the hashtag is no longer usable (Jones, 2019).

Unfortunately, as the field of artificial intelligence advances, bots become even harder to detect. According to a study published in 2021, humans retweet bot and non-bot posts at the same rate, suggesting that even now, users do not distinguish significantly between automated and manual accounts (Martini et al., 2021) . The structure and policies of social media platforms further complicate the ability of researchers to gather information about bots and technologies to detect and combat them. Often, Twitter does not release data on statistics regarding bot accounts, making it difficult to curate datasets to use in training algorithms. Also, techniques of bots, social media platforms, and trending topics constantly evolve, so datasets quickly become outdated, and algorithms need to be re-trained (Martini et al., 2021).

Case Study: Bots in the Russia-Ukraine conflict

In 2022, Smart, et al., completed a study on bot and non-bot identities of tweets on the invasion of Ukraine by Russia from February 23 to March 8, 2022. Researchers identified pro-Ukraine and pro-Russia accounts through searching hashtags on Twitter. They classified 90.16% of these accounts as pro-Ukraine and only 6.80% as pro-Russia, but Russia has a much longer history of using bots. They found that Russian non-bot accounts have the highest information flows to other groups, and then bots are used to amplify the messaging coming from non-bot accounts. For example, following Russia's capture of Kherson, an increase in #(I)StandWithPutin correlated with an increase in bot activity on the Twitter platform, suggesting that this trend originated at least partially by artificial means. Further, while Russian accounts have significant out-group activity, Pro-Ukrainian groups have high in-group flow and minimal out-group flow, indicating Ukrainian accounts infiltrate discussions of their less than Russian accounts do. Finally, researchers found that self-declared bots increase discussions on many topics and have an especially strong influence on discussions surrounding politics and the government. As these accounts were mostly in English, the researchers speculated that they could be targeting users from Western countries with these tweets (Smart, 2022).

This research is interesting as it points to the different roles bot and non-bot accounts can play in influencing online discussions. Automated and non-automated accounts can work in tandem to amplify messages through in-group and out-group interactions. It also points to the need to consider audience when analyzing the impacts and credibility of tweets. For example, language can give clues to what the intent behind tweets and accounts might be, giving social media users clues to their credibility and how to interpret the information being shared.

Case Study: Bots in the Gulf Crisis of 2018

Another conflict in which bots had a central role was a dispute between Qatar, and the Quartet: the United Arab Emirates, Egypt, Saudi Arabia, and Bahrain. The conflict erupted when Qatar's head of state released a tweet in support of Iran in the spring of 2017. This angered the Quartet as they felt it violated the foreign policy objectives of the Gulf Cooperation Council, but Qatar alleged someone hacked the emir's account, and that the emir himself had not written these comments (Jones, 2019).

For the five years leading up to this conflict, about 318 accounts were created on Twitter per month, but in May 2017, this number jumped to 3,347. Furthermore, many of these accounts had similar metadata: they were all created on the web and had similar profiles, including poor-quality images for their profile

pictures (Jones, 2019). There are suspicions that these bot swarms had been set up in advance or that Saudi Arabia hired commercial bots to amplify their messaging (Jones, 2019; Nimmo, 2018). The fact that General Supervisor al-Qahtani of Saudi Arabia published many of the hashtags that went viral in support of the Quartet's campaign further implicates Saudi Arabia in using social media to launch a coordinated campaign against Qatar. For example, of the 2,116 retweets Turki Al-Shekh received for an anti-Qatar tweet, 1,600 were attributed to bots (Jones, 2019). These amplification methods had far-reaching implications as even BBC Arabic picked up on these trending hashtags and reported on these false impressions of popularity in mainstream news (Jones, 2019). This emphasizes the large-scale impact Twitter bots can have on shaping media coverage. The engineers behind the bots also used other techniques to give the false impression of genuine public backlash, such as setting the location of accounts to Qatar to make the users seem like citizens of Qatar (Jones, 2019; Nimmo, 2018).

Although Saudi Arabia and the Quartet likely heavily manipulated social media to their advantage, Qatar employed social media to promote their own agenda as well. Pro-Qatar tweets often received hundreds of retweets in just two seconds, and researchers traced these to commercial bot accounts rather than real people (Nimmo, 2018). This case study points to further techniques for identifying bots spreading misinformation. First, many of the accounts purchased by both sides had repetitive Twitter handles and grainy or stock images for photos, and had previously posted on non-political topics (Nimmo, 2018). Additionally, those setting up accounts often chose locations for the profiles to give the false impression of a specific nation's support of a particular ideology or claim, cautioning those assessing the veracity of tweets to rely on metrics such as location when making quick judgments (Jones, 2019). This research on social media's role in the crisis also suggests that rapid spikes in retweets or account creation potentially indicate bot involvement. Finally, the role of BBC Arabic spreading the false trends on social media cautions that even mainstream media organizations can misinterpret Twitter trends and contribute to the spread of dis/misinformation.

Solutions

Distinguishing bots from humans

While humans cannot surpass algorithms in their ability to distinguish bots from non-bots, some simple metrics can prove useful when making quick judgements (Martini et al., 2021). While bots do attempt to imitate human behavior, they tend to include more links in their tweets and focus more on retweeting content compared to human users (Stieglitz et al., 2017). However, programmers adjust other aspects of bots' behavior to mimic that of humans. For example, some bots take breaks to imitate sleep, and some bots can slightly modify the text of reposts, so as not to trigger Twitter's bot detectors (Stieglitz et al., 2017). Lastly, bot accounts often post with much greater frequency than non-bot accounts. While some debate exists on what cutoff to use, some researchers suggest that users posting more than 144 times a day are likely bots (Nimmo, 2018). Single-indicator methods, such as number of daily posts, do have limitations, however. Most single-indicators imprecisely partition data, and they will exclude bot accounts with low activity. Furthermore, they are unlikely to surpass the performance of algorithms using the indicator as a parameter (Martini et al., 2021).

Technology for detecting bots and its pitfalls

Given the challenges with detecting bots manually, automated tools are a promising option, but they come with their own limitations. Martini et al. published a study in 2021, comparing two prominent bot detection tools: Botometer and Tweetbotornot. Many well-known organizations, such as PEW Research Center, use Botometer, and the programmers who designed the tool used a diverse training set with many types of bots. When tested, Botometer produced both high false positive and high false negative rates. Since then, the creators of Botometer have modified the algorithm to predict specific types of bots rather than the general labels of bot or non-bot, improving the accuracy as the algorithm no longer has to generalize to patterns only seen across all types of bots. In contrast, Tweetbotornot mostly focuses on identifying bots involved in disinformation campaigns within the United States, limiting its usefulness for detecting bots in other contexts. When Martini et al. ran these algorithms on the same dataset, they found that there was minimal overlap in the bots detected by these two algorithms, emphasizing the need for caution when interpreting results from such online tools (Martini et al., 2021).

As discussed previously, one principal limitation in designing such tools is the availability of relevant data to train the algorithms. Algorithms make classifications based on what they have already seen, so if the training data does not represent the test data, the predictive model will have low accuracy. For example, Tweetbotornot likely would perform poorly when detecting bots not designed to spread disinformation domestically, and Botometer would struggle to detect bots not falling under one of the six bot types currently identified. Additionally, as bots continually evolve, the training datasets and thus the algorithms quickly become outdated. Genetic algorithms, designed to mimic evolution, show promise in detecting evolving bots and compensating for some of the limitations in available training data (Schuchard, 2019).

For these reasons, offices in the State Department using these tools must carefully consider the training data used and the types of bots each algorithm can detect when selecting the tool(s) that will best meet the goals of an office's projects. For instance, a bot detector focused on US elections would be a poor choice for Watch Officers focused on detecting bots meddling in international affairs. One way to elucidate the effectiveness of several tools for a particular project is to compile a dataset of correctly classified sample tweets. Comparing the accuracies of several bot detection algorithms in classifying bot versus non-bot could serve as a proxy for each algorithms' likely performance on future data. Following this testing, the State Department could inform employees about which tools they should use with the most accurate tools generally being the ones that have the highest accuracy on the State Department's custom datasets. A similar method could be used to determine the optimal classification threshold for each algorithm.

Additionally, sometimes bot-detecting algorithms provide information on their training dataset and scope which can be used to judge which tools seem most relevant; however, often, some of this data gets withheld by the algorithm creators as too much transparency can enable bots to evade detection (Martini et al., 2021). More extensive listings of bot detection tools can be found here: https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html#q=bot+detection. If all algorithms perform unsatisfactorily, it might be worthwhile for a group to design their own algorithm trained on bot classification data more relevant to the scope of the project.

Recommendations

- Use caution when using profile characteristics such as the location or number of followers for an account as it is easy for bot accounts to manipulate or inflate these.
- Look to past characteristics of bot accounts to make quick judgments on the reliability and source of information, such as repetitive Twitter handles, spikes in activity, and profile pictures.
- Supplement human judgment with digital tools, such as bot-detection algorithms.
- Choose which tools to use based on their performance on sample data representative of the types of bots the State Department will be detecting.
- Set classification thresholds for these tools based on which thresholds optimize classification accuracies on sample data representative of the types of the data the State Department will be examining.
- Routinely revisit decisions on protocols for bot identification tool use as bots continually evolve and develop new strategies to evade detection.

Astroturfing and Trolling

Astroturfing occurs when a centralized disinformation campaign takes on the disguise of a grassroot movement, often with the goal of convincing outsiders that popular support exists for an actor or idea. Typically, astroturf networks follow the principal-agent theory: an overseer (principal) works with several agents, who do a task for the overseer. There is an information asymmetry between the overseer and the agents because the overseer cannot constantly watch the agents, creating an incentive for the agents to take shortcuts in their work (Keller et al., 2020). The result is many low-quality accounts that barely fulfill the overseer's requirements. Flaws in these accounts and their posts can provide clues that its source is unreliable. Ideally, astroturfing campaigns can most accurately be detected by looking at the patterns in networks of accounts, versus individual accounts (Keller et al., 2022).

Trolling occurs when an account posts inflammatory, emotionally-charged content to rile up other users and encourage them to respond in a similar manner (Paavola et al., 2016). Because of this, tense political and social climates can provide a breeding ground for trolls. Trolls' comments often align with or target certain ideological grounds, and in a volatile sociopolitical content, real users are more sensitive to comments they perceive as attacking their viewpoints (Sanfilippo, 2017). Because people are more reactive to posts singling out their ideological identities, trolling can also be used to amplify aggression surrounding political beliefs.

Case Study: Russia

Russian astroturfing efforts tend not to deviate from the principal-agent model. The co-tweet network of the Russian Internet Research Agency (IRA) offers insight into how the government coordinates digital action. False accounts post similar or identical content, even when such overlap may be contradictory; for example, during the 2016 election, "the IRA campaign targeted both ends of the political spectrum and therefore posted very different messages. But… research showed that left-wing trolls impersonating Black Lives Matter activists and the right-wing accounts posted 1,661 identical tweets," (Keller et al., 2019). Because these false accounts often receive instructions from the principal at the same time, many

Russian astroturfing accounts will post similar content within a short period - interestingly, this short period frequently occurs during office hours in St. Petersburg (Keller et al., 2019). When trying to identify astroturfing efforts, contradictory or identical content is a red flag.

Russian trolling techniques can be seen when studying health disinformation. A study of health topics tweeted about by Russian trolls during the 2016 US presidential election revealed discussions on over 40 health-related topics, including 17 with significant differences between trolls mimicking left and right political ideologies. For instance, right trolls, supporters of Donald Trump, focused much more on topics such as Hillary Clinton's health and health insurance policy than left trolls, while left trolls, supporters of Bernie Sanders but not Hillary Clinton, emphasized topics like LGBT conversion therapy and the Flint Water Crisis (Karami et al., 2021). When talking about these topics, trolls covered a wide variety of topics and sought topics that would promote divisive discourse.

Case Study: China

International perspectives on Chinese astroturfing are largely concerned with the colloquial "50c party," named after rumors that ordinary citizens are paid 50 cents to combat agitating comments. This prevailing view is incorrect; most of the 50c participants appear to be government employees and they do not engage in online debate. The overwhelming majority of the posts are simply government cheerleading, as "most 50c posts from [the] data appear in highly coordinated bursts around events with collective action potential—either after unexpected events or before periods of time such as the Qingming festival and political meetings when collective action is perceived by the regime to be more likely," (King et al., 2017).

Posts appear to share similar content, occur in a short time frame, and are coordinated by a principal. Chinese astroturfing is a widespread government effort; in the years preceding 2017, 50c participants annually write about 448 million posts. 52.5% of the posts are on government sites, while 212 million posts are inserted into social media sites. The efforts rarely have a general focus, often featuring "specific intent and content," (King et al., 2017). In this case, red flags for astroturfing include similar posts that spike around times when collective actions are likely and if the posts are replies to government social media accounts.

China also utilizes 50c trolls for disruption and chaos. Similar to Russia, sometimes party-backed trolls engage in opinion wars with foreign audiences, creating rifts and spreading confusion. Additionally, they are very organized, with entire command structures of volunteers, which can increase the effectiveness of these operations (Fedasiuk, 2021). Just like Russia, they are used both abroad but also domestically as well and are the driving force behind boycotts and harassment operations on domestic networks against foreign businesses and actors. Finally, it is important to understand that just because networks like Twitter can catch many sock puppets and trolls used by the CCP, they are still a threat and there are more that have not been found (Fedasiuk, 2021). China's trolls are organized very efficiently and in a military format, do not expect Chinese trolls to work alone.

Case Study: Iran

Iranian usages of astroturfing are significant, due in part to the government's early forays into digital misinformation. Networks emphasize Iranian moral propaganda while downplaying international criticisms, especially of human rights concerns and political repression. Iranian efforts seem to be focused on influencing the opinion of Muslim populations in countries of indirect strategic importance (i.e.,

Indonesia, Nigeria). However, when Iranian disinformation networks focus on the US and its Western allies, it is "to achieve definable foreign policy objectives" (Brooking & Kianpour, 2020). The level of coordination between the networks of accounts is complex because of the structural robustness of Iranian media apparatuses:

> "Broadly, Iran's foreign influence efforts evidence a level of routinization that distinguish it from Russia, China, Saudi Arabia, or any other nation that has built a digital influence apparatus. This is a result of the early integration of digital manipulation into Iranian government and military functions. In 2009, Ayatollah Khamenei stated that 'content promotion' was 'the most effective international weapon' against foreign adversaries. In 2011, the head of the IRIB bragged that he had developed seven cyber battalions of 'media experts and specialists,' supposedly consisting of 8,400 members. Tehran's IRGC headquarters has trained thousands of recruits in 'content production,' teaching them social media strategy and graphic design" (Brooking & Kianpour, 2020).

As of January 2020, there have been 1,114 Facebook and 344 Instagram accounts attributed to Iran, which have been followed by 439,000 users; on Twitter, there have been 7,896 accounts that have sent nearly 8.5 million messages (Brooking & Kianpour, 2020). Suspect accounts tend not to disseminate obvious disinformation, but do not put much effort into the creation of false identities. Past accounts "relentlessly promoted their own material, willing to rapidly switch from one persona to another if it could improve their chances of engagement. In one case, an account that was called "Liberty Front Press" (named after an Iranian propaganda front) abruptly changed its username to "Berniecrats." Even with the sudden identity switch, however, the account's content remained the same," (Brooking & Kianpour, 2020). In other cases, Iranian accounts have created fake news sources or masqueraded as legitimate news sources to spread domestic propaganda abroad. After examining Iranian disinformation efforts, likely red flags for astroturfing include abrupt account name changes or obvious use of false accounts.

Iran uses trolls in Canadian politics, though overall Russian and Chinese trolls get more attention. While Russian trolls attacked Canadian politicians like Justin Trudeau, Iranian trolls "disseminated false reports on Stephen Harper shortly before the 2015 Canadian election, suggesting that the CIA installed him as prime minister and that he was an ISIS supporter" (Al-Rami, 2021). The interesting thing about Iranian trolls is that, unlike Chinese and Russian trolls who sowed disruption by playing both sides or by having a very strict command structure, Iranian trolls were mainly sympathetic tweets for refugees and turning those stories against countries such as Canada. (Al-Rami, 2021). Iranian trolls utilize humanitarian crises and human rights issues in their attacks, pay attention to why the user is bringing up the human rights issues and who they are attacking with it.

Recommendations

When determining if an account or post is suspect, the following questions may useful in the Watch Officer's evaluation:

- Does the post clearly identify an owner, geographical location, or government affiliation?
- Does this post post similar or identical content to its peers?
- Is the post within a cluster of other posts that have posted similar content within a short time frame?

- Is the post affiliated with a government employee of a country with a known history of online disinformation?
- Has the account abruptly changed names, content, language, or tone?
- Did the post appear during routine office hours for its originating country?

Conclusion

As technology continues to evolve and improve, online dis/misinformation is only becoming a more pressing problem. Watch Officers monitoring social media and trying to assess the accuracy of posts should lean on a solid understanding of the state of disinformation when making quick decisions. While disinformation can reside in the most inconspicuous parts of the digital sphere, many types of disinformation are patterned and utilized routinely by state actors, such as China, Russia, and Iran. Familiarity with these patterns and common biases in Watch Officers' assessments is critical to quickly and accurately identifying disinformation. There are various recommended techniques and questions that Watch Officers can utilize to sculpt a mindset that is conducive to spotting irregularities that are indicative of disinformation.

Utilizing tools to detect dis/misinformation in the forms of multimedia manipulation and bots can assist with determining the reliability of new information. There are many open-source tools available that can help spot disinformation, as well as algorithms to detect bot activity. It is important to note that when utilizing these tools, there should be frequent preventative checks to ensure that the information is up-to-date due its constant evolution.

Addendum A – Watch Officer "Cheat Sheet"

Multimedia Manipulation

- Consider that sometimes the most harmful deepfakes and memes are not the highest quality, but are simply the most popular.
- Pay attention to the quality of videos (large pixels, poor audio, etc.), if the actions or words said seem out of place, and if there are any inconsistencies.
- Consider the account posting the deepfake or meme to understand if it is part of a political media campaign. Consider user following, follower interactions, what kinds of accounts are posting similar information.

Bots

- Use caution when using profile characteristics such as the location or number of followers for an account as it is easy for bot accounts to manipulate or inflate these.
- Look to past characteristics of bot accounts to make quick judgments on the reliability and source of information, such as repetitive Twitter handles, spikes in activity, and profile pictures.

Astroturfing and Trolls

- Does the post clearly identify an owner, geographical location, or government affiliation?
- Does this post post similar or identical content to its peers?
- Is the post within a cluster of other posts that have posted similar content within a short time frame?
- Is the post affiliated with a government employee of a country with a known history of online disinformation?
- Has the account abruptly changed names, content, language, or tone?
- Did the post appear during routine office hours for its originating country?

Addendum B – Watch Officer "Online Toolkit"

Social Listening Dashboard

- https://diplomacy-lab.lib.purdue.edu/tools/dashboard

Edited Image Identification Practice

- Which Face is Real?

Reverse Image Checker

- TinEye
- Google Image Reverse Search

Information Literacy Modules

- https://diplomacy-lab.lib.purdue.edu/education/literacy-modules

Flowchart and Checklists

- https://diplomacy-lab.lib.purdue.edu/tools/flowchart

Bot Detection

- Botometer
- Tweetbotornot
- More tools can be found here: https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html#q=bot+detection.

References

Adame, B. J. (2016). Training in the mitigation of anchoring bias: A test of the consider-the-opposite strategy. *Learning and Motivation, 53,* 36–48. https://doi.org/10.1016/j.lmot.2015.11.002

Al-Rami, A. (2021). How did Russian and Iranian trolls' disinformation toward Canadian issues diverge and converge? *Digital War*, *2*(1-3), 21–34. https://doi.org/10.1057/s42984-020-00029-4

Apuzzo, M., & Satariano, A. (2019, May 12). *Russia is targeting Europe's elections. so are far-right copycats*. The New York Times. Retrieved November 3, 2022, from https://www.nytimes.com/2019/05/12/world/europe/russian-propaganda-influence-campaign-european-elections-far-right.html

BBC. (n.d.-a). GOV.UK. Retrieved November 5, 2022, from https://www.gov.uk/government/organisations/bbc#:~:text=BBC%20is%20a%20public%20corporation,%2C%20Culture%2C%20Media%20%26%20Sport.

BBC. (n.d.-b). Licence fee and funding. Retrieved November 5, 2022, from https://www.bbc.co.uk/aboutthebbc/governance/licencefee/

Bleck, J., & Michelitch, K. (2017). Capturing the Airwaves, Capturing the Nation? A Field Experiment on State-Run Media Effects in the Wake of a Coup. *The Journal of Politics*, *79*(3), 873–889. https://doi.org/10.1086/690616

Bogart, L. M., & Thorburn, S. (2005). Are HIV/AIDS conspiracy beliefs a barrier to HIV prevention among African Americans?. *Journal of acquired immune deficiency syndromes*, *38*(2), 213–218. https://doi.org/10.1097/00126334-200502010-00014

Borenstein, E. (2022). Meanwhile, in Russia : Russian memes and viral video culture. *Bloomsbury Academic*. https://doi.org/10.5040/9781350181564

Broinowski, A. (2022). Deepfake Nightmares, Synthetic Dreams: A Review of Dystopian and Utopian Discourses Around Deepfakes, and Why the Collapse of Reality May Not Be Imminent—Yet. *Journal of Asia-Pacific Pop Culture, 7*(1), 109-139. https://www.muse.jhu.edu/article/857647.

Brooking, E. T. & Kianpour, S. (2020). Iranian digital influence efforts: Guerrilla broadcasting for the twenty-first century. *Atlantic Council*. https://www.atlanticcouncil.org/in-depth-research-reports/report/iranian-digital-influence-efforts-guerrilla-broadcasting-for-the-twenty-first-century/

CISA. (n.d.). Mis, dis, malinformation. https://www.cisa.gov/mdm

Doherty, T. & Carroll, A. E. (2020). Believing in Overcoming Cognitive Biases. *AMA Journal of Ethics.* https://journalofethics.ama-assn.org/article/believing-overcoming-cognitive-biases/2020-09

Downing, J., & Ahmed, W. (2019). MacronLeaks as a "warning shot" for European democracies: challenges to election blackouts presented by social media and election meddling during the 2017 French presidential election. *French Politics*, *17*(3), 257–278. https://doi.org/10.1057/s41253-019-00090-w

Farris, R., Benkler, Y., & Roberts, H. (2018). *Network Propaganda*. Oxford University Press.

Fazio, L. (2020, February 14). *Out-of-context photos are a powerful low-tech form of misinformation.* The Conversation. https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959

Fedasiuk, R. (2021). *China's internet trolls go global*. Council on Foreign Relations. Retrieved November 1, 2022, from https://www.cfr.org/blog/chinas-internet-trolls-go-global

Ferreira, R. R. (2022). Liquid Disinformation Tactics: Overcoming Social Media Countermeasures through Misleading Content. *Journalism Practice*, *16*(8), 1537–1558. https://doi.org/10.1080/17512786.2021.1914707

Greenberg, A. (2019, June 12). *Alphabet-owned Jigsaw bought a Russian troll campaign as an experiment.* Wired. Retrieved November 3, 2022, from https://www.wired.com/story/jigsaw-russia-disinformation-social-media-stalin-alphabet/

Gregory, S. (2022). Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism (London, England)*, *23*(3), 708–729. https://doi.org/10.1177/14648849211060644

Harold, S. W., Beauchamp-Mustafaga, N., & Hornung, J. W. (2021). Chinese disinformation efforts on social media. *Rand Corporation.* https://doi.org/10.7249/rr4373.3

Jones, M. O. (2019). Propaganda, Fake News, and Fake Trends: The Weaponization of Twitter Bots in the Gulf Crisis. *International Journal of Communication*, *13*, 1389–1415.

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, *23*(1), 130–137. https://doi.org/10.1038/s41593-019-0549-2

Karami, A., Lundy, M., Webb, F., Turner-McGrievy, G., McKeever, B. W., & McKeever, R. (2021, February 23). Identifying and analyzing health-related themes in disinformation shared by conservative and liberal Russian trolls on Twitter. *MDPI*. Retrieved November 1, 2022, from https://www.mdpi.com/1660-4601/18/4/2159/htm

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2022). Coordination patterns reveal online political) astroturfing across the world. *Scientific Reports*, *12*(1), 4572–4572. https://doi.org/10.1038/s41598-022-08404-9

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2020). Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication*, *37*(2), 256–280. https://doi.org/10.1080/10584609.2019.1661888

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2019). It's not easy to spot disinformation on Twitter. Here's what we learned from 8 political 'astroturfing' campaigns. *The Washington Post*. https://www.washingtonpost.com/politics/2019/10/28/its-not-easy-spot-disinformation-twitter-heres-what-we-learned-political-astroturfing-campaigns/

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument. *The American Political Science Review*, *111*(3), 484-501. https://doi.org/10.1017/S0003055417000144

Lam, N. (2021, October 5). Library Guides: News: Fake News, Misinformation & Disinformation. Guides.lib.uw.edu. https://guides.lib.uw.edu/c.php?g=345925&p=7772376

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). "THE RUSSIANS ARE HACKING MY BRAIN!" investigating Russia's internet research agency twitter tactics during the 2016 United States presidential campaign. *Computers in Human Behavior*, *99*, 292–300. https://doi.org/10.1016/j.chb.2019.05.027

Madrigal, A. C. (2018, October 26). Iranian Propaganda Targeted Americans With Tom Hanks. It's yet another attempt by a government to use Facebook to sow discord in the United States. *The Atlantic*. Retrieved November 1, 2022, from https://www.theatlantic.com/technology/archive/2018/10/irans-facebook-propaganda-targeted-americans-tom-hanks/574129/

Martini, F., Samula, P., Keller, T. R., & Klinger, U. (2021). Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data & Society*, *8*(2), 205395172110335–. https://doi.org/10.1177/20539517211033566

Ndlela, M. N., & Mano, W. (2020). Social Media and Elections in Africa, Volume 2 Challenges and Opportunities (M. N. Ndlela & W. Mano, Eds.; 1st ed. 2020.). *Springer International Publishing*. https://doi.org/10.1007/978-3-030-32682-1

Neyazi, T. A. (2020). Digital propaganda, political bots and polarized politics in India. *Asian Journal of Communication*, *30*(1), 39–57. https://doi.org/10.1080/01292986.2019.1699938

Nimmo, B. (2018). Robot wars: How bots joined battle in the Gulf. *Journal of International Affairs (New York)*, *71*(1.5), 87–96.

Nimmo, B., Buziashvili, E., Seldon, M., Karan, K., Aleksejeva, N., Bandeira, L., & Andriukaitis, L. (2019, June 22). *Top takes: Suspected Russian Intelligence Operation.* Medium. Retrieved November 3, 2022, from https://medium.com/dfrlab/top-takes-suspected-russian-intelligence-operation-39212367d2f0

Paavola, J., Helo, T., Jalonen, H., Sartonen, M., & Huhtinen, A.M. (2016). *Understanding the trolling phenomenon: The automated detection of bots*. Retrieved November 1, 2022, from https://www.jstor.org/stable/26487554

Patterson, T. (2020). Election Beat 2020: How news outlets become misinformation superspreaders. *Journalist's Resource*. Retrieved November 6, 2022, from https://journalistsresource.org/politics-and-government/news-misinformation-superspreaders/

Paul, C., & Matthews, M. (2016). The Russian "firehose of falsehood" propaganda model: Why it might work and options to counter it. *Rand Corporation.* Retrieved November 6, 2022, from https://doi.org/10.7249/pe198

Sanfilippo, M. R., Fichman, P., & Yang, S. (2017). Multidimensionality of online trolling behaviors. *The Information Society*, *34*(1), 27–39. https://doi.org/10.1080/01972243.2017.1391911

Seta, D. G. (2021, July 30). Huanlian, or changing faces: Deepfakes on Chinese digital media platforms. *Sage Journals*, 27(4), 935-953. https://journals.sagepub.com/doi/10.1177/13548565211030185

Schuchard, R., Crooks, A. T., Stefanidis, A., & Croitoru, A. (2019). Bot stamina: examining the influence and staying power of bots in online social networks. *Applied Network Science*, *4*(1), 1–23. https://doi.org/10.1007/s41109-019-0164-x

Selvage, D. (2020, May 26). Lessons from Operation "Denver," the KGB's massive AIDS disinformation campaign (Interview by M. Kramer) [Transcript]. The MIT Press Reader. Retrieved November 4, 2022, from https://thereader.mitpress.mit.edu/operation-denver-kgb-aids-disinformation-campaign/

Smart, B., Watt, J., Benedetti, S., Mitchell, L., & Roughan, M. (2022). #IStandWithPutin Versus #IStandWithUkraine: The Interaction of Bots and Humans in Discussion of the Russia/Ukraine

War. In *Social Informatics* (pp. 34–53). Springer International Publishing. https://doi.org/10.1007/978-3-031-19097-1_3

Stieglitz, S., Brachten, F., Berthelé, D., Schlaus, M., Venetopoulou, C., & Veutgen, D. (2017). Do Social Bots (Still) Act Different to Humans? – Comparing Metrics of Social Bots with Those of Humans. *Social Computing and Social Media. Human Behavior*, 379–395. https://doi.org/10.1007/978-3-319-58559-8_30

Sunstein, C. R. (2006). The availability heuristic, intuitive cost-benefit analysis, and climate change. *Climatic Change*, *77*(1-2), 195–210. https://doi.org/10.1007/s10584-006-9073-y

Treyger, E., Cheravitch, J., & Cohen, R. S. (2022). Russian disinformation efforts on social media. *Rand Corporation*. Retrieved November 6, 2022, from https://doi.org/10.7249/rr4373.2

Tsohou, A., Karyda, M., & Kokolakis, S. (2015). Analyzing the role of cognitive and cultural biases in the internalization of information security policies: Recommendations for information security awareness programs. *Computers & Security*, *52*, 128–141. https://doi.org/10.1016/j.cose.2015.04.006

US House of Representatives Permanent Select Committee on Intelligence. (n.d.). *Exposing Russia's effort to sow discord online: The Internet Research Agency and advertisements* [Press release]. https://intelligence.house.gov/social-media-content/default.aspx

Uyheng, J., & Carley, K. M. (2020). Bot Impacts on Public Sentiment and Community Structures: Comparative Analysis of Three Elections in the Asia-Pacific. In *Social, Cultural, and Behavioral Modeling* (pp. 12–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-61255-9_2 ‘

Uzonwanne, F. C. (2016). Rational Model of Decision Making. In: Farazmand, A. (eds) *Global Encyclopedia of Public Administration, Public Policy, and Governance*. Springer, Cham. https://doi.org/10.1007/978-3-319-31816-5_2474-1

Walker, C., & Orttung, R.W. (2014). Breaking the News: The Role of State-Run Media. *Journal of Democracy*, *25*(1), 71-85. https://doi.org/10.1353/jod.2014.0015

West, J., & Bergstrom, C. (2019). Which Face is Real?. *University of Washington*. https://www.whichfaceisreal.com/about.html

Wood, D., McMinn, S., & Feng, E. (2019). China Used Twitter To Disrupt Hong Kong Protests, But Efforts Began Years Earlier. *NPR*. Retrieved November 6, 2022, from https://www.npr.org/2019/09/17/758146019/china-used-twitter-to-disrupt-hong-kong-protests-but-efforts-began-years-earlier

K, M., Scheirer, W., & Weninger, T. (2021). Meme warfare: AI countermeasures to disinformation should focus on popular, not perfect, fakes. *Bulletin of the Atomic Scientists*, *77*(3), 119–123. https://doi.org/10.1080/00963402.2021.1912093

Zhōngguó xīnwén shè zhōng xīn shè, zhōng xīn wǎng Twitter zhànghào tuīguǎng zhèngfǔ cǎigòu xiàngmù zhòngbiāo gōnggào [China News Agency, China News Agency, and China News Network Twitter account to promote the announcement of winning the bid for government procurement projects]. (2019, August 16). Retrieved November 3, 2022, from https://web.archive.org/web/20190822085751/http://www.ccgp.gov.cn/cggg/dfgg/zbgg/201908/t20190816_12699714.html