PSAA 608 Paper

Political Disinformation on the Internet and legalities

"One nation controlled by the media" may not be part of the pledge of allegiance but is a verse from a 2004 song about a dystopian government by punk rock band Green Day, and the sad reality that this country is headed towards. The modernization of the Smith-Mundt act in 2012 brought the question of domestic use of propaganda back in the spotlight (Hudson 2013). With the 2024 election season rapidly approaching, major conflicts happening in the Middle East, Ukraine and tensions in the South China Sea, America must brace for the worst regarding political disinformation coming its way. While social media has enabled widespread communication of disinformation and facilitated its distribution, such campaigns existed long before the internet came to be. Disinformation scholars like Thomas Rid often attribute the first ever disinformation campaign to Imperial Russia, with the publication of the Protocols of the Elders of Zion which later served as a base for the Holocaust and antisemitism throughout the world (Rid 2021). Trends in disinformation have shifted as well, while still using similar methods to the 20th century, moving us from forged documents like the Tanaka memo, outlining Japan's ambition for total Asian control (Stephan 1973), to generative AI models being used to create fake images of Donald Trump being arrested and tackled by police (Arijeta Lajka 2023). Facing the reality that disinformation will never end but rather adapt to our technologies and contemporary domestic political issues, we can wonder if any laws or technical advancements can help us fight a cyber-attack aimed directly at the user, and not the machine they are on. We can examine different strategies adversarial nations have used, technologies employed, similarities with other campaigns, research done, and legal steps taken to prevent an information age of hysteria.

First, it's important to understand what a political disinformation campaign actually is. The objective of a disinformation campaign is to "influence others, stir them to action, and cause harm" which can be done through a multitude of techniques CISA outlined in a concise publication (CISA 2022). Some of them include astroturfing, the cultivation of online personas, targeted content and the exploitation of information gaps. Many different terms exist to address the same issue, leading to confusion and sometimes misuse and misunderstanding of the problem. Historical scholars often talk about "active measure" campaigns to refer to events that happened in the 20th century that did not involve the internet; it's the term used by the KGB to describe some of their influence operations during the Cold War (Rid 2021). Journalists prefer the term "fake news" to differentiate with their own work. Then comes the three biggest buzzwords used by the media and congressional representatives are "misinformation" which is information that is false but built on a core of truth, "disinformation" which is completely false although my exploit a lack of understanding and sensationalism to get the viewer to believe it, and "malinformation" which is true information but overshared to cause harm.

Iran, Russia and China all have different strategies for political disinformation and use of social media for propaganda purposes when turned outwards.

Endless Mayfly is an Iranian led disinformation campaign, aimed at causing division among the American population "in advance of the 2020 elections" (Vavra 2020). Experts including Gabrielle Lim estimate the operation began as early as 2016 when multiple accounts were created all linking back to a page called "Peace, Security and Justice Community" (Lim 2020), impersonating a small organization. The main method used for the campaign is known as "astroturfing", described by CISA as posting overwhelming amounts of content with "similar messaging from several inauthentic accounts"(CISA 2022). Narratives spread are to sow political discord, especially at

time when tensions between nations might be critical. One of the articles quoted by Lim mentions "Europe[‘s] fear of Erdogan's anger" (Lim n.d.), which came at a time of heightened tension between the Turkish and Dutch governments, when the Dutch prime minister was up for reelection. What makes Iranian disinformation methods unique is their direct targeting of foreign journalists first through their astroturfing, making it difficult for the press to find contradicting opinions then by their use of screenshots with shortened links from bit.ly and url spoofing and misspelling, like sending social media users to like "indepednent.co" instead of "independent.co.uk" (Lim n.d.). After their message has gained enough traction, agents will delete their link, or change it to a url leading to the homepage of the original website, causing users (including the press) to be redirected to the homepage of a legitimate news site and making them believe they interacted with a real but now deleted page. By cultivating their online personas as activists and journalists, agents were able to directly interact with members of the press and researchers about current issues (Lim et al. 2019). The main effects of the campaign, and its impact were felt when political personalities like France's Marion Marechal Le Pen, a prominent far-right figure shared an article that claimed "Saudi Arabia was supporting then-presidential candidate Emmanuel Macron" in the upcoming 2017 election (Yates, Rogers, and Rocha 2019).

Russian tactics for disinformation are slightly different to Iran's, while still trying to obtain similar results, discord within the population and overall manipulation of the political conversation in the United States. While Iran targets journalists, Russia focuses on the users of social media and trying to cause physical events, mainly using uncontrolled social media sites, or platforms that the federal government can't seem to regulate and that do very little censoring and controlling of the content on them. With operations like Endless Mayfly being directly attributed to "Iran-backed" groups (Lim et al. 2019) (likely the cyber division of the IRGC), Russian disinformation goes through a

private company, the Wagner group, infamous for their role in the invasion of Ukraine and their failed coup against Vladimir Putin. Unlike the Iranians who impersonate real news sites, Russian disinformation operators create their own “hyper partisan” pages and share divisive content, mainly aimed to divide American along strong domestic issues like race relations, which was highlighted by the Senate Select Committee for Intelligence (US Senate 2018). The objective was to create riots and direct physical confrontation between different groups of Americans, with different floors of the Internet Research Agency targeting different groups and able to pitch them against each other (Sanger 2019). Recently, with the invasion of Ukraine underway, Russian disinformation tactics have shifted in support of the war effort and have focused efforts on exposing “Ukrainian war crimes” (Sardarizadeh 2022), emphasizing how much money has been spent by Western powers on the conflict and conspiracies about the Nord stream pipeline (Al Jazeera 2024).

Chinese efforts mainly focus on shifting attention away from their actions, whether in the South China sea, online or to deflect from current issues and national problems, like China’s growing pollution and air quality. As Asian cultures often try avoiding to “lose face”, China has adopted a disinformation strategy of whataboutism, or the art of bringing up a whole other issue when faced with accusations like the Chinese government pointing to Fort Dietrich, the Army’s biological weapon study center as the origin of COVID-19 when faced with numerous investigations. Current tensions in South Asian seas and lessons learned from the current events in Ukraine lead many to believe that disinformation tactics and campaigns of mass influence will be used as a preemptive strike against America and every other nation that aims at protecting Taiwan. As Putin mentioned in his interview with Tucker Carlson that if the flow of weapons stopped coming into Ukraine (Al Jazeera 2024), the fighting would end, we could see similar narratives being employed by China

to try to deescalate or gain an advantage in the conflict. Part of whataboutism campaigns is trying to get social media users to look at the mistakes of their own government before looking at those of a foreign one or refreshing their motivations by "echoing the current problems faced by our nation" (Robic 2024). China also has access to a great weapon to be used against our youth, TikTok. With attention spans decreasing and social media platforms offering rapidly digestible formats like reels on Instagram and Facebook or shorts on Youtube, China has understood that the next disinformation battle will be won with short videos and not long articles. TikTok in China is completely different than what is presented to an American audience, with Chinese youth being offered CCP-approved "science, educational and historical content" while American teens are shown "stupid dance videos with the main goal of making us imbeciles" (Schlott 2023).

Our adversaries may have different strategies at trying to sow discord and division in our nation, or convincing our next generation their nation is one to be ashamed of but all of them have had a significant effect on our institutions and way of life already. One measure already taken by Texas governor Greg Abbott was banning the use of TikTok on government owned devices, citing that the app "harvests significant amounts of data from a user's device" (Press Release 2023). While writing this paper, Congress passed a bill forcing Chinese company ByteDance to sell TikTok to "a buyer that satisfies the U.S. government" (McCabe and Maheshwari 2024), due to concerns that the app could be used to spread disinformation, especially so close to the elections. Seeing this as a first step to preventing both the collection of data from government devices and the dissemination of foreign backed political messages to State employees, we can therefore look at other measures to prevent disinformation from reaching the American public. As these campaigns are launched on the internet, using digital means, they can be more easily analyzed, especially with modern Machine Learning capabilities.

One of the recommendations of the Cyberspace Solarium Commission was to fund "non-governmental disinformation researchers", among others (Cyberspace Solarium Commission 2020). Reverse engineering misinformation campaigns, from a technical point of view has worked in the past and led to the development of numerous methods of analysis and detection of misleading posts on the internet. A survey of currently available Computer Science literature on the topic shows that some of the most used techniques include Linguistic-based analyses (Moura, Sousa-Silva, and Lopes Cardoso 2021; de Oliveira et al. 2021; Schuster et al. 2020) with some looking at term frequency to determine the content type (Kaur, Kumar, and Kumaraguru 2020). The biggest issue facing research is images, as some generative AI models can embed hidden messages in them. Because of this, optical character recognition software (OCR) can't properly identify which text is in the image and has led to online communities being able to circumvent censorship mechanisms. OpenAI, the same company behind the infamous GPT and DallE generative models has recently released a new text-to-video model, Sora, able to convert a user input into a nearly picture-perfect video. This, combined with the astroturfing method could open the floodgates for new, AI-enabled campaigns in the near future, as AI can "lower the cost to produce" disinformation (Hu 2024). Because of such technical trends, future research will have to focus on human-like recognition of content, and recognizing text that has "extra letters" or "chang[ed] spelling so that the pronunciation is different enough to still be understood as the original message" (Robic 2024).

Future research will have to especially focus on memes and images, especially trying to grasp an understanding of internet linguistics and vocabularies that different online communities use. A quick survey of the infamous 4chan's politically incorrect board shows a wide array of words that would otherwise be censored on social media. Furthermore, some of the most interacted with posts include language that would be qualified by researchers as extremely emotional, with phrases like

"I would get out of France ASAP" (As Soon As Possible) as French President Macron discusses the future of warfare in Eastern Europe. Other users are all adding to it, posting phrases like "it's happening", creating a sense of urgency and immediacy of the current situation. While most projects have focused on a single platform, or with datasets easily obtainable from large social media companies, many researchers have failed to look at the smaller ones or make cross-platform connections. Research will therefore have to evolve more to identify campaigns at their source, on smaller uncontrolled platforms before they make their way into the public.

The biggest issue facing policymakers when dealing with disinformation campaigns on social media is the ownership of such platforms and first amendment concerns. Before section 230, social media sites were not held "liable as a publisher if any of that content was defamatory" but would be liable if they decided to remove "any third-party material" (here defined as material not posted by the platform itself, but rather by individual users, or third parties) (Department of Justice 2020). Section 230 was therefore enacted to provide immunity to platforms for "third-party content on their services" and the "removal of certain categories" of online content. The main issue regarding section 230, as outlined by the Department of Justice's is that these platforms are now "some of the nation's largest and most valuable companies" and are no longer simple platforms for posting "third-party content" but now have access to sophisticated algorithms to recommend content to users and further connections. Because of this rise in connectivity and the possibility to interact with nearly anyone in the world, numerous illegal activities "including child sexual exploitation, selling illicit drugs, cyberstalking, human trafficking and terrorism" have flourished (Department of Justice 2020). In wake of the Covid-19 pandemic, Americans are spending a lot more time on the internet, especially social media, reaffirming the need for "a reassessment of America's laws governing the internet", especially now that social media companies are no longer "nascent or

fragile" (Department of Justice 2020). The question therefore arises of modifying the blanket immunity of section 230, and putting a bit more responsibility for the censorship of content in the hands of social media companies but raises the concern, especially in times of elections, of the political bias some people at the top of those enterprises may have. A good case study, and one that made national headlines is the Hunter Biden laptop story, and the coverage of it on Twitter and subsequent censorship by tech companies of the story. The article first mentioning it was censored by employees since it was deemed "unclear whether the material was authentic", a valid concern to have during an election season (Bond 2022). With access to such large datasets of content flagged as potentially harmful, national security services and social media companies alike have created solutions to automatically "detect and reject" content, as the United Kingdom's Home office did in 2018 with a tool that boasts a surprising "99.99 percent success rate" (Macdonald, Correia, and Watkin 2019). This is mainly limited to posting content that "promotes [terrorist] acts, incites violence or celebrates [terrorist] acts" (Macdonald, Correia, and Watkin 2019) which all platforms and lawmakers, including in section 230 a part about immunity for moderation of content that promotes "violence or terrorism" (Department of Justice 2020), and does not include anything about posting disinformation. Instead, platforms are bound to their terms of service which, again, are at the mercy of whoever runs the social media company.

As part of a strong military alliance, the United States might be tempted to cooperate with NATO regarding disinformation campaigns as a form of cyber attack against a nation but this would require consensus among members regarding such a definition.

The Tallinn manual, as NATO's main rule book and "objective restatement of international law" (CCDCOE n.d.) for cyber operations and conflicts of digital nature can be further analyzed in the context of disinformation campaigns, especially rules that may relate to civilians, private

infrastructure, and cyber espionage. An interesting rule to start with is number 36, stating that cyber terror attacks can be defined as attacks or threats meant to "spread terror among the civilian population" but states below that "a false tweet […] sent out in order to cause panic [..] does not violate this rule" (Schmitt 2017). Rule 27, states that civilians that engage in cyber operations as part of a "Levée en masse" (mass rising) benefit from the same privileges as prisoners of war and armed combatants in a conflict, similar to that of a conventional army (Schmitt 2017). Rule 32 states that "the civilian population […] shall not be the object of cyber-attack" but does not go into detail about the type of attack, likely implying direct threat to an individual's devices and machines, as opposed to accounts and data stored on such devices (Schmitt 2017). Rule 37 adds to that with the prohibition against civilian objects, stating that only "cyber infrastructure […] if they are military objectives" may be targeted (Schmitt 2017). Rule 49 brings in further detail, adding that attacks "without distinction" are prohibited, meaning that targeting an entire population without concern for differentiation would fall into an unsanctioned act (Schmitt 2017). What makes this situation interesting is that, as defined by the Tallinn manual, disinformation operations would not qualify under rule 66 as they are not designed to "gather (or attempt to gather) information". Rule 82 might have a direct application to the question of disinformation campaigns as it specifically prohibits the use of "digital cultural property for military purposes" (Schmitt 2017), meaning the spreading of memes and other forms of culture on the internet for propaganda purposes, especially in trying to create a 5^{th} column effect through the internet. While all these rules seem to prohibit NATO member states from engaging in disinformation campaigns on foreign social media sites, they do not help qualify such campaigns as attacks. Rule 30 states that a cyber operation is an attack if it intends in causing "injury or death" to a person or "damage or destruction" to objects (Schmitt 2017). In this case, influence would not qualify as an injury,

meaning adversarial nations can freely launch disinformation campaigns against NATO member countries without any fear or this being considered an attack, at least according to the Tallinn manual version 2.0. The next version of the Tallinn manual, version 3.0 will therefore have to address disinformation campaigns, international law regarding social media (both domestic networks and foreign ones) as well as the involvement of third-party groups like the Internet Research Agency in such campaigns.

We haven't exactly "lost" the information war as authors like Nina Jankowicz may suggest, but we certainly are not winning it either (Jankowicz 2020). Looking at case studies from other nations like Estonia, it seems like the solution would simply rely in the education of citizens about not believing every piece of information that is thrown at you and knowing accurate historical information and world affairs. This does however raise a new set of issues, that of the American public education system and requiring a consensus regarding current world affairs, which may not always be possible due to conflicting opinions in government and from educators alike. The Solarium white paper does mention establishing a "Civic Education Task Force", in other words raising "public awareness about foreign disinformation" (Cyberspace Solarium Commission 2020) but this idea would require bipartisan support and most importantly a consensus regarding what can be labeled as foreign disinformation. Such a bill has been introduced by Rep. Donald Beyer (D-VA-8) and aims at supporting "information and media literacy" as well as preventing "misinformation and disinformation" (Rep. Beyer, Donald S., Jr. [D-VA-8] 2022). Parties may be tempted to use this to gain popularity if the campaign gives them an advantage in the next election, regardless of moral and ethical concerns. A disinformation campaign should be regarded as a cyber-attack regardless of definition, as it can be seen as a "compromise on information integrity" (Dupuis and Williams 2019).

Sources:

Al Jazeera. 2024. “What Did Vladimir Putin Say to Tucker Carlson? Five Key Takeaways.” *Al Jazeera*. https://www.aljazeera.com/news/2024/2/9/five-key-moments-from-tucker-carlsons-interview-with-vladimir-putin.

Arijeta Lajka. 2023. “Trump Arrested? Putin Jailed? Fake AI Images Flood the Internet, Increasing ‘Cynicism Level.’” *Sydney Morning Herald*. https://www.smh.com.au/world/north-america/trump-arrested-putin-jailed-fake-ai-images-flood-the-internet-increasing-cynicism-level-20230324-p5cuup.html.

Bond, Shannon. 2022. “Elon Musk Is Using the Twitter Files to Discredit Foes and Push Conspiracy Theories.” *NPR*. https://www.npr.org/2022/12/14/1142666067/elon-musk-is-using-the-twitter-files-to-discredit-foes-and-push-conspiracy-theor.

CCDCOE. “The Tallinn Manual.” *NATO Cooperative Cyber Defence Centre of Excellence*. https://ccdcoe.org/research/tallinn-manual/.

CISA. 2022. “Tactics of Disinformation.” https://www.cisa.gov/sites/default/files/publications/tactics-of-disinformation_508.pdf.

Cyberspace Solarium Commission. 2020. “Cyberspace Solarium Commission White Paper #6: Countering Disinformation in the United States.” https://www.solarium.gov/public-communications/disinformation-white-paper.

Department of Justice. 2020. “DEPARTMENT OF JUSTICE’S REVIEW OF SECTION 230 OF THE COMMUNICATIONS DECENCY ACT OF 1996.” https://www.justice.gov/archives/ag/department-justice-s-review-section-230-communications-decency-act-1996.

Dupuis, Marc J., and Andrew Williams. 2019. “The Spread of Disinformation on the Web: An Examination of Memes on Social Networking.” In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Leicester, United Kingdom: IEEE, 1412–18. doi:10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00256.

Hu, Charlotte. 2024. “How AI Bots Could Sabotage 2024 Elections around the World.” *Scientific American*. https://www.scientificamerican.com/article/how-ai-bots-could-sabotage-2024-elections-around-the-world/.

Hudson, John. 2013. “U.S. Repeals Propaganda Ban, Spreads Government-Made News to Americans.” *Foreign Policy*. https://foreignpolicy.com/2013/07/14/u-s-repeals-propaganda-ban-spreads-government-made-news-to-americans/.

Jankowicz, Nina. 2020. *How to Lose the Information War: Russia, Fake News, and the Future of Conflict*. London New York Oxford New Delhi Sydney: I.B. Tauris.

Kaur, Sawinder, Parteek Kumar, and Ponnurangam Kumaraguru. 2020. "Automating Fake News Detection System Using Multi-Level Voting Model." *Soft Computing* 24(12): 9049–69. doi:10.1007/s00500-019-04436-y.

Lim, Gabrielle. 2020. "Copycat Websites: The Endless Mayfly Network." *mediamanipulation.org*. https://mediamanipulation.org/case-studies/copycat-websites-endless-mayfly-network.

Lim, Gabrielle. "Case Study: Attributing Endless Mayfly." *datajournalism.com*. https://datajournalism.com/read/handbook/verification-3/investigating-platforms/11a-case-study-attributing-endless-mayfly.

Lim, Gabrielle, Etienne Maynier, John Scott-Railton, Alberto Fittarelli, Ned Moran, and Ron Deibert. 2019. "Burned After Reading: Endless Mayfly's Ephemeral Disinformation Campaign." *citizenlab.ca*. https://citizenlab.ca/2019/05/burned-after-reading-endless-mayflys-ephemeral-disinformation-campaign/.

Macdonald, Stuart, Sara Giro Correia, and Amy-Louise Watkin. 2019. "Regulating Terrorist Content on Social Media: Automation and the Rule of Law." *International Journal of Law in Context* 15(2): 183–97. doi:10.1017/S1744552319000119.

McCabe, David, and Sapna Maheshwari. 2024. "What to Know About the TikTok Bill That the House Passed." *New York Times*. https://www.nytimes.com/2024/03/13/technology/tiktok-ban-law-congress.html.

Moura, Ricardo, Rui Sousa-Silva, and Henrique Lopes Cardoso. 2021. "Automated Fake News Detection Using Computational Forensic Linguistics." In *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, eds. Goreti Marreiros, Francisco S. Melo, Nuno Lau, Henrique Lopes Cardoso, and Luís Paulo Reis. Cham: Springer International Publishing, 788–800. doi:10.1007/978-3-030-86230-5_62.

de Oliveira, Nicollas R., Pedro S. Pisa, Martin Andreoni Lopez, Dianne Scherly V. de Medeiros, and Diogo M. F. Mattos. 2021. "Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges." *Information* 12(1): 38. doi:10.3390/info12010038.

Press Release. 2023. "Governor Abbott Announces Statewide Plan Banning Use Of TikTok." *Office of the Texas Governor | Greg Abbott*. https://gov.texas.gov/news/post/governor-abbott-announces-statewide-plan-banning-use-of-tiktok.

Rep. Beyer, Donald S., Jr. [D-VA-8]. 2022. *Educating Against Misinformation and Disinformation Act*.

Rid, Thomas. 2021. *Active Measures: The Secret History of Disinformation and Political Warfare*. First Picador paperback edition. New York: Picador.

Robic, Alex. 2024. “Uncontrolled Social Media Sites: Vectors for Misinformation.” *Ready Room*. https://www.readyroom.online/post/uncontrolled-social-media-sites-vectors-for-misinformation.

Sanger, David E. 2019. *The Perfect Weapon: War, Sabotage, and Fear in the Cyber Age*. First paperback edition. New York: Broadway Books.

Sardarizadeh, Shayan. 2022. “Ukraine War: False TikTok Videos Draw Millions of Views.” *BBC*. https://www.bbc.com/news/60867414.

Schlott, Rikki. 2023. “China Is Hurting Our Kids with TikTok but Protecting Its Own Youth with Douyin.” *New York Post*. https://nypost.com/2023/02/25/china-is-hurting-us-kids-with-tiktok-but-protecting-its-own/.

Schmitt, Michael N., ed. 2017. *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. 2nd ed. Cambridge University Press. doi:10.1017/9781316822524.

Schuster, Tal, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. “The Limitations of Stylometry for Detecting Machine-Generated Fake News.” *Computational Linguistics* 46(2): 499–510. doi:10.1162/coli_a_00380.

Stephan, John J. 1973. “The Tanaka Memorial (1927): Authentic or Spurious?” *Modern Asian Studies* 7(4): 733–45.

US Senate. 2018. “Report on Russian Active Measures.” https://www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures.

Vavra, Shannon. 2020. “Someone Duped Twitter Verification to Spread Racist Disinformation on US Coronavirus Vaccine.” *Cyberscoop*. https://cyberscoop.com/twitter-verification-racist-disinformation-coronavirus-vaccine-iranian-endless-mayfly/.

Yates, Jeff, Kaleigh Rogers, and Roberto Rocha. 2019. “How a Suspected Iran-Based Campaign Tried to Get Canadian Media to Spread Fake News.” https://www.cbc.ca/news/science/how-a-suspected-iran-based-campaign-tried-to-get-canadian-media-to-spread-fake-news-1.5143913.