Analyzing the Usability of Natural Language Processing for Detecting Disinformation Tactics, Techniques, and Procedures

by

Helen Landwehr

B.S., Computer Science, US Air Force Academy (2020)

Submitted to the The Institute for Data, Systems, and Society and Department of Political Science
in partial fulfillment of the requirements for the degrees of

Master of Science in Technology and Policy

and

Master of Science in Political Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Technology and Policy Program
and
Political Science
May 6, 2022

Certified by .....
Una-May O'Reilly
Principal Research Scientist
Thesis Supervisor

Certified by .....
Kenneth Oye
Professor of Political Science and Professor of Data, Systems and Society
Thesis Supervisor

Accepted by .....
Noelle Eckley Selin
Professor, Institute for Data, Systems, and Society and
Department of Earth, Atmospheric and Planetary Sciences
Director, Technology and Policy Program

Accepted by .....
Fotini Christia
Ford International Professor in the Social Sciences and
Director of Graduate Studies in Political Science

Analyzing the Usability of Natural Language Processing for Detecting Disinformation Tactics, Techniques, and Procedures

by

Helen Landwehr

Submitted to the The Institute for Data, Systems, and Society and Department of Political Science
on May 6, 2022, in partial fulfillment of the
requirements for the degrees of
Master of Science in Technology and Policy
and
Master of Science in Political Science

Abstract

The purposeful manipulation of information for political gain by powerful state actors is a threat to security and democracy that is challenging to address without infringing upon freedom of expression. By qualitatively analyzing the evolution of Russian media manipulation in the early 21st century, we see that the threat is not a product of, but is exacerbated by, technology such as social media, which increases the speed and reach of malicious information. State strategies for information manipulation co-evolve with internet and communication technology to take advantage of the new platform affordances of social media. We analyze the history of international disinformation policy in the European Union and find that policies fail because they attempt to regulate based on the effect of information manipulation rather than developing tractable definitions and characterizations of illicit information manipulation. As such, this thesis proposes that the persistence of this threat to information security is not primarily a result of technological advancements but rather a failure of policy to adequately define information manipulation. Also, we build a protype, machine learning enabled pipeline to investigate the capabilities and limits of using software techniques to characterize disinformation in a standardized manner. This pipeline offers speed and consistency to process large volumes of disinformation texts. Results indicate that even a prototype of a pipeline can detect important characteristics of disinformation. Standardized characterization of disinformation generated by pipelines such as this prototype could then potentially be used to build legal precedents, supporting a quilt-work policy approach. A technology enabled policy solution is thus a potentially feasible and effective path forward to prevent and combat state-sponsored information manipulation.

Thesis Supervisor: Una-May O'Reilly
Title: Principal Research Scientist

Thesis Supervisor: Kenneth Oye
Title: Professor of Political Science and Professor of Data, Systems and Society

Acknowledgments

First and foremost, my sincerest thank you to Dr. Steve Hadfield, who selfless sacrificed his precious time to teach me to enjoy coding, coach me through graduate school applications, and encouraged me to come to Cambridge. I am also beyond grateful for the opportunity to work with MIT Lincoln Laboratory and my advisors from the lab who gave me the academic freedom to explore new topics. I am especially grateful for the Political Science Department at MIT who gave me a chance to broaden my educational horizons. Una-May and Ken, thank you for taking me in as your graduate student and guiding me throughout the entire research process. Barb, you're great! Thanks for your enthusiasm and energy, even over zoom.

Shout out to Maya, Braxton, Jon, Justin, Coen, and Juneau for being the best crew of "Bostonians" housemates out there (even though we found out we live in Cambridge). I appreciate the random debates, experimental food production, and friendship. Thank you to all my friends in TPP for being the most hardworking p-set partners, brave running buddies, and fellow adventurers. I cherish our many memories from sitting on the 4th floor to sitting at the docks. Ashley, thank you for your never-ending, unmatched thoughtfulness and kindness. Our friendship always keeps me grounded and reminds me of what really matters in life. Also, thank you to the BU Fencing Club, for welcoming me to the team and reminding me of my love for fencing.

Table of Contents

List of Figures

List of Tables

1 Introduction

The internet has undeniably increased information accessibility through platforms such as online news media and social networks. However, increased availability does not necessarily beget high quality information. In fact, lower barriers to entry for acting as an information contributor or distributor decreases the overall veracity of online content. Paradoxically, democratizing technology such as the internet both empowers individuals to seek the truth and enables those in power, such as state actors, to exert control through the purposeful manipulation of information.

At least 81 countries have spread computational propaganda.[1] Specifically, there has been evidence of election meddling in countries such as France, the UK, the US, and South Korea. Similarly, manipulation of information has been used in attempts to cover atrocities such as the downing of MH-17.[2] During times of conflict, information manipulation seeks to obfuscate the details of the war and to win over hearts and minds of populations. While the existence of information manipulation is not new, the weaponization of information in combination with globalization and reliance upon internet technology is a major cause of concern. Information manipulation, often referred to as propaganda or psychological operations, are considered a hybrid threat and part of the grey zone of warfare.[3] Clearly, the distortion of information, especially by state actors, is a serious threat.

[1] Samantha Bradshaw, Hannah Bailey, and Phillip Howard, "Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation" (Oxford, UK: Programme on Democracy & Technology), accessed April 21, 2022, demtech.oii.ox.ac.uk.
[2] Eliot Higgins, *We Are Bellingcat: An Intelligence Agency for the People* (Bloomsbury Press, 2021), 63–108.
[3] Guillem Colom-Piella, "Cyber Activities in the Grey Zone: An Overview of the Russian and Chinese Approaches," *STRATEGIES XXI International Scientific Conference The Complex and Dynamic Nature of the Security Environment*, November 5, 2020, 189–98.

In response to this, researchers from a variety of fields have intensified efforts to investigate information manipulation. Computer and data scientists are largely focused on leveraging machine learning to quickly and automatically discover, remove, or flag disinformation.[4] Meanwhile, political scientists, psychologists, and journalists are performing qualitative analysis of information manipulation to uncover trends and improve overall understanding of the phenomenon.[5] Some researchers are attempting to create standardized frameworks by which to describe disinformation such as the DISARM Framework[6] and the Media Manipulation Case Book.[7]

Although these efforts to understand and counter the harmful effects of disinformation are imperative, they are reactionary, forensic, and not preventative or anticipatory. They only treat the symptoms of a much larger problem: the international community lacks legislation to prohibit information manipulation and prescribe appropriate responses. Even at a national level, many affected nations do not yet have effective policy.[8]

Creating disinformation policy on a national or international scale is not trivial. Efforts are challenged by difficulties in proving intent and attribution, the inherent trade-off between information security and freedom of speech, the role of commercial enterprises, cultural differences in the understanding of truth and acceptable levels of information obfuscation, and the lack of detailed, standardized definitions of information manipulation typologies for policy makers to use.

---

[4] For example, Alim Al Ayub Ahmed et al., "Detecting Fake News Using Machine Learning : A Systematic Literature Review," *ArXiv:2102.04458 [Cs]*, February 8, 2021, http://arxiv.org/abs/2102.04458.
[5] For example, Thomas Rid, *Active Measures: The Secret History of Disinformation and Poltical Warfare*, First Edition (New York: Farrar, Straus and Grioux, 2020).
[6] "DISARM Foundation," accessed April 21, 2022, https://www.disarm.foundation/.
[7] "About Us," Media Manipulation Casebook, March 17, 2020, https://mediamanipulation.org/about-us.
[8] Chris Tenove, "Protecting Democracy from Disinformation: Normative Threats and Policy Responses," *The International Journal of Press/Politics* 25, no. 3 (July 1, 2020): 517–37, https://doi.org/10.1177/1940161220918740.

Broadly, information manipulation is an umbrella term which includes many forms of the purposeful alteration of facts with the intention of causing harm or generating political gains. Under the broad category of information manipulation, there exists a variety of more specific categories such as disinformation, propaganda, and psychological warfare. Even these more specific categories can be problematically vague. For example, the European Union defines disinformation as "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm." [9] Achieving consensus on what constitutes disinformation for policy purposes with this definition is nearly impossible, there is too much room for interpretation. Instead, a "quilt-work" approach[10] to policy, founded upon setting boundaries for narrowly defined types of disinformation may be helpful. To apply this approach, there must be systematic, detailed, and agreed upon characterizations of information manipulation typologies. Furthermore, there must be agreed upon methodologies for determining when certain examples meet or fail a standard. While creating these typologies is challenging, they present practical definitions upon which policy can be based to improve the overall information manipulation policy.

This thesis asserts that, although technology influences the evolution of information manipulation, the continued prevalence and exacerbated severity of disinformation is primarily due to a lack of anticipatory and effective international regulations capable of adapting to the new media communications landscape, rather than solely a symptom of the proliferation of emerging technologies such as social media. The lack effective policy is a result of our inability to predict the effects of emerging communication technology and is perpetuated by the inherent

[9] "Action Plan on Strategic Communication," archive.ph, November 23, 2016, 1, http://archive.ph/iaGkd.
[10] Xuan W Tay, "Reconstructing The Principle of Non-Intervention and Non-Interference – Electoral Disinformation, Nicaragua, and the Quilt-Work Approach," n.d., 47.

challenges in creating agreed upon definitions of information manipulation typologies that align with variations in national values. A quilt-work policy approach will help by allowing signatories to progressively prohibit specific types of information manipulation activities. Potential preliminary policies could include prohibiting governments from creating fake accounts and banning governments from publishing directly contradictory or confusing content. Additionally, there could be standard rate limits to control information spread and collaborative tracking policies to monitor state-sponsored information outlets and specific types of information. To facilitate this policy approach, standardized methodologies for characterizing disinformation are essential.

Accordingly, in the following interdisciplinary thesis, I contribute modest comparative analysis of Russian state-sponsored disinformation during three major events, the 2008 Russo-Georgian War, the 2014 annexation of Crimea, and the 2022 invasion of Ukraine, to highlight disinformation's co-evolution with technology, the advantages, and the short comings of existing descriptive frameworks for information manipulation campaigns, and the need for improved information manipulation policy. I then describe the history of the European Union's information manipulation policy from 1936 in the era of mass media to present day in the era of social media. I pay specific attention to the novel EUvsDisinfo Database and its potential for supporting disinformation research and policy. Next, I describe the construction of a prototype for a software-based, data-science oriented digital pipeline that would allow for scalable, artificial intelligence enabled, quantitative evaluation and standardized characterization of typologies of information manipulation. I provide a discussion of the prototype pipeline's capabilities and propose potential, preliminary policies for information manipulation at the international scale in accordance with the quilt-work approach. Then, I explain important qualitative and quantitative

limitations of this thesis. Finally, I present potential avenues of future work to extend the systematized characterization of disinformation.[11]

[11] The appendix will include a more detailed explanation of the coding used to compare Russian disinformation qualitatively and short-term policy recommendations for the EUvsDisinfo initiative to immediately improve the usability of the data.

2 Qualitative Comparison of Russian Information Manipulation

The following section compares the Russian information manipulation operations across time starting with the Russo-Georgian War in 2008, followed by the annexation of Crimea in 2014, and finally the Russian invasion of Ukraine in 2022. Although the use of information manipulation is not new in and of itself, there has been an evolution of its organization and strategy of employment as is the case with other, more conventional weapons. Similarly, the changes in technology which have altered global media consumption habits, have undeniably influenced how, when, and why information manipulation is used. This has created a vicious cycle in which technological evolutions of publicly available, everyday technologies, such as the internet and social media, drastically shape the strategic use of information manipulation by state actors as means to achieve political ends in times of peace and turmoil. The analysis is conducted using the framework and coding from the Media Manipulation Case Book.[12]

The three cases of information manipulation are only a few examples of a much larger problem. Russia's use of information manipulation occurs in previous eras and other nations not analyzed in these cases. Furthermore, other state actors also rely on domestic and international information manipulation in times of peace, political turmoil, and war. Therefore, this section motivates the creation of better information manipulation policy despite the obvious regulatory challenges. This section also highlights the limitations of qualitative comparison of information manipulation campaigns through manual coding. The time intensive process is subject to personal biases and competing definitional frameworks inhibit standardized comparison. Therefore, software aided methods are crucial to preventing and countering state-sponsored information manipulation.

[12] "The Media Manipulation Case Book: The Code Book," Media Manipulation Casebook, October 15, 2020, https://mediamanipulation.org/code-book.

2.1 Scope, Definitions, Data and Methodology

2.1.1 Scope and Definitions

To make meaningful observations regarding the consistencies and changes in the use of information manipulation, a sufficiently narrow and simultaneously inclusive definition must be identified. Drawing lines between that which is information manipulation and that which is not is challenging due to the typically clandestine nature of these activities, the lack of universal ground truth by which to determine veracity, and a lack of commonly agreed upon definitions of what constitutes information manipulation. Furthermore, the cases chosen for this study all coincide with conventional conflict. As such, the lines between cyber operations and information manipulation are often blurred.

This study is scoped to media manipulation defined as

> a process where actors leverage specific conditions or features within an information ecosystem in an attempt to generate public attention and influence public discourse through deceptive, creative, or unfair means. Media is a reference to artifacts of communication and not simply a description of news.[13]

Throughout this section, the term media manipulation will be used interchangeably with information manipulation and information warfare. This definition is justified because it allows for the consideration of manipulated content as well as notable efforts to control the flow of information. Especially during times of conflict, as is the case in 2008, 2014, and 2022, adversaries may attempt to gain an advantage by blocking or dominating critical information pathways.

With this definition in mind, the following section will be scoped to include activities attributed to the Russian state as well as modest analysis of how targeted nations responded in the information domain. By considering only Russian, state-sponsored activities, we can largely

[13] "The Media Manipulation Case Book: The Code Book," 3.

eliminate cross-national variation in the use of information manipulation. Furthermore, Russia has long relied upon the use of disinformation to achieve political ends. As such, their information warfare activities are often considered highly effective because of consorted effort to improve and adapt.[14] Therefore, analyzing evolutions and adaptations in the "state-of-the-art" information manipulation may be more likely to provide insight into cutting edge information warfare adaptations with technology. Finally, for the cases analyzed, the Russian motivation for the use of information warfare is highly similar. Russia appears to follow a larger strategic plan of maintaining a sizable security buffer from NATO in eastern Europe through the acquisition and control of neighboring regions, especially those with large proportions of ethnic Russians as seen in Figure 1.

Figure 1 Map of ethnic Russian populations in eastern Europe[15]

[14] Keir Giles, "Russian Information Warfare," in *The World Information War*, ed. Timothy Clack and Robert Johnson, 1st ed. (Abingdon, Oxon ; New York, NY : Routledge, [2021] | Series: Routledge advances in defence studies: Routledge, 2021), 33–36, https://doi.org/10.4324/9781003046905-12.
[15] Image from "Russian-Majority Areas Watch Moscow's Post-Crimea Moves," *BBC News*, March 26, 2014, sec. Europe, https://www.bbc.com/news/world-europe-26713975.

These cases therefore highlight emerging technology's role in the evolution of information manipulation and minimize the influence of confounding factors, such as the prevalence of kinetic operations, cross-national variation, differing strategic objectives.

2.1.2 Data and Methodology

This section relies upon qualitative analysis of secondary sources such as credible news media and scholarly articles to compare evolutions in information manipulation with respect to technological developments between 2008 and 2022. Unfortunately, primary sources such as articles of Russian state-sponsored media from each case, are not always publicly available. Therefore, analysis from researchers and journalists during and after the events will be synthesized to paint a picture of the overall information landscape for each case. Each scenario will be constrained to the timeframe between the start of Russian physical presence in the region and the signing of an accord to end the conflict.[16] This section attempts to characterize three campaigns by the most prominent features of information manipulation activities based on publicly available news and research. There are undoubtedly activities not accounted for in each case either because the activity was not detected or made public by available sources.

A general overview of the context in which the information activities occurred is provided for each case. Following this summary, the coding standards defined by the Media Manipulation Case Book[17] are applied to allow for a standardized evaluation of information manipulation strategy and tactics. A summary of the coding for each of the three cases is available in Table 1. Finally, a comparison of all three cases and a discussion of the technologically influenced evolution will summarize the observed consistencies and changes.

[16] In the case of the 2022 invasion of Ukraine, no agreement to end conflict has been signed, therefore the analysis will attempt to cover as much of the information manipulation activities as possible to April 30th, 2022.
[17] "The Media Manipulation Case Book: The Code Book."

This summary of changes can be useful for security experts, researchers, and policy makers combatting information manipulation.

The Media Manipulation Case Book breaks down a media manipulation campaign into five stages as pictured in Figure 2. For the following analysis, there will be no coding for "Campaign Planning" since there is little to no openly available data of the behind-the-scenes efforts of Russian leadership to plan and release campaigns. Most of the analysis will be comparing the campaigns themselves. Note that the field of "Campaign Adaptation" will not be coded but general campaign adaptations between cases will be discussed throughout. The Appendix Sections 7.1.1-7.1.3 includes further justification for each coding decision.

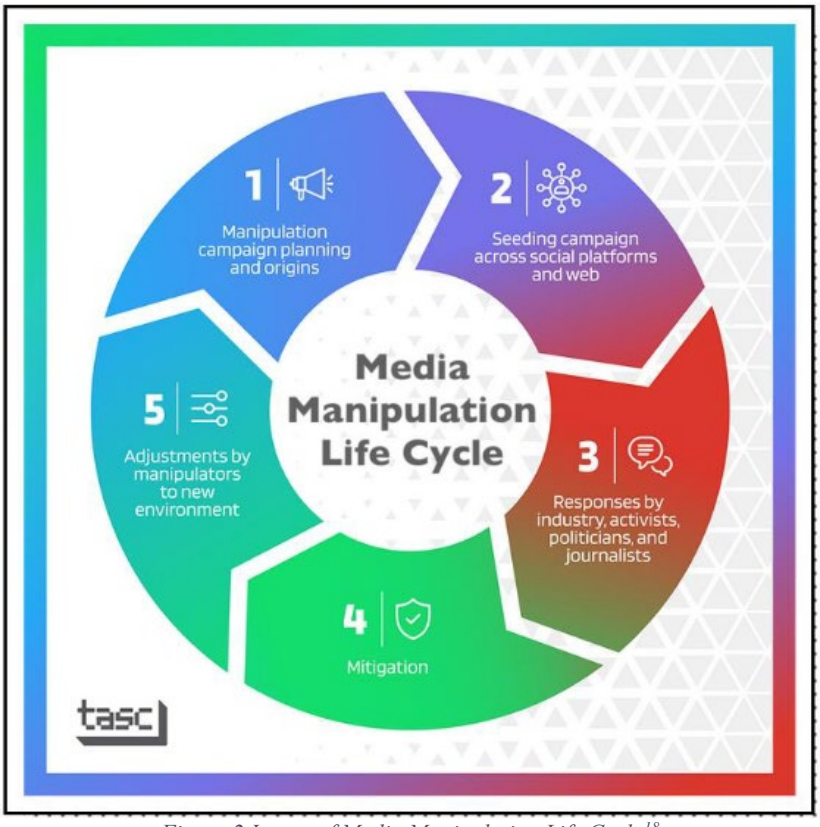

Figure 2 Image of Media Manipulation Life Cycle[18]

[18] Image from "The Media Manipulation Case Book: The Code Book," 8.

2.2 Georgia 2008

2.2.1 Summary of Conflict

The five day, Russo-Georgian War of 2008 consisted of both kinetic and information operations in Georgia to seize South Ossetia. The map in Figure 3 provides an overview of the battleground upon which the short war took place.

Figure 3 Russo-Georgia War geography[19]

There is some debate over who is responsible for starting the conflict and whether Russian intervention was justified. However, evidence suggests that the Russian actions were not entirely responsive, as was claimed by Russian media and leadership. Rather they were planned and executed after careful provocation of Georgian military forces who were suppressing separatist in South Ossetia.[20] Following the conflict, western states and international organizations, such as NATO, largely side with the Georgian perspective. According to some scholars, this is evidence that Georgia dominated the information war.[21]

[19] Image from "A Scripted War," *The Economist*, August 16, 2008, The Economist Historical Archive.
[20] Roy Allison, "Russia Resurgent? Moscow's Campaign to 'Coerce Georgia to Peace,'" *International Affairs (Royal Institute of International Affairs 1944-)* 84, no. 6 (2008): 1148.
[21] Emilio Iasiello, "Russia's Improved Information Operations: From Georgia to Crimea," *The US Army War College Quarterly: Parameters* 47, no. 2 (June 1, 2017): 57, https://press.armywarcollege.edu/parameters/vol47/iss2/7.

Russian involvement in the area and analysis of possible motivations for the outbreak of conflict are explained by Roy Allison in his article "Russia resurgent? Moscow's campaign to 'coerce Georgia to peace'". From his analysis, Russia's formal involvement in the region was established in June of 1992 when the Sochi Agreement allowed for Russia to conduct peace-keeping operations between Georgia and Ossetia. From this point onward, Russia was performing acts that could be seen as pursuing a policy of absorption such as issuing Russian passports to Ossetians. In the early 2000's, the relatively peaceful situation became more unfriendly as Georgia became increasingly aligned with NATO, especially after the election of President Saakashvili in 2004 (who had an agenda to reconsolidate Georgian territory) and the NATO summit in Bucharest in April of 2008 (which promised Georgia a plan for accession to NATO). Military buildup occurred on both sides and conflict broke out on August 8th, 2008, between Georgian forces sent by President Saakashvili to control Tskhinvali, the capital of South Ossetia, and Russian troops backing the separatists in the same area.[22] The conflict formally ended when the Russian president, Dmitry Medvedev, signed the cease-fire agreement a few days later.[23]

The events from 1992 to the breakout of the conflict in 2008 illustrate that Russia had been involved in South Ossetia for many years in an effort to maintain, if not strengthen, control over the region. The threat of Georgia re-establishing control over the area and becoming a member of NATO would damage Russian security. The fear of a loss of strength drove Russia to employ escalatory measures including conventional conflict and information manipulation. The

[22] Allison, "Russia Resurgent?"; Sarah Pruitt, "How a Five-Day War With Georgia Allowed Russia to Reassert Its Military Might," HISTORY, accessed April 28, 2022, https://www.history.com/news/russia-georgia-war-military-nato.
[23] "Russo-Georgian War," in *Wikipedia*, April 14, 2022, https://en.wikipedia.org/w/index.php?title=Russo-Georgian_War&oldid=1082738325.

conflict ended in with a cease-fire that allowed Russian forces to continue conducting peacekeeping in South Ossetia under the oversight of the EU in the positions they had held before the conflict.

2.2.2 Media Manipulation Summary

The Russo-Georgian War is one of the first conflicts to blend cyber, information, and conventional warfare.[24] Extensive examination of both the cyber-attacks and the information manipulation activities was conducted by the US Cyber Consequences Unit.[25] While the full report is not publicly available, the summary released in 2009 provides insight into the Russian operations during the war. Notably, Russian manipulation of information appears to be relatively rudimentary. Rather than winning over the hearts and minds of Georgians by posing as Georgians, information blocking was used to allow time and space for the Kremlin to spread false narratives justifying the war through official channels.[26] For example, Russian leaders legitimized the war claiming they were responding to a genocide in Georgia.[27]

Website defacement was conducted using crowd sourcing of individuals to obfuscate connections to the Kremlin. The primary goals of the website defacement were to 1) continue outsourcing and perpetuating cyberattacks and 2) to create emotional disturbance among Georgian citizens. Cyber-attacks focused on defacing important sites, attempting to take down news and media outlets to block information flow to and from Georgia. This created confusion for Georgian citizens during the crisis and slowed the response of the international community.

---

[24] Lionel Beehner et al., "Analyzing the Russian Way of War," 2008, 63.
[25] John Bumgarner and Scott Borg, "Overview by the US-CCU of the Cyber Campaign against Georgia in August of 2008" (US Cyber Consequences Unit, 2009).
[26] Beehner et al., "Analyzing the Russian Way of War," 67; Ronald J. Deibert, Rafal Rohozinski, and Masashi Crete-Nishihata, "Cyclones in Cyberspace: Information Shaping and Denial in the 2008 Russia–Georgia War," *Security Dialogue* 43, no. 1 (2012): 9.
[27] Vasile Rotaru, "'Mimicking' the West? Russia's Legitimization Discourse from Georgia War to the Annexation of Crimea," *Communist and Post-Communist Studies* 52, no. 4 (October 19, 2019): 319, https://doi.org/10.1016/j.postcomstud.2019.10.001.

According to the CCU report, five Georgian government websites and the national bank were defaced with propaganda. [28]

The information tools used by Russians during this time were blunt. Instead of nuanced manipulation of information, Russian forces mostly attempted to block information flows within and from Georgia while flooding the international community with their own narrative. Russians relied heavily upon the use of state-sponsored media sources to push a narrative that suggested Russia's involvement was humanitarian intervention as a response to Georgian aggression and atrocities.[29] These narratives were constructed using reporters on the ground and then circulated internationally to win over populations around the world.[30] A qualitative analysis of news articles published through two state-sponsored media outlets revealed that the same few sentences of text, which contained potentially exaggerated statistics regarding casualties caused by the crisis, was repeated in multiple news articles.[31]

At the time of the conflict, somewhere between 10% and 20% of Georgian citizens had internet access.[32] This likely impacted Russian strategy by pushing them to seed fake news stories via print media or television outlets instead of online. During the war, the Georgian officials demanded that Russian television channels be blocked.[33] Although popular social network sites such as Twitter and Facebook were invented prior to the conflict in 2004 and 2006 respectively, blogs were the primary online network leveraged during the information war to recruit patriotic cyber-attackers rather than to target civilians with disinformation. Infamously,

---

[28] Bumgarner and Borg, "Overview by the US-CCU of the Cyber Campaign against Georgia in August of 2008," 6.
[29] Deibert, Rohozinski, and Crete-Nishihata, "Cyclones in Cyberspace," 9.
[30] Iasiello, "Russia's Improved Information Operations," 53.
[31] Topuria Revaz, "Russia's Weapon of Words in Numbers. Evolution of Russian Assertive (Dis)Information Actions: Comparative Analysis of the Cases of Russo-Georgian War 2008 & Annexation of Crimea 2014 .," *Ante Portas*, 2020, 56.
[32] "Georgia Internet Users," accessed April 16, 2022, https://www.internetlivestats.com/internet-users/georgia/.
[33] "Georgia Cuts Access to Russian Websites, TV News," *Reuters*, August 19, 2008, sec. Internet News, https://www.reuters.com/article/us-georgia-ossetia-media-idUSLJ36223120080819.

the domain name StopGeorgia.ru was purchased and launched in conjunction with ground operations to coordinate the attacks on Georgian websites. Although the international community has not formally attributed the activity on StopGeorgia.ru, open-source intelligence reports are confident that support for the effort originated from the Kremlin.[34]

Despite being relatively rudimentary in nature, these campaigns were not adhoc. For example, evidence suggests that the graphic used by those defacing Georgian sites was developed in 2006, indicating that there was significant planning and preparation for the campaign despite the apparently rapid execution.[35] Although the use of memes was not yet widespread or popular, some defacement carried meme-like nature, such as that depicted below in Figure 4, comparing the Georgian President to Adolf Hitler.

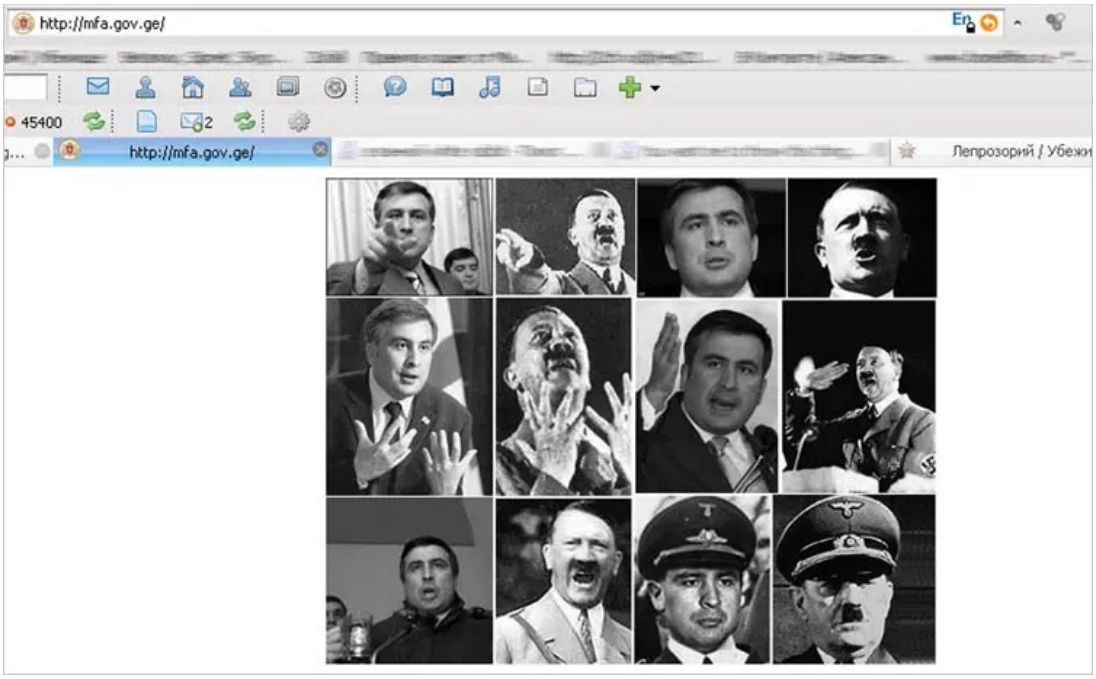

Figure 4 2008 website defacement[36]

[34] "Project Grey Goose Phase II Report: The Evolving State of Cyber Warfare" (Greylogic, March 20, 2009).
[35] Bumgarner and Borg, "Overview by the US-CCU of the Cyber Campaign against Georgia in August of 2008," 5.
[36] Image from John Markoff, "Before the Gunfire, Cyberattacks," *The New York Times*, August 12, 2008, sec. Technology, https://www.nytimes.com/2008/08/13/technology/13cyber.html.

2.3 Crimea 2014

2.3.1 Summary of Conflict

The strategic interest in Crimean territory leading up to it annexation by Russia in 2014 largely mirrored the situation of the Russo-Georgian War of 2008. Russia has long held ethnic ties to the Crimean population, just as they did South Ossetians. Additionally, Crimea provides a geographic buffer between Russia and NATO countries as well as access to the Black Sea which offers natural gas resources that help satisfy Russian energy needs.[37]

According to a Russian official who was a speaker in the lower house of parliament in July of 2014, Crimea was annexed by Ukraine in 1991.[38] This statement is contrary to the widely accepted view that Crimea was made part of Ukraine in 1954, before the dissolution of the Soviet Union, demonstrating that Russia has long felt entitled to the region despite the international community's opposing perspective. Crimea is important to Russian geopolitical stability because of its geographic location which makes it a buffer between the west and Russia. Crimea also provides access to the Black Sea. In contrast to the coordination of cyber, information and open conflict in 2008, the annexation consisted of little to no open conflict.

Despite ethnic ties to Russia, pro-EU sentiment in Ukraine was on the rise according to public surveys.[39] The president of Ukraine at the time, President Viktor F. Yanukovych, did not sign a bill to join the European Union, even though it had significant citizen support. Public outrage led to the bloody Euromaidan Protests and the flight of President Yanukovych in February of 2014. The EU and the US both supported the new, pro-west Ukrainian government

[37] John Biersack and Shannon O'Lear, "The Geopolitics of Russia's Annexation of Crimea: Narratives, Identity, Silences, and Energy," *Eurasian Geography and Economics* 55, no. 3 (May 4, 2014): 248, https://doi.org/10.1080/15387216.2014.985241.
[38] Biersack and O'Lear, 1.
[39] Biersack and O'Lear, 250.

which replaced Yanukovych. The sudden switch from a pro-Kremlin government to a widely supported pro-western government threated to diminish Russia's security buffer.

On February 27th, unmarked Russian forces moved into Crimea and began to spread across the peninsula in early March before the Kremlin official declared they have the right to invade Crimea on March 3rd.[40] The parliament of Crimea voted to secede from Ukraine on March 6th and the partnership with Russia was further formalized by overwhelming public support on March 16th.[41] The west generally condemned Russian actions leading to the annexation however, little action was taken outside of sanctions. Following the annexation of Crimea, pro-Russian separatist violence continued in the Donbass region. In September of 2014, the Minsk Protocol was signed to end violence in the region.[42]

2.3.2 Media Manipulation Summary

The first and most blatant examples information manipulation were the official statements from the Kremlin that denied sending forces into Crimea at the start of the conflict. In early 2014 there was a serious effort to convince both the Russian and international population that Crimea was rightfully Russian and a victim of misguided Ukrainian leadership. In Ukraine,

> the Russian state and media engaged in an extremely sophisticated and expensive effort to project its message in a variety of formats. Increased control over a 'free' media and the acquisition/expansion of media holdings amounts to the dissemination of Kremlin-friendly and approved coverage except for few independent outlets within Russia.[43]

[40] "Timeline: Political Crisis in Ukraine and Russia's Occupation of Crimea," *Reuters*, March 8, 2014, sec. Emerging Markets, https://www.reuters.com/article/us-ukraine-crisis-timeline-idUSBREA270PO20140308.
[41] "Ukraine - The Crisis in Crimea and Eastern Ukraine | Britannica," accessed April 17, 2022, https://www.britannica.com/place/Ukraine/The-crisis-in-Crimea-and-eastern-Ukraine.
[42] "A 5-Minute Guide to Understanding Ukraine's Euromaidan Protests," accessed April 20, 2022, https://www.opensocietyfoundations.org/explainers/understanding-ukraines-euromaidan-protests.
[43] Biersack and O'Lear, "The Geopolitics of Russia's Annexation of Crimea," 253.

Over the entire period of information manipulation, Russian media took advantage of the active crisis in Crimea, protests, the ongoing referendum, and breaking news events such as the downing of MH17.[44]

A major adaptation from the disinformation campaign of the Russo-Georgian War in 2008 was the expansion of the network terrain to include social media such as YouTube, Facebook and Twitter. The use of social media allowed for the incorporation of bots and trolls into Russian information manipulation operations to increase amplification[45] and spread disinformation to a more global community. Furthermore, social media operated in tandem with traditional media sources such as news outlets published in print, aired on TV, and posted on networks. News media could now post journalistic articles to social media communities as well as deploy new forms of attention-grabbing media, such as memes. to attract viewership and support.[46] Finally, advertisements on social media were used to sway public opinion via astroturfing campaigns.[47]

One of the most dominant Russian news media outlets, RT, joined Twitter in August of 2009.[48] Studies analyzing the information released by RT on Twitter during and after the annexation of Crimea in 2014 revealed that RT faced serious competition and was not able to control the global conversation in terms of number of tweets. However, RT had a strong impact on Russian speaking populations with the most retweets. Not all RT tweets contained disinformation, but scholars believe this is a sign of highly sophisticated information

[44] Higgins, *We Are Bellingcat: An Intelligence Agency for the People*.
[45] Todd C. Helmus, *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe*, Research Report (Rand Corporation), RR-2237-OSD (Santa Monica, Calif: RAND Corporation, 2018), 22.
[46] Bradley E Wiggins, "Crimea River: Directionality in Memes from the Russia–Ukraine Conflict," 2016, 35.
[47] Ahmed Al-Rawi and Anis Rahman, "Manufacturing Rage: The Russian Internet Research Agency's Political Astroturfing on Social Media," *First Monday*, August 16, 2020, https://doi.org/10.5210/fm.v25i9.10801.
[48] "RT (@RT_com) / Twitter," Twitter, accessed April 20, 2022, https://twitter.com/RT_com.

manipulation that intersperses false narratives with truth to gain credibility.[49] According to survey data, most Russians believed that state media, such as RT, was trustworthy in 2014.[50]

At the time, social media was still a relatively new technology. Russian efforts leveraging social media took advantage of a lack of protocol to flag and remove false and misleading claims. GRU operatives created fake accounts and spread rumors about westerners (condescendingly referred to as zapadentsy) to create distrust and fear among Ukrainians such as "Brigades of zapadentsy are now on their way to rob and kill us. It is very clear that these people hold nothing sacred."[51] To exacerbate these attacks, many social media sites had not yet implemented mechanisms for detecting and blocking bot and troll activity. Finally, governments and populations lacked general media literacy and were not well prepared to face information manipulation threats, as many of the national and international efforts to counter disinformation arose in response to the investigations which surfaced following the annexation of Crimea.[52]

Information manipulation campaigns sought to justify Russian actions through legal frameworks,[53] denying military activities,[54] and discrediting the post-revolutionary Ukrainian political party.[55] The propagation of information activities was largely state initiated but was able

[49] Yevgeniy Golovchenko, "Measuring the Scope of Pro-Kremlin Disinformation on Twitter," *Humanities and Social Sciences Communications* 7, no. 1 (December 11, 2020): 1–11, https://doi.org/10.1057/s41599-020-00659-9.
[50] Niklas Granholm, Johannes Malminen, and Gudrun Persson, "Ramifications of Russian Aggression Towards Ukraine," n.d., 33.
[51] Ellen Nakashima, "Inside a Russian Disinformation Campaign in Ukraine in 2014," *Washington Post*, December 25, 2017, sec. National Security, https://www.washingtonpost.com/world/national-security/inside-a-russian-disinformation-campaign-in-ukraine-in-2014/2017/12/25/f55b0408-e71d-11e7-ab50-621fe0588340_story.html.
[52] Golovchenko, "Measuring the Scope of Pro-Kremlin Disinformation on Twitter," 2.
[53] Rotaru, "'Mimicking' the West?"
[54] Carl Schreck, "From 'Not Us' To 'Why Hide It?': How Russia Denied Its Crimea Invasion, Then Admitted It," *Radio Free Europe/Radio Liberty*, 17:01:50Z, sec. Russia, https://www.rferl.org/a/from-not-us-to-why-hide-it-how-russia-denied-its-crimea-invasion-then-admitted-it/29791806.html.
[55] Sam Sokol, "Russian Disinformation Distorted Reality in Ukraine. Americans Should Take Note.," *Foreign Policy* (blog), accessed April 20, 2022, https://foreignpolicy.com/2019/08/02/russian-disinformation-distorted-reality-in-ukraine-americans-should-take-note-putin-mueller-elections-antisemitism/.

to rely heavily on social media and online networks to further disseminate information in Ukraine, Russia, and beyond.

In summary, Russian information manipulation in 2014 with regards to the annexation of Crimea and subsequent violence until the Minsk Protocol was cutting edge, well-coordinated, and expansive. Some scholars argue that Russian failure in the information domain for the Russo-Georgian War led to the innovation of Russian information warfare strategy and tactics.[56] Russian President, Vladimir Putin, publicly announced his effort to bolster information control in 2013 saying that he wanted to break up the Western monopoly on news media.[57]

Many of the strategies, tactics, network terrain, vulnerabilities, attribution, and targets categories of the Media Manipulation coding from 2008 carry over to the case of 2014. Although cyber-attacks were still prevalent, summaries suggest they focused on hacking accounts and shutting down critical infrastructure rather than making attempts to deface, disrupt and destroy information sources.[58] It appears Russia wanted to allow for continued information flow, especially over social media, to dominate the narratives**.** Russia still attempted to suppress dissent through blocking media outlets[59] and harassing journalists.[60] Furthermore, the state needed to rely less on recruiting individuals via web forums to amplify cyber-attacks and website defacement. Instead, social media networks allowed for more pernicious, less direct spread of malicious information.

---

[56] Iasiello, "Russia's Improved Information Operations," 51; Revaz, "Russia's Weapon of Words in Numbers. Evolution of Russian Assertive (Dis)Information Actions: Comparative Analysis of the Cases of Russo-Georgian War 2008 & Annexation of Crimea 2014 .," 38.
[57] Helmus, *Russian Social Media Influence*, 15.
[58] "Russian Cyber-Operations in Ukraine and the Implications for NATO," Canadian Global Affairs Institute, accessed May 6, 2022, https://www.cgai.ca/russian_cyber_operations_in_ukraine_and_the_implications_for_nato.
[59] "Ukraine: Police Attacked Dozens of Journalists, Medics," *Human Rights Watch* (blog), January 30, 2014, https://www.hrw.org/news/2014/01/30/ukraine-police-attacked-dozens-journalists-medics.
[60] "Russia: Halt Orders to Block Online Media," *Human Rights Watch* (blog), March 23, 2014, https://www.hrw.org/news/2014/03/23/russia-halt-orders-block-online-media.

The increase in technology penetration, such as televisions and personal computers, as well as the adoption of social media platforms, also likely inspired changes in Russian disinformation. At the time of the annexation, roughly 43% of Ukrainians were internet users.[61] Surveys of those who participated in the protests demonstrate the importance of social media for organization, as 49% of respondents got information regarding the protest via social media. Furthermore, respondents generally agreed that social media and internet news were more reliable than television.[62] This makes online news platforms and social media prime targets for Russian information manipulation. Although social media had been invented in 2008, social media usage was not yet significant enough to be a serious target.

2.4 Ukraine 2022

2.4.1 Summary of Conflict

The Russian invasion of Ukraine, much like the annexation of Crimea, is a result of continuous tension in eastern Europe dating back to the dissolution of the Soviet Union in 1989. A summary written by NPR details the military and political movements which led to the invasion. Starting in April of 2021, Russia moved 100,000 troops to the border of Ukraine. Although some troops were quickly removed, thousands remained. In November, Russian military presence was renewed. In December, President Vladimir Putin listed his demands for the removal of troops. These demands included the permanent ban of Ukraine from NATO membership. In January of the new year, the Deputy of Foreign Affairs in Russia told the United States that there were no plans to invade Ukraine. The United States and western allies attempted to negotiate with Russia on President Putin's request to ban Ukraine from NATO with little

[61] "Ukraine Internet Users," accessed April 17, 2022, https://www.internetlivestats.com/internet-users/ukraine/.
[62] "Social Networks and Social Media in Ukrainian 'Euromaidan' Protests," *Washington Post*, accessed April 17, 2022, https://www.washingtonpost.com/news/monkey-cage/wp/2014/01/02/social-networks-and-social-media-in-ukrainian-euromaidan-protests-2/.

success. Diplomates were evacuated from Ukraine on January 23rd and both NATO and US troops were readied to deploy. In February negotiations between Russia and the west continued with little success. On February 21st, Putin officially recognized the separatist regions of the Donetsk People's Republic and the Luhansk People's Republic and deployed troops for what Russia claimed is a "peace-keeping" mission. This marks what Biden considered to be the start of the invasion. More dramatically, on February 24th, Russian force made a full-scale attack on Ukrainian cities.[63] As of April of 2022, there is no treaty to end the conflict in sight.

2.4.2 Media Manipulation Summary

Russia's war with Ukraine is still ongoing at the time this thesis is being written. However, news articles, government publications, and preliminary scholarly work still bring insight into the nature of the information war. In general, the information warfare continued to expand the network terrain to more social media platforms while still channeling disinformation through state-sponsored media and government accounts while using cyber-attacks to block information flow. For example, in the early part of the conflict Russia targeted a Ukrainian television tower with a missile strike.[64] Notably, the use of multimedia, especially through TikTok, has become increasingly prevalent. Additionally, there is potential evidence that Russia coordinated with China to push pro-Kremlin narratives to the Chinese population despite tight government media controls.[65]

In the 2008 invasion of Georgia, Russia relied heavily upon the use of cyber warfare to deface and block government websites as described in Section 2.2.2. Russia did not abandon the

[63] Becky Sullivan, "Russia's at War with Ukraine. Here's How We Got Here," *NPR*, February 24, 2022, sec. World, https://www.npr.org/2022/02/12/1080205477/history-ukraine-russia.
[64] Microsoft Digital Security Unit, "An Overview of Russia's Cyberattack Activity in Ukraine," April 27, 2022, 12.
[65] Li Yuan, "How China Embraces Russian Propaganda and Its Version of the War," *The New York Times*, March 4, 2022, sec. Business, https://www.nytimes.com/2022/03/04/business/china-russia-ukraine-disinformation.html.

use of cyber-attacks against media outlets and information sources but seems to be generally less reliant upon them as their ability to manipulate, rather than cut off, information expands. One notable attack meant to block communications was the Viasat Outage, which blocked communications at the outset of the conflict.[66]

Unlike in the annexation of Crimea, official channels did not blatantly deny military buildup leading up to the war. Russian troop presence and preparation for war was used in official international channels as a threat to bring NATO countries to accept President Vladimir Putin's demands. The west was unwilling to negotiate with this demand and was better prepared than in the past to fight Putin's information war, even before the invasion. Unlike previous conflicts in which the targets of information manipulation acted in response to media manipulation, the west made efforts to "pre-bunk" by releasing sensitive intelligence about planned Russian actions including cooperation with China to avoid interfering with the 2022 Olympics[67] and the filming of a video[68] which staged Russia as the victim of a Ukrainian assault to start the war.

The Russian information manipulation eco-system has become increasingly complex and convoluted while expanding their reach through adding UK, French, and German channels[69] all of which have been blocked in the EU.[70] In a similar effort to further their reach, Russia

---

[66] "Tracking Cyber Operations and Actors in the Russia-Ukraine War," Council on Foreign Relations, accessed May 4, 2022, https://www.cfr.org/blog/tracking-cyber-operations-and-actors-russia-ukraine-war.
[67] Edward Wong and Julian E. Barnes, "China Asked Russia to Delay Ukraine War Until After Olympics, U.S. Officials Say," *The New York Times*, March 2, 2022, sec. U.S., https://www.nytimes.com/2022/03/02/us/politics/russia-ukraine-china.html.
[68] Julian E. Barnes, "U.S. Exposes What It Says Is Russian Effort to Fabricate Pretext for Invasion," *The New York Times*, February 3, 2022, sec. U.S., https://www.nytimes.com/2022/02/03/us/politics/russia-ukraine-invasion-pretext.html.
[69] "Report: RT and Sputnik's Role in Russia's Disinformation and Propaganda Ecosystem," *United States Department of State* (blog), 21, accessed March 4, 2022, https://www.state.gov/report-rt-and-sputniks-role-in-russias-disinformation-and-propaganda-ecosystem/.
[70] "EU Imposes Sanctions on State-Owned Outlets RT/Russia Today and Sputnik's Broadcasting in the EU," accessed March 4, 2022, https://www.consilium.europa.eu/en/press/press-releases/2022/03/02/eu-imposes-sanctions-on-state-owned-outlets-rt-russia-today-and-sputnik-s-broadcasting-in-the-eu/.

continues to use individual accounts of government officials to publish false narratives as they are treated differently than the government pages banned by social media sites such as Twitter.[71] Reports also indicate that bots and trolls are still a mechanism by which Russian operatives amplify pro-Kremlin narratives across platforms[72] despite efforts by social media platforms to de-platform inauthentic accounts.

The war and the information manipulation associated with the conflict is now more personal than in the past. Russian narratives target individuals, as was the case when state-sponsored outlets claimed the photograph in Figure 5 was manipulated and the Ukrainian women pictured was uninjured. Individuals and major news events are potential targets of manipulation.

Figure 5 Photo of women which Russia claims is manipulated[73]

Using social media and multimedia as a primary channel of communication during the conflict reflects modern media consumption habits as well as strategic information warfare developments. According to a 2021 survey, the most popular social media sites for Ukrainians to get news were Facebook, Youtube, and Telegram.[74] All three of these sites have significant

[71] "How Kremlin Accounts Manipulate Twitter," *BBC News*, March 19, 2022, sec. Technology, https://www.bbc.com/news/technology-60790821.
[72] Microsoft Digital Security Unit, "An Overview of Russia's Cyberattack Activity in Ukraine," 14.
[73] Image from "How Kremlin Accounts Manipulate Twitter."
[74] "Social Networks for News in Ukraine 2021," Statista, accessed May 4, 2022, https://www.statista.com/statistics/1029018/social-networks-for-news-in-ukraine/.

Russian information manipulation activities. Telegram, which was initially designed as an encrypted messaging app, is one of few social media apps still accessible to the Russian population during the conflict. Reports suggest that this may be because Russia did not have the technical means to enforce a ban.[75] As such, many Russians turn to Telegram for unbiased updates of the 2022 conflict while even the Ukrainian military uses the app to evade Russian surveillance.[76] However, the flow of Russian consumers to a single channel is convenient for Russian operatives who now need to penetrate fewer channels of communication to spread falsehoods. The New York Times reports that, "Rather than stifling Telegram, the Kremlin tries to control the narrative there, not just through its own channels but by paying for posts."[77] By using social media channels instead of known Russian state-sponsored outlets, operatives delivering disinformation avoid the sweeping bans pre-emptively put in place by the European Union. Furthermore, using social media for malicious information dissemination may maximize viewership without reducing the believability of said information and reaching those who are not even looking for news. In 2021, only 2% of Ukrainians used TikTok as their primary source for news.[78] However, this has not discouraged Russian operatives from using the platform to publish deep fakes and manipulated content while President Zelensky attempts to use the same medium to drum up international support for Ukraine.[79] The use of TikTok, which is a relatively new platform, demonstrates Russia's flexible and adaptive information manipulation strategy.

---

[75] "Russia Lifts Ban on Telegram Messaging App after Failing to Block It," *Reuters*, June 18, 2020, sec. Technology News, https://www.reuters.com/article/us-russia-telegram-ban-idUSKBN23P2FT.
[76] "Why Telegram — despite Being Rife with Russian Disinformation — Became the Go-to App for Ukrainians," *Nieman Lab* (blog), accessed May 4, 2022, https://www.niemanlab.org/2022/03/why-telegram-despite-being-rife-with-russian-disinformation-became-the-go-to-app-for-ukrainians/.
[77] Valeriya Safronova, Neil MacFarquhar, and Adam Satariano, "Where Russians Turn for Uncensored News on Ukraine," *The New York Times*, April 16, 2022, sec. World, https://www.nytimes.com/2022/04/16/world/europe/russian-propaganda-telegram-ukraine.html.
[78] "Social Networks for News in Ukraine 2021."
[79] Kyle Chayka, "Ukraine Becomes the World's 'First TikTok War,'" *The New Yorker*, March 3, 2022, https://www.newyorker.com/culture/infinite-scroll/watching-the-worlds-first-tiktok-war; "TikTok's Algorithm

2.5 Qualitative Case Study Conclusions

The three cases analyzed in this section shed light on the nature of media manipulation and its relationship with emerging technologies. The shifts in tactics and strategies of Russian information manipulation campaigns took advantage of the expanding network terrain enabled by the increasing internet penetration and the adoption of social media platforms. Globalization of information through news media sites in multiple languages may have heavily influenced the expansion of targeted communities to reach international leaders and publics. Media manipulation has continually become a more and more precise weapon as Russian operatives have transitioned from primarily blocking media and official communication channels to flooding social media, messaging, and entertainment outlets with pro-Kremlin propaganda. These messages feign objectivity and credibility while attempting to mislead, confuse, and stimulate extreme emotions of specific populations. Targeted messages have adapted to prey on human psychology with attention-grabbing designs such as memes and videos to transmit fake news to consumers as entertainment on social media platforms.

Social media platforms have allowed for the integration of mass media with social media to increase the spread of deceptive information by operatives and unknowing citizens alike. In general, the information ecosystem has become decentralized and connections to state media have been obfuscated. Like other military means that develop in response to counter-measures, Russian information manipulation evaded social media platform controls by publishing through different accounts and new platforms.

Shows Users Fake News on Ukraine War," Fortune, accessed May 4, 2022, https://fortune.com/2022/03/21/tiktok-misinformation-ukraine/.

Information manipulation also become an increasingly influential strategy as the state-sponsored perpetrators expanded from secret organizations and ad hoc hacker groups to cross national coordination. In general, Russian information manipulation forces have continuously expanded their capabilities without unlearning previous strategies. The relatively low cost of media manipulation and cyber-attacks has allowed Russia to constantly increase and expand their playbook. For example, the war on Ukraine in 2022 witnessed efforts to block communications via cyber-attacks, which was a novel technique in 2008 in addition to substantial flooding of social media with nuanced algorithmic manipulation and deepfakes. The continuous expansion of methods is visible in Table 1.

Qualitative analysis of trends in Russian disinformation sheds light on the importance of creating standards such as those proposed by the Media Manipulation Case Book for insightful, cross-case comparison of disinformation campaigns. While the manual coding of disinformation cases is useful, the process is subjective and time consuming. This inhibits the use of the results for rapid responses to ongoing disinformation campaigns, and timely policy innovation and implementation. Similarly, data collection for this analysis was limited. A lack of sources, specifically primary sources, may leave several key strategies or tactics mistakenly not accounted for in the final coding. Finally, the Media Manipulation Code Book is limited in its ability to account for the blocking of information flows with conventional and cyber attacks targeted at media infrastructure, which is often integrated with information operations in Russian military doctrine.[80]

Fortunately, the continued movement of disinformation to online platforms, especially social media networks, enables better data collection efforts, much of which are naturally

[80] Iasiello, "Russia's Improved Information Operations," 51.

publicly available. Databases of disinformation created from online news media and social networks can also include insightful meta-data as well as the text and multi-media content. Access to these primary sources will enable the study of disinformation. Larger datasets of different types of information manipulation, such as disinformation, require more efficient methods of analysis. The resource intensive time-consuming nature of manual annotation, as performed here, will be a limiting factor. The development of computationally enabled methods to perform large scale and standardized analysis of information manipulation activities is essential to preventing and combatting the abuse of information in the future.

	Russo-Georgian War 2008	Annexation of Crimea 2014	Invasion of Ukraine 2022
Region	Georgia International community Russia	Ukraine Russia International community	China International community Russia Ukraine
Date	August 8, 2008 – August 16, 2008	February 27 – September 5, 2014	February 23, 2022 - present
Strategy	Distributed amplification Reputation management Targeted harassment	Astroturfing Meme war Reputation management Targeted harassment	Astroturfing Distributed amplification Gaming an algorithm Meme war Reputation management Targeted harassment
Tactics	Bots Copypasta Misinfographic Swarming	Bots Copypasta Evidence collage Memes Recontextualized media Trolling	Advertising Bots Cheap fake Memes Recontextualized media Testimonial Viral sloganeering
Network Terrain	Media outlets State controlled media Websites	State controlled media Media outlets YouTube Facebook Twitter Websites	Facebook Media outlets Reddit State controlled media Telegram TikTok Twitter Websites YouTube
Vulnerabilities	Active crisis Lax security environment Prejudice Wedge issues	Active crisis Breaking news event Election period Inconsistent regulatory Enforcement Lax security environment Prejudice Wedge issues	Active crisis Inconsistent regulatory Enforcement Prejudice
Attribution	Networked factions State actor	Conspiracists Partisans State actor Trolls	Influencers State actor Trolls
Targets	Activist groups Individuals Political party Social identity group	Activist groups Individual Political party Social identity group	Activist groups Individual Political party Politician Social identity group
Observable Outcomes	Harassment Media exposure Muddy the waters Recognition by target	Harassment Media exposure Muddy the waters Political adoption Recognition by target	Harassment Media exposure Muddy the waters Recognition by target

Mitigation	Critical press Media blackout Research and investigation	Critical press Media blackout Research and investigation	Blocking Content removal Critical press Debunking Flagging Labeling Media blackout Research and investigation

Table 1 Summary of Media Manipulation Case Book codings

3 History of EU Disinformation Policy

The cases in Section 2 in conjunction with evidence information manipulation in previous decades such as the Cold War to more recent cases of election meddling demonstrate that information manipulation is an old problem that is continuously evolving with technology. Efforts to manipulate information continue to plague the international community due to a lack of ability for international norms and policy to deter the use of information manipulation campaigns. Next, we analyze the recent history of disinformation policies to understand how and why these policy failures occur. Our scope is limited to the European Union, in keeping with our focus on Russian information manipulation campaigns.

Policy related to disinformation began before the second World War prompted in a large part by the rise of mass media. *The International Convention Concerning the Use of Broadcasting in the Cause of Peace Broadcasting of 1936* was created by the League of Nations, the predecessor of the United Nations, and is still in effect today. A minority of current EU states (Bulgaria, Denmark, Estonia, Finland, Hungary, Ireland, Latvia, Luxembourg, Malta, and Sweden) remain as signatories.[81] Apparently motivated by the expanding use of radio as a mass medium, the convention outlines the responsibilities of signatories regarding the accuracy of information sent over broadcasts. Article I demands that nations protect their information environment and their populations and

> stop without delay the broadcasting within their respective territories of any transmission which to the detriment of good international understanding is of such a character as to

[81] "International Convention Concerning the Use of Broadcasting in the Cause of Peace - Wikipedia," accessed December 9, 2021, https://en.wikipedia.org/wiki/International_Convention_Concerning_the_Use_of_Broadcasting_in_the_Cause_of_Peace.

> incite the population of any territory to acts incompatible with the internal order or the security of a territory of a High Contracting Party.[82]

Article II proceeds to specifically define illegal information from within a territory to be that which is "an incitement either to war against another High Contracting Party or to acts likely to lead thereto."[83] Article III extends the previous two articles by adding that nations must rectify incorrect information "at the earliest possible moment and the by the most effective means."[84] Article IV and Article V are focused on establishing what information ought to be shared "to promote a better knowledge of the civilization and the conditions of life of his own country as well as of the essential features of the development of his relations with other peoples and of his contribution to the organization of peace."[85] Article VI establishes that states are responsible to apply these rules to government and autonomous media within their territories.[86] Article VII gives the authority to settle disputes to the Permanent Court of Justice.[87]

The convention thus places responsibility on governments to ensure that the information broadcast within and from their territories is accurate, does not incite war, and is not harmful to good order and peace. Governments are also responsible for correcting false information and sharing information that will further peace. Individuals or companies are not held responsible by an international body for the accuracy of their broadcasts but may be accountable to their local government, depending on national policy. The document describes the types of information which are permissible and prohibited based largely upon the effect or potential effect of the information transmitted. The first article emphasizes that information which causes harm to good

[82] "International Convention Concerning the Use of Broadcasting in the Cause of Peace Broadcasting," Pub. L. No. 4319, 186 303 (1936), 309, https://treaties.un.org/doc/Publication/UNTS/LON/Volume%20186/v186.pdf.
[83] International Convention concerning the Use of Broadcasting in the Cause of Peace Broadcasting, 309.
[84] International Convention concerning the Use of Broadcasting in the Cause of Peace Broadcasting, 309.
[85] International Convention concerning the Use of Broadcasting in the Cause of Peace Broadcasting, 309.
[86] International Convention concerning the Use of Broadcasting in the Cause of Peace Broadcasting, 311.
[87] International Convention concerning the Use of Broadcasting in the Cause of Peace Broadcasting, 311.

order is in violation of the treaty. Only later, in the third article, is a measure of accuracy of information explained. Conversely, information that should be communicated under the treaty is that which will promote knowledge, peace, and quality of life. Again, permissibility is determined by examining effects, not necessarily veracity or content.

The convention went into effect in April of 1938. In the time before and during World War II, 28 countries signed and ratified the document.[88] Notably, the Soviet Union did not ratify it while the United Kingdom did. After WWII, when the League of Nations was dissolved, the United Nations adopted the convention and it remained in effect.[89]

Shifts in the international order during the Cold War led to Soviet aligned nations to ratify the convention while western nations started to denounced the convention. During the Cold War, the Soviet Union began jamming western signals, arguing that the information being broadcast from the west was harmful to the USSR. The Soviet Union decided to ratify the convention and encouraged other Soviet aligned nations to do the same. For example, the Soviet-aligned German Democratic Republic signed and ratified the convention, claiming that no changes would have to be made to their broadcasting or blocking protocols. In response, several western-aligned states, including Australia, France, and the United Kingdom, denounced the convention.[90]

As is evident by the ongoing global concern of state-sponsored information, the convention has been relatively unsuccessful in maintaining benevolent information

[88] Elizabeth A Downey, "A Historical Survey of the International Regulation of Propaganda," n.d., 344.
[89] Björnstjern Baade, "Fake News and International Law," *European Journal of International Law* 29, no. 4 (December 31, 2018): 1366, https://doi.org/10.1093/ejil/chy071.
[90] Baade, 1367.

environments. For example, Russia is still a party to the convention[91] but is often accused of using disinformation as a weapon for political gain and to "confuse, blackmail, demoralize, subvert, and paralyze."[92] Although the treaty was meant to promote accurate broadcasting within and between nations, it did not provide a mechanism for determining what information was in violation. The definition of what is and is not harmful to "good international understanding" is not clear and is largely based on the effects of information, which are difficult to measure. It is presumed nations will agree upon what constitutes information in violation of the convention. However, as is evident by the turbulent history of convention, achieving any level of international agreement is not simple. The convention demonstrates that the threat of inaccurate information is not a new phenomenon resulting from the internet. However, initial attempts to address this threat, such as the *Broadcasting Convention*, have been ineffective.

The *United Nations (UN) Charter (1945)* binds it signatories, including all 27 EU members.[93] It sets forth principles that prohibit states from interfering with the autonomy of other states (the principle of non-intervention), and it outlines the fundamental human and political rights of individuals (the freedom of expression). Foreign disinformation has the potential to violate the principles of non-intervention. At the same time, the censoring of information, to defend national sovereignty and block disinformation spread, may infringe upon the freedom of expression.

[91] "International Convention Concerning the Use of Broadcasting in the Cause of Peace - Wikipedia."
[92] Peter Pomerantsev and Michael Weiss, "How the Kremlin Weaponizes Information, Culture and Money," n.d., 44.
[93] The EU has a collective seat as a permanent observer to the UN. "European Union and the United Nations - Wikipedia," accessed December 9, 2021, https://en.wikipedia.org/wiki/European_Union_and_the_United_Nations.

The *UN Charter* does not define the principle of non-intervention; nor does the *Vienna Convention of Laws and Treaties.*[94] In general, the principle, which is an aspect of the principle of sovereignty, is meant to prevent states from interfering with the affairs of other states. The ruling of the International Court of Justice (ICJ) in *Nicaragua* (1986)[95], directly linked actions beyond the scope of traditional military action to violations of the principle of non-intervention. This ruling set legal precedent, further solidifying the principle's definition. In the case, the ICJ found the US in violation of the principle of non-intervention for supplying insurgents in Nicaragua. The court explained,

> A prohibited intervention must . . . be one bearing on matters in which each State is permitted, by the principle of State sovereignty, to decide freely. One of these is the choice of a political, economic, social and cultural system, and the formulation of foreign policy. Intervention is wrongful when it uses methods of coercion in regard to such choices, which must remain free ones.[96]

This principle of non-intervention and the ruling of the ICJ is of importance to the evolution of disinformation policy in the EU in two ways. First, the capacity for disinformation to influence the decision of voters in democracy, as explained by the EU,[97] could be considered in violation of the non-intervention as redefined by the findings of the ICJ in *Nicaragua*. Second, the process for refining the definition of the principle of non-intervention through legal precedent and practical examples of actions taken in violation could serve as a model for defining

[94] Maziar Jamnejad and Michael Wood, "The Principle of Non-Intervention," *Leiden Journal of International Law* 22, no. 2 (June 2009): 347, https://doi.org/10.1017/S0922156509005858.
[95] Case Concerning Military and Paramilitary Activities In and Against Nicaragua (Nicaragua v. United States of America) (International Court of Justice (ICJ) June 27, 1986).
[96] *Nicaragua*, *supra* note 4, para. 205. as cited by Jamnejad and Wood, "The Principle of Non-Intervention," 348.
[97] High Representative of the Union for Foreign Affairs and Security Policy, "Joint Communication to the European Parliment, the European Council, the Council, The European Economic and Social Committee, and the Commitee of Regions: Action Plan against Disinformation" (Brussels, Belgium: European Commission, May 12, 2018), 1, https://eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf.

disinformation in practice. This process of continuous redefinition by legal precedent is termed the "quilt-work approach".[98]

The lack of explicit definition of non-intervention, despite a UN consensus on the importance of the principle, mirrors what the international community is currently experiencing regarding disinformation. Nations seem to generally agree that the spread of some types of untrue information should be prevented and regulated. This is evident by the early creation and support of the *Broadcast Convention* as well as the appearance of misinformation policies around the world.[99] However, the explicit definition of what constitutes the illegal spread of harmful information in practice is elusive. To improve the understanding of what constitutes illegal spread of information a quilt-work approach, as proposed by Xuan W. Tay, can be used in obtaining a practical and operational definition of illegal disinformation related activities. This approach, "asks that States work towards identifying and prohibiting specific manifestations, or 'sub-norms', of dangerous interventions and interferences."[100] Through this processes, instances of disinformation defined by the actions taken to create and spread inaccuracies could be more systematically found in violation of UN principles, such as the principle of non-intervention.

The *UN Charter* also guarantees rights to individual citizens in the UN. These rights protect citizens but also limit the ability of a nation's leaders to counter disinformation. In the international community at large, freedom of expression is guaranteed through *The International Bill of Human Rights*. The *International Bill of Human Rights* contains both *the Universal Declaration of Human Rights* and the *International Covenant on Civil and Political Rights*. In

[98] Tay, "Reconstructing The Principle of Non-Intervention and Non-Interference – Electoral Disinformation, Nicaragua, and the Quilt-Work Approach."
[99] "A Guide to Anti-Misinformation Actions around the World," *Poynter* (blog), accessed May 8, 2021, https://www.poynter.org/ifcn/anti-misinformation-actions/.
[100] Tay, "Reconstructing The Principle of Non-Intervention and Non-Interference – Electoral Disinformation, Nicaragua, and the Quilt-Work Approach."

the EU, the *European Charter on Fundamental Rights* reaffirms many of the freedoms promised by the United Nations, including the freedom of expression.

The *Universal Declaration of Human Rights* (1948) continues to be fundamental to the United Nations. Article 19 expresses that all humans are entitled to "hold opinions without interference and to seek, receive and impart information and ideas through any media regardless of frontiers."[101] In 1966, this position was reaffirmed in similar language by the *International Covenant on Civil and Political Rights* (ICCPR). With only slightly more specificity, the freedom of expression is described in Article 19 of the ICCPR as "freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice."[102] A notable addition to the UDHR is found in Article 20 of the ICCPR which states "Any propaganda for war shall be prohibited by law."[103] The convention fails to define propaganda for war. Still, it is important to note that information that may fuel war, such as propaganda, is explicitly mentioned and prohibited. Much like the broadcast convention, unacceptable activities are defined predominately by the potential effect of the information. However, the use of the term propaganda, which is normally understood to be persuasive and manipulative or deceptive information[104] does help to specify the more exact properties which make the information illegal.

In 2000, the EU explicitly defined the freedom of expression in Article 11 of the *European Charter on Fundamental Rights*. The charter largely draws from the language of the

[101] United Nations, "Universal Declaration of Human Rights" (1948), 5, https://www.un.org/sites/un2.un.org/files/udhr.pdf.
[102] "International Covenant on Civil and Political Rights," 2200A § (1966), 11, https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx.
[103] International Covenant on Civil and Political Rights, 11.
[104] Caroline Jack, "Lexicon of Lies," Data & Society (Data & Society Research Institute, August 9, 2017), 6, https://datasociety.net/library/lexicon-of-lies/.

UDHR and the ICCPR. These multinational declarations regarding the freedom of expression and information are essential to preserving the rights of people around the world. At the same time, they also hamper a nation's ability to block disinformation from abroad since signatories are obligated to maintain open communication and media for citizens. The charter acknowledges this tension between preserving sovereignty and protecting the rights of citizens by stating, "Combating disinformation represents a major challenge because it needs to strike the right balance between maintaining fundamental rights to freedom and security and encouraging innovation and an open market."[105]

In 2001, the *Budapest Convention on Cybercrime* was created to outline

> (i) the criminalisation of conduct ranging from illegal access, data and systems interference to computer-related fraud and child pornography; (ii) procedural law tools to investigate cybercrime and secure electronic evidence in relation to any crime; and (iii) efficient international cooperation."[106]

Disinformation, fake news, propaganda, of information operations are not specifically referenced in the main body of the convention.

In 2003, an additional protocol, ETS No. 189, was added to the *Budapest Convention* in response to "national and international law need to provide adequate legal responses to propaganda of a racist and xenophobic nature committed through computer systems,"[107] noting

[105] The European Court of Auditors, *Audit preview Information on an Upcoming Audit: EU action plan against disinformation*, March 2020, pg. 4 accessed at https://www.eca.europa.eu/lists/ecadocuments/ap20_04/ap_disinformation_en.pdf

[106] Cybercrime Convention Committee (T-CY), *The Budapest Convention on Cybercrime: benefits and impact in practice*, Strasbourg, 13 July 2020, pg.4 accessed at https://rm.coe.int/t-cy-2020-16-bc-benefits-rep-provisional/16809ef6ac

[107] "Additional Protocol to the Convention on Cybercrime, Concerning the Criminalisation of Acts of a Racist and Xenophobic Nature Committed through Computer Systems," Pub. L. No. 189 (2001), 1.

that racist and xenophobic propaganda is destabilizing to democracies. This protocol defines and criminalizes "acts of a racist and xenophobic nature committed through computer systems."[108]

Unlike the disinformation legislation in the past, the additional protocol is very specific and relies upon a narrow category, defined by the content of information rather than the effect of the information, to set boundaries. There is certainly a subset of disinformation that relies on racist and xenophobic narratives and that may fall under the definition outlined in the additional protocol of the *Budapest Convention*.[109] The explicit definition of and criminalization of racist and xenophobic propaganda is an important step in the creation of future laws, policies, and norms regarding disinformation. Even though it does not address many other types of harmful and inaccurate information, it is an example for how to practically discern what constitutes illegal information manipulation by narrowly establishing a characteristic that, when observed, clearly signals that the spread of this information is unlawful.

An update of the impacts of the convention published at the end of June in 2021 reported that 66 countries had signed the *Budapest Convention* and 11 had been invited to accede. Furthermore, the update estimates that 82% of countries worldwide have leveraged the convention as a guideline or inspiration for cybercrime legislation.[110] By contrast, only 33 nations have ratified the additional protocol dealing with racist and xenophobic propaganda.[111] The far-reaching impacts of the main text of the convention are promising. Similarly, generating an increasingly larger consensus on what constitutes the unlawful spread of information, even if

[108] Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems, 2.
[109] For detailed information about how disinformation targets and effects minorities in the EU see https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/653641/EXPO_IDA(2021)653641_EN.pdf
[110] Cybercrime Programme Office of the Council of Europe, *The global state of cybercrime legislation 2013-2021: A cursory overview*, 30 June 2021, pg. 5
[111] "Full List," Treaty Office, accessed December 9, 2021, https://www.coe.int/en/web/conventions/full-list.

it is a relatively small subset of the spectrum of harmful information, would be a step forward in the process of expanding and refining the practical understanding of unlawful use of information for future policies and rulings.

New technological capabilities enabled by the internet have sparked the need for discussion regarding how legislation created before the internet applies to criminal international actions today. The first edition of *The Tallinn Manual on International Law Applicable to Cyber Operations* was published in 2013. In 2017, the second edition, *The Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations,* was published to update the conclusions drawn earlier in the decade.[112] The manual is not authoritative, but it summarizes the policies, laws, and norms of NATO members apply to the cyber realm of conflict.

The Tallinn Manual 2.0 is examined here to understand what norms are being established and what the application of international law to disinformation could look like in the near future. Specifically, the manual attempts to summarize the international community's stance on what operations are permissible in times of peace, those which constitute a breach in international law but are not enforceable, and those which qualify as an armed attack and therefore should be subject to the Laws of Armed Conflict.

The manual does not mention disinformation but does refer to propaganda as well as psychological operations when defining an attack, explaining violations of sovereignty, clarifying prohibition of intervention, outlining conduct during war, and specifying the use of force. Neither propaganda nor psychological operations are defined within the document or its

[112] Mabel Shaw, "Guides: International and Foreign Cyberspace Law Research Guide: Tallinn Manual & Primary Law Applicable to Cyber Conflicts," accessed December 9, 2021, https://guides.ll.georgetown.edu/c.php?g=363530&p=4821482.

appendices. Simply updating the document to include more specific definitions is not yet an effective method to further policy for two reasons. First, the document is not binding, while updating the definitions may help establish norms, there is no motivations for nations to acknowledge and accept *Tallinn Manual* definitions. Secondly, there are still significant challenges in creating standardized definitions of disinformation typologies. As such, a widely accepted framework to categorize disinformation is needed to define disinformation related activities with adequate specificity. Technology enabled solutions, such as that proposed in Section 4 may make adding definitions to the *Tallinn Manual* a more realistic option in the near future.

During peacetime, the Group of Experts who compiled *The Tallinn Manual* agree that propaganda which is coercive and causes civil unrest may violate the principle of intervention but does not necessarily violate the principle of sovereignty. Within the Group of Experts, there is disagreement regarding what forms of information campaigns qualify as an intervention based on coercion. The definition of a coercive act must "have the potential for compelling the target State to engage in an action that it would otherwise not take."[113] However, since it is not always obvious what action a state would have taken, especially in the case of public decision in a democracy, some of the experts believe that each potentially coercive act must be carefully considered within its context and consequences. It may be difficult to categorize what should be considered a violation of non-intervention when the act is coercive but not forceful.

Despite the disagreement among experts regarding when information may be a breach of non-intervention, Rule 69 clearly explains that "psychological operations intended solely to

[113] *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, 2nd ed. (Cambridge: Cambridge University Press, 2017), https://doi.org/10.1017/9781316822524.

undermine confidence in a government" do not qualify as a use of force and are presumptively legal.[114] The combination of Rule 66 and Rule 69 suggest that coercing another state's action at large is illegal but undermining the public's trust in their government, which may, in the future lead to different political outcomes, is permissible.

The Tallinn Manual 2.0 also addresses in what instances psychological operations would qualify as a cyber-attack, pushing states from peace to conflict. Generally, "psychological cyber operations and cyber espionage, do not qualify as attacks,"[115] meaning that the Laws of Armed Conflict would not apply. However, there are conditions which would elevate the conduct of psychological operations to the level of an attack such as causing "severe mental suffering that are tantamount to injury"[116] against a civilian population. Actions below the threshold of severe mental suffering, even against a civilian population, are permissible under the principle of distinction during armed conflict.

The principle of distinction, which protects citizens from being used as military targets during war, is outlined in Rule 93. Propaganda and psychological operations, even when targeted at civilians, are not generally considered to violate the principle of distinction unless they breach the threshold of what constitutes an attack as described in Rule 92. Rule 93 explains,

> Certain operations directed against the civilian population are lawful. For instance, psychological operations such as dropping leaflets or making propaganda broadcasts are not prohibited even if civilians are the intended audience. In the context of cyber warfare, transmitting email messages to the enemy population urging capitulation would likewise comport with the law of armed conflict. Only when a cyber operation against civilians or civilian objects (or other protected persons and objects) rises to the level of an attack is it prohibited by the principle of distinction and those rules of the law of armed conflict that derive from the principle.[117]

[114] *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*.
[115] *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*.
[116] *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*.
[117] *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*.

Furthermore, in the context of war, Rule 100 explains that psychological operations which harm civilian morale is not a determining factor in determining if the object of an attack was a military objective because "civilian morale is not a 'military advantage.'"[118] Information manipulation that is meant to create confusion during wartime, including psychological operations, are often called "ruses of war", and are explicitly allowed by Rules 123. Finally, in times of war, Rule 136 prohibits the use of detained persons to "mount a psychological operation."[119] This restricts the conduct of psychological operations and sets expectations for how prisoners of war ought to be treated but does not prohibit psychological operations themselves.

The precision used throughout *The Tallinn Manual* goes far beyond that of existing legislation for disinformation, and it provides examples to clarify definitions. The examples of information still rely upon the potential effects or effects of the information; however, there are considerably more factors considered. For example, lines of legal and illegal information are drawn based on the context (peace or war), targets (civilian or military), and techniques (psychological operations from a detained person are illegal). These details could serve to create more powerful international regulations regarding the use of false information in the future.

3.1 Conclusions from Policy Prior to 2015

The analysis international legislation regarding disinformation applicable to the EU from 1936 to present has highlighted the following key ideas. First, disinformation has been and is of concern to governments of EU member states. Governments felt compelled to address the spread of harmful information from the beginning of mass media in 1936 onward. The threat of

[118] *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*.
[119] *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*.

inaccurate information is not new; however, it has grown more apparent and intense with new technology including the radio and the internet because of their reach and velocity of spread.

Second, the relevant policy up to this point has not waivered on the broad understanding of harmful information. Rather, the limiting factor is achieving international consensus on what is considered the unlawful spread of untrue information in practice. Policies and legislation largely presume that nations will agree on a ground truth. However, nations with opposing interests are unlikely to agree upon what information is harmful to good order and intentionally untrue. This was the case between the Soviet Union and western states in the Cold War under the *Broadcasting Convention*.[120] In a similar vein, new Russian state policy justifies punishing journalists who spread "false information" about the war in Ukraine. Simply calling the war in Ukraine a war, rather than a special military operation, could be enough grounds for punishment.[121] Furthermore, the use of the relatively poorly defined principle of non-intervention implicitly relies upon validating that a state's disinformation was coercive, yet we have no consistent method to prove the effects of disinformation were coercive. Legislation of disinformation cannot focus solely on the immeasurable effect of disinformation.

Third, there are major barriers which will continue to make building policy that prevents and defends against disinformation challenging. Nations are, and will remain, in disagreement about the ground truth in many situations. Furthermore, powerful defenses against disinformation in the EU are limited by the concern of encroaching upon human and political

[120] This problem is still apparent today, even within the smaller, more homogenous context of the EU. Individual nations are motivated to create differing disinformation responses due to the variations threat landscapes and societal norms. Polyneter provides a good overview of differences in disinformation responses. Additionally, Freedom House explains how Hungary has aggressively controlled the media landscape.

[121] Anton Troianovski, "Russia Takes Censorship to New Extremes, Stifling War Coverage," *The New York Times*, March 4, 2022, sec. World, https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html.

rights outlined in the EU and UN Charters. These challenges exist in addition to the inherent characteristics of disinformation that also make creating disinformation policy difficult as explained in the introduction.

Considering the above, a feasible, well-reasoned path forward is to focus attention on creating precise and narrow facets of quilt-work policy for disinformation. These policies should be based on regulating more than the effect of the information. They should include criteria regarding the context, targets, and methods used by the perpetrators, in addition to the intent and potential effects of the information as was done in *The Tallin Manual*. Creating policies of this type can be supported by more research and classifications of disinformation characteristics by its specific tactical methods and argumentative techniques. This would support a quilt-work approach of progressively, in a piece-by-piece way, establishing what types of information spread are in violation of international principles based on the specific details and characteristics of the information. This can be done retrospectively, with historical perspectives and rational thinking. Since we document and curate disinformation campaigns (e.g., EUvsDisinfo), we can examine these previous campaigns, characterize them narrowly, and draft legislation accordingly to prevent future campaigns of a similar nature.

This process is already naturally occurring as international definitions of disinformation grow more precise and granular. They depend more heavily on the methods of information manipulation and the content (such as racist or xenophobic content) rather than its estimated effects. The *Broadcasting Convention* was arguably the most vague conceptualization of harmful information spread. The *UN Charter* more specifically prohibited propaganda for war. The *Budapest Convention* has attempted to explicitly prohibit information of racist or xenophobic nature, and *The Tallinn Manual* outlines very finite typologies of harmful information in

different contexts that could be used to judge what information spread is in violation of international law.

A main advantage of continuing and expediting this quilt-work like processes is that it mostly avoids encroaching upon the rights of citizens. Rather than using wide-spread blocking and media censorship to prevent the dissemination of disinformation, this approach generally allows for the free flow of information to continue while searching for specific violations by state actors and prescribing punishments. Additionally, it does not necessarily require wide consensus on truth, but rather relies upon a consensus of what characterizations of information manipulation ought to be illegal and what standardized methods can be used to detect and evaluate potential information manipulation. Although policy that effectively deters information manipulation may be ideal, this thesis does not examine strictly deterrent legislation.

The following section will provide an overview of measures taken by the EU since 2015, many of which are well aligned with the general policy prescriptions described here. Following the explanation of the current EU initiatives, more specific recommendations will be given to the EUvsDisinfo project, which is one of the projects undertaken by the EU to catalog disinformation for educational and research purposes.

3.2 Explanation of EUvsDisinfo and Database

In March of 2015, the EU Council "stressed the need to challenge Russia's ongoing disinformation campaigns and invited the High Representative, in cooperation with Member States and EU institutions, to prepare by June an action plan on strategic communication."[122] By

[122] "European Council Conclusions, 19-20 March 2015," 5, accessed December 10, 2021, https://www.consilium.europa.eu/en/press/press-releases/2015/03/20/conclusions-european-council/.

June of the same year, the Action Plan on Strategic Communication was created with three primary objectives:

> 1. Effective communication and promotion of EU policies and values towards the Eastern neighbourhood; 2. Strengthening of the overall media environment including support for independent media; 3. Increased public awareness of disinformation activities by external actors, and improved EU capacity to anticipate and respond to such activities[123]

The timeline below Figure 6 shows the progression of actions taken by the EU since the creation of the East StratCom Task Force.[124] Overall, the set of actions taken since 2015 is diverse but internally and defensively focused. Much attention is paid to securing electoral integrity as well as increasing transparency, education, media literacy, and disinformation research. The East StratCom Task Force does not develop new definitions of disinformation, nor do they create rigid rules of reprisal for responding to foreign disinformation. The EU directly states their primary motivation for their response to disinformation is the threat of Russian or pro-Kremlin disinformation. For example, one initiative created by the task force, the EUvsDisinfo project, was launched as "the hub of our campaign to raise awareness of pro-Kremlin disinformation."[125] The EUvsDisinfo website considers news media to be pro-Kremlin disinformation when it is "verifiably false or misleading, according to publicly available factual evidence' and 'originates in a Kremlin funded media outlet."[126] As of 2018, EUvsDisinfo does not search all European media outlets for pro-Kremlin disinformation in response to an EEAS policy change in 2018.[127] The list of specific outlets searched for disinformation messages is not

[123] "Action Plan on Strategic Communication," 2.
[124] Image source is https://www.disinfobservatory.org/soma-officially-in-the-european-commissions-plan-to-tackle-disinformation/
[125] "'To Challenge Russia's Ongoing Disinformation Campaigns': The Story of EUvsDisinfo," EU vs DISINFORMATION, April 22, 2020, https://euvsdisinfo.eu/to-challenge-russias-ongoing-disinformation-campaigns-the-story-of-euvsdisinfo/.
[126] "Disinformation Review," EU vs DISINFORMATION, accessed December 9, 2021, https://euvsdisinfo.eu/disinfo-review/.
[127] "Disinformation Review."

published as EUvsDisinfo attempts to highlight the message of disinformation rather than the outlet.[128]

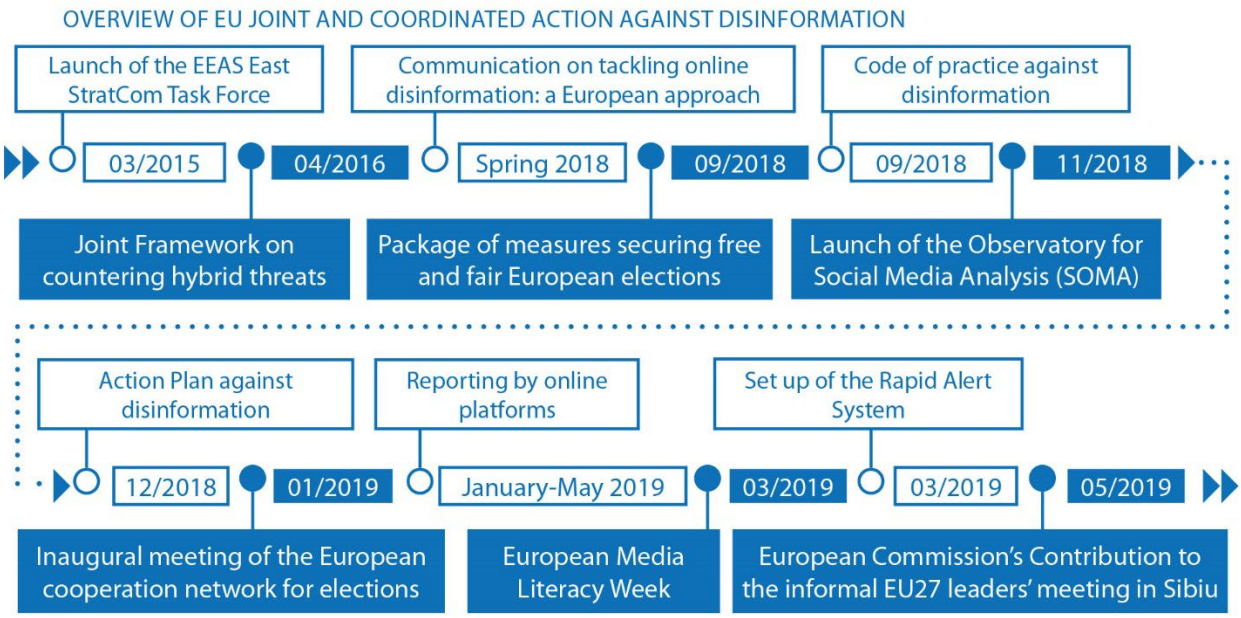

Figure 6 Timeline of actions taken by the EU[129]

In December of 2018, a comprehensive plan entitled the "Action Plan against Disinformation" was published. It extended and strengthened aspects of the previously published "Action Plan on Strategic Communications." The main goals of the "Action Plan against Disinformation" are:

> (i) improving the capabilities of Union institutions to detect, analyse and expose disinformation; (ii) strengthening coordinated and joint responses to disinformation; (iii)

[128] "Change of Terminology in the EUvsDisinfo Database," EU vs DISINFORMATION, January 24, 2018, https://euvsdisinfo.eu/change-of-terminology-in-the-euvsdisinfo-database/.
[129] Image from "SOMA Officially in the European Commission's Plan to Tackle Disinformation," *SOMA Disinfobservatory* (blog), October 4, 2019, https://www.disinfobservatory.org/soma-officially-in-the-european-commissions-plan-to-tackle-disinformation/.

> mobilising private sector to tackle disinformation; (iv) raising awareness and improving societal resilience.[130]

Within each of these pillars, there are several specific tasks outlined in Figure 7.[131] Although the detailed overview of all the initiatives which are part of this plan is outside the scope of this report, the Rapid Alert System and the Code of Practice are two ground-breaking initiatives that should be mentioned. The Rapid Alert System is a platform to allow the rapid sharing of disinformation threat information across states of the EU. The Code of Practice encourages internet platforms to accept responsibility for the implementation of counter disinformation measures to monitor and report the effects of those measures.[132]

[130] High Representative of the Union for Foreign Affairs and Security Policy, "Action Plan against Disinformation" December 5, 2018, pg. 5 available from https://eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf

[131] Image source is "Audit Preview: EU Action Plan against Disinformation," March 2020, 8, https://www.eca.europa.eu/en/Pages/DocItem.aspx?did=53299.

[132] "Code of Practice on Disinformation | Shaping Europe's Digital Future," accessed November 15, 2021, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

Pillar	Actions
I. Improving the capabilities of Union institutions to detect, analyse and expose disinformation	(1) **Strengthen** the **StratCom** task forces and EU Delegations with additional resources (human and financial) to detect, analyse and expose disinformation activities (2) **Review** of the Task Force South and Task Force Western Balkans **mandates**
II. Strengthening coordinated and joint responses to disinformation III. Mobilising private sector to tackle disinformation	(3) Establish by March 2019 a **Rapid Alert System** that works closely with other existing networks (EP, NATO and G7) (4) Step up **communication** pre-EP elections (5) Strengthen **strategic communications** in the **Neighbourhood**. (6) Close and continuous monitoring of the implementation of the **Code of Practice**, including push for rapid and effective compliance, and a comprehensive assessment after 12 months
IV. Raising awareness and improving societal resilience	(7) With Member States, organise targeted campaigns for to **raise awareness** of the negative effects of disinformation, and support work of independent media and **quality journalism** (8) Member States should support the creation of teams of multi-disciplinary independent **fact-checkers** and researchers to detect and expose disinformation campaigns (9) Promotion of **media literacy**, including through Media Literacy Week (March 2019) and rapid implementation of the relevant provisions of the Audio-visual Media Services Directive (10) Effective follow-up of the **Elections Package**, notably the Recommendation, including monitoring by the Commission of its implementation

Source: Action Plan against Disinformation, JOIN(2018) 36 final.

Figure 7 Table of pillars and actions taken in the "EU Action Plan against Disinformation"

This section provides an overview of the EUvsDisinfo, which offers the capacity to further disinformation research that could, in the long term, lead to the better creation of narrow, method-based, international disinformation policy pieces. Recommendations for its improvement are included in the Appendix Section 7.2. It should be noted that there are many trailblazing initiatives being supported under the East StratCom Task Force, but herein specific attention is being paid to this public-facing, research-oriented, educational initiative.

EUvsDisinfo launched in 2016 directly following the creation of the East StratCom Task Force in 2015.[133] The project aims to raise public awareness of the disinformation problem. The EU specifically scopes the EUvsDisinfo objectives to be "an analytical product made available publicly by the EU."[134] EUvDisinfo is meant to inform the public, journalists, and policy makers.

To carry out this mission, the project EUvsDisinfo maintains a website that is available in six languages. The most notable features of the website are the tabs for "News & Analysis", "Disinfo Review", "Disinfo Database", "Studies & Reports", "In the Media" and "Quiz & Games" which will each be reviewed in further detail below.

The "News & Analysis" section of the database and the "Disinfo Review" feature EUvsDisinfo and EEAS authored articles and reports that qualitatively analyze pro-Kremlin disinformation. There are dedicated resources to topics that are targets of disinformation such as the European elections in 2019 and the COVID-19 pandemic. To reach a wide audience, the project maintains an active social media presence of Facebook and Twitter, where they often post information about disinformation. The "Disinfo Review" lists the weekly publications released by EUvsDisinfo. These newsletters summarize "the main pro-Kremlin disinformation trends observed across the disinformation cases collected throughout the week and includes our latest news and analysis."[135] The "In the Media" section of the website links users to outside news outlets who reference the EUvsDisinfo data and reports. The "Quiz & Games" section

[133] The task force is meant to address the following three objectives according to their FAQ found at (https://www.eeas.europa.eu/eeas/questions-and-answers-about-east-stratcom-task-force_en) "Effective communication and promotion of EU policies towards the Eastern Neighbourhood, Strengthening the overall media environment in the Eastern Neighbourhood and in the EU Member States, including support for media freedom and strengthening independent media, Improved EU capacity to forecast, address and respond to disinformation activities by external actors"

[134] "'To Challenge Russia's Ongoing Disinformation Campaigns.'"

[135] "News and Analysis," EU vs DISINFORMATION, accessed May 9, 2021, https://euvsdisinfo.eu/news/.

offers several educations tools in the form of games that allow users to test their media literacy and understand how disinformation spreads.

The "Studies & Reports" section of the website is more policy and academically oriented. It features "a wide range of studies, articles and reports relating to the spread of pro-Kremlin disinformation."[136] This page is useful for policy makers and investigators who are looking for a detailed and in depth understanding of disinformation and existing disinformation policies. In general, all areas of the EUvsDisinfo website are user friendly and informative. They are easily comprehensible for users. They offer simple mechanisms which allow users to share the resources provided by EUvsDisinfo.

Finally, the most important feature of the website for the purposes of this report is the "Disinfo Database" which is a publicly available dataset of pro-Kremlin disinformation curated by the monitoring of news sources in 15 languages. The website explains that it is the "only searchable, open-source repository of its kind."[137]

The database has a friendly user interface which allows users to easily search for articles of disinformation based on the date of detection, language, country (or region) discussed in the disinformation, or keywords. Each instance of disinformation contains a disinformation headline (which describes the content disinformation), a summary of the original article, a detailed disproof of the article, an achieved link to the full article in its original language, the "Disinfo Review Issue" which analysts mentioned the article, the article's original language, and keywords associated with the article. A screen shot of one article from the database is in Figure 8

[136] "STUDIES AND REPORTS," EU vs DISINFORMATION, accessed December 9, 2021, https://euvsdisinfo.eu/reading-list/.
[137] "About," EU vs DISINFORMATION, accessed May 9, 2021, https://euvsdisinfo.eu/about/.

to provide a visual of the general layout and information available. It is important to note that below the disproof of the disinformation, EUvsDisinfo provides buttons marked with "Facebook", "Twitter", "Copy", and "Embed". These buttons streamline the process of sharing the summary and the disproof of the disinformation via social media, email, or on a website.

Figure 8 Screen shot of data in EUvsDisinfo

The database is simple, and well suited for general citizens who want to search for articles of disinformation. Beyond the publicly available dataset, EUvsDisinfo has been developing an application programming interface (API) to allow researchers to sort and query data within the Disinfo Database. This API is not yet publicly available. Access to the API was granted upon request after using the "Contact Us" function at the bottom of the EUvsDisinfo website. Once contacted, the experts who responded were extremely helpful. They provided the API, a document with the underlying database schema, further instructions for querying the API, and a copy of the data in .csv format. Those at EUvsDisinfo asked me to test the API as part of

my work regarding the study of disinformation and provide API specific feedback. Feedback specifically related to the API is available in the Appendix Section 7.2.

EUvsDisinfo makes major contributions to global disinformation research and education. It establishes credibility among researchers, policy makers, and outside news sources. It demonstrates constant evolution to meet new challenges and threats posed by disinformation, as well as expanding the reach and resources of their services.

First, the website has established itself as a credible source of metrics about disinformation as is evident from its use in news, numerous EU policy documents,[138] reports,[139] and academic research papers.[140] The wide use of EUvsDisinfo expands awareness of the initiative, strengthens media literacy, and heightens awareness of the threat of disinformation. Finally, the EUvsDisinfo website is easily accessible to the public. From a user's perspective it is easy to navigate and clearly organized. Much of the website is interconnected via links. These links provide further details and support the claims being made by the news analysts.

A second major success is that the EUvsDisinfo project has continued to expand since its creation in 2015 in both resources and reach. In 2018, EUvsDisinfo the European Parliament expanded and earmarked the budget for EUvsDisinfo to make "a systematic media monitoring service"[141] which replaces the volunteer network that was operating since 2016. EUvsDisinfo

[138] Many of the disinformation measures taken by the EU after 2015 rely on reports and statistics from EUvsDisinfo as motivation for new initiatives for example, the Democracy Action Plan available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

[139] James Pamment, "The EU's Role in Fighting Disinformation: Taking Back the Initiative," Carnegie Endowment for International Peace, accessed November 15, 2021, https://carnegieendowment.org/2020/07/15/eu-s-role-in-fighting-disinformation-taking-back-initiative-pub-82286.

[140] Svitlana Volkova and Jin Yea Jang, "Misleading or Falsification: Inferring Deceptive Strategies and Types in Online News and Social Media," in *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18* (Companion of the The Web Conference 2018, Lyon, France: ACM Press, 2018), 575–83, https://doi.org/10.1145/3184558.3188728.

[141] "'To Challenge Russia's Ongoing Disinformation Campaigns.'"

reports that the expanded network of news monitors also expanded the number of languages they can monitor. The website reports "As of 2019, our monitoring capabilities also expose disinformation spread in the Western Balkans and the EU's Southern neighbourhood."[142] As of December of 2021, the database contains 13,349 instances of pro-Kremlin disinformation. As the threats of disinformation evolve, EUvsDisinfo adapts. The website has dedicated space to disinformation concerning the COVID-19 and the Elections of 2019.

The data in the EUvsDisinfo database currently contains an abundance of information which is qualitatively analyzed and synthesized by diligent professionals at EUvsDisinfo. However, large scale, aggregated analysis of the data with a standardized framework and methodology is not yet being conducted due to time and resource constraints. Therefore, the following section explains the creation of the prototype for a software and machine learning enabled pipeline constructed to extract characteristics of disinformation in a standardized, time and resource efficient manner.

[142] "Disinformation Review."

4 Pipeline Construction, Data Description and Results

4.1 Pipeline Construction

Previous sections of this thesis provided a glimpse into the severity of the malicious information manipulation by state actors, especially as strategies co-evolve with new media technology. Still, the persistence and expansion of information manipulation on the global scale demonstrates that technology is only partially responsible for its use as a weapon of foreign policy. Historical analysis of international policies attempting to regulate the veracity of information show that our inability to generate definitional consensus for what constitutes the illegal manipulation is a critical short coming in the fight against information manipulation. The quilt-work policy approach, which is dependent upon precise definitions of illicit activities setting legal precedent over time, is a potential path forward. To enable the quilt-work approach, this section presents a pipeline powered by natural language processing (NLP) techniques to demonstrate the feasibility of using machine learning (ML) models for standardized disinformation characterization.

A ML and data science driven approach, such as the prototype pipeline, to characterize and define information manipulation is novel and potentially promising method for furthering the establishment of information manipulation norms. A competent pipeline would offer speed, objectivity, and partial transparency. ML algorithms provide the ability to rapidly analyze large volumes of data in a standardized matter. Furthermore, the models used can be partially transparent through agreed upon training data and code.

The prototype pipeline constructed for this thesis was intentionally designed to leverage the wealth of information available in the EUvsDisinfo Database[143] and the unique strengths of

[143] "DISINFO DATABASE," EU vs DISINFORMATION, accessed April 21, 2022, https://euvsdisinfo.eu/disinformation-cases/.

data science and ML approaches including speed and consistency. The data in the EUvsDisinfo Database is mostly text based, as opposed to images or videos. This calls for automated text processing and, fortuitously, state of art in NLP as a topic area in machine learning and artificial intelligence makes NLP techniques well suited to the task. Trained ML models, even if computationally expensive, are far faster than hand labeling at preforming text analysis. They allow larger quantities of disinformation articles to be analyzed, which makes it easier for researchers to observe important properties and distinctions in disinformation. Additionally, these models may be more objective than groups of individual analysts. Manual labeling of articles is almost inevitably influenced by personal biases. While these ML models inherit biases through their training data, they potentially offer standardization. These reasons motivate the exploration and assessment of the potential of using ML techniques for disinformation characterization and allow us to gauge their current limitations. The pipeline integrates three different NLP techniques: auto-translation, sentiment analysis, and argumentation analysis. An overview of the pipeline can be found in Figure 9.

Figure 9 Overview of prototype pipeline

Auto-translation is the process of translating text between two languages. In this case, all texts, from languages including Russian, Arabic, and French, are translated to English. Sentiment analysis classifies a text according to the overall feeling attitude communicated by the of the text. For example, a text which states "The horrible dictator is going to cause harm to our nation" should be labeled as negative while a phrase such as "We are grateful for our fearless leader" should be labeled as positive. Some forms of sentiment analysis classify texts into discrete categories such as positive, negative, or neutral, while other models are more granular and offer a continuous score of the sentiment. Finally, argumentation analysis or argument mining, is the "automatic identification and extraction of the structure of inference and reasoning expressed as arguments presented in natural language."[144] In the following subsections, each specific model chosen for the prototype pipeline is explained with justification for its selection as part of the overall pipeline.

4.1.1 Auto-Translation

The downstream tasks of sentiment analysis and argumentation analysis are conducted using models trained for English datasets. Therefore, I translated the original full articles into English, one of the most widely supported language for natural language processing. It is important to note that using any form of translation adds noise to the results. Translation of articles by native speakers would be more accurate, however the task of manual translation is costly and time intensive. Therefore, auto-translation is essential to making large scale disinformation analysis feasible. In real contexts, this component of the pipeline should be upgraded and enhanced to include human checking for the quality of the translation. Hand

[144] John Lawrence and Chris Reed, "Argument Mining: A Survey," *Computational Linguistics* 45, no. 4 (January 1, 2020): 765–818, https://doi.org/10.1162/coli_a_00364.

checking of translations is not an aspect of the prototype pipeline. Alternatively, future pipelines could perform downstream tasks in the original language of the disinforming text.

The Python Library googletrans was selected to translate the full texts from the original language to English because it offers free, unlimited use and has a relatively large limit of 15,000 characters per translation.[145] The underlying model of this library and Google Translate is Google's pre-trained Neural Machine Translation (GNMT).[146] This model is built to reduce computational inefficiencies while still generating natural translations of texts, even those that include rare words. According to the developers,

> Using human-rated side-by-side comparison as a metric, we show that our GNMT system approaches the accuracy achieved by average bilingual human translators on some of our test sets. In particular, compared to the previous phrase-based production system, this GNMT system delivers roughly a 60% reduction in translation errors on several popular language pairs.[147]

It is important to recognize the limitations of auto-translation. Even the most advanced auto-translation models still generate inaccurate results. The quality of a translation is especially important in the context of disinformation, where authors carefully construct an article to incite specific emotional reactions, subtly mislead readers, or take facts out of context.

4.1.2 Sentiment Analysis

Sentiment analysis "is the computational study of people's opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes."[148] For the task of analyzing the sentiment of disinformation and disinformation related texts, there are many potential NLP models

[145] SuHun Han, "Googletrans Documentation," n.d., 3.
[146] Yonghui Wu et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *ArXiv:1609.08144 [Cs]*, October 8, 2016, http://arxiv.org/abs/1609.08144.
[147] Wu et al., 20.
[148] Lei Zhang, Shuai Wang, and Bing Liu, "Deep Learning for Sentiment Analysis: A Survey," *WIREs Data Mining and Knowledge Discovery* 8, no. 4 (2018): 1, https://doi.org/10.1002/widm.1253.

available offering varying levels of granularity and robustness. To create the pipeline for this thesis, open-source models that could be easily integrated were considered. These models included the Natural Language Tool Kit Sentiment Polarity Analyzer, FLAIR, and SpaCy. For each of these libraries, the pre-trained model, which includes the stemming and tokenization steps automatically, were considered. The underlying structures of the models and their training data vary. NLTK is a very simple model, based on VADER, a bag of words approach.[149] The FLAIR model implemented relies on GloVe embeddings at the sentence level.[150] SpaCy's pretrained model was trained on a large corpus of web data and relies upon a "tok2vec" embedding at the document level.[151]

Qualitative analysis was conducted to determine which label or polarity score was most appropriate for several texts. Although the polarity score was not always correct, NLTK labeled texts appropriately more often than the other models when discrepancies between the labels produced by NTLK and SpaCy or FLAIR were compared. Another unique advantage of NLTK is the higher degree of granularity resulting from scoring each text between -1 and 1. In contrast, the other models simply label texts as positive, negative, or neutral. Therefore, the NLTK model was chosen to perform sentiment analysis for the pipeline. Still, it must be noted that there may be errors in which the model misinterprets the sentiment of a given text.

---

[149] C. Hutto and Eric Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media* 8, no. 1 (May 16, 2014): 216–25.

[150] Alan Akbik et al., "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (Minneapolis, Minnesota: Association for Computational Linguistics, 2019), 2, https://doi.org/10.18653/v1/N19-4010.

[151] "Trained Models & Pipelines · SpaCy Models Documentation," Trained Models & Pipelines, accessed April 14, 2022, https://spacy.io/models.

4.1.3 Argumentation Analysis

Within the field of argument mining of NLP, there are several models which are trained to identify various types of arguments across different lengths and structures of texts. Broadly, these models are built to find argumentative spans within texts and label the spans with a type of argument. Sometimes, these models may also classify the entire text into a broad category of text type based on the proportions and types of spans detected. Figure 10 shows the steps involved in an argument mining model.

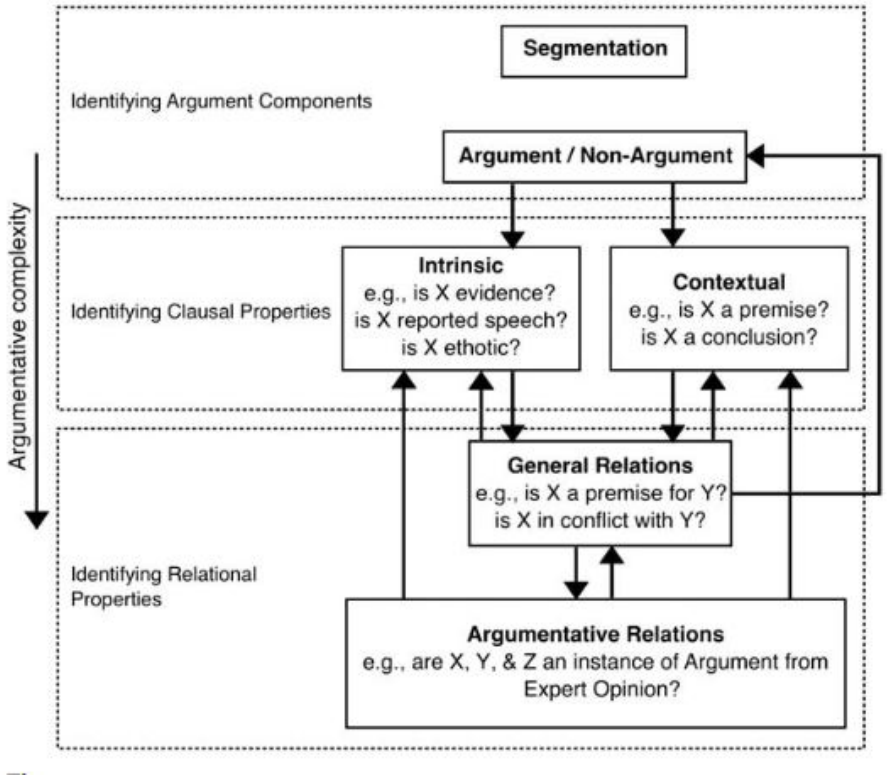

Figure 10 Overview of argument mining model[152]

One such model is named the Propaganda Persuasion Techniques Analyzer (PRTA).[153] This model detects spans of any one of the following 18 different propaganda techniques: loaded

[152] Image from Lawrence and Reed, "Argument Mining," 787.
[153] Giovanni Da San Martino et al., "Prta: A System to Support the Analysis of Propaganda Techniques in the News," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System

language, name calling / labeling, repetition, exaggeration / minimization, doubt, appeal to fear/prejudice, flag-waving, causal oversimplification, slogans, appeal to authority, black-and-white fallacy, obfuscation / intentional vagueness / confusion, thought terminating cliches, whataboutism, reduction ad hitlerum, red herring, bandwagon, and strawman. After identifying the argumentative spans in the text, PRTA also classifies the entire text as propaganda or not.

This model was chosen for the pipeline because it offers a high degree of granularity, and it is specifically trained for disinformation related texts. A high degree of granularity is essential to the quilt-work policy approach described in Section 3. Access to the API and unlimited usage privileges for PRTA was graciously granted by the Tanbih Project.[154]

Like other models, PRTA is not perfect. The task of detecting propaganda techniques within the texts is not trivial and it is reasonable to expect errors in performance. There are spans of text labeled incorrectly. For instance, when test sentences for each of the 18 techniques were passed through the model, not all the expected spans were appropriately labeled. Further training and fine-tuning of models such as PRTA must be conducted to improve the robustness of the pipeline and the confidence in our results.

4.2 Data Description

Several hundred articles of disinformation from EUvsDisinfo were selected for analysis to demonstrate the pipeline's ability to rapidly characterize disinformation texts in a standardized manner. Subsets of disinformation articles were selected based on the topic discussed in the disinformation or the target of the disinformation. The topic discussed by the disinformation was determined using the keywords associated with each article as labeled by the news analysts at

Demonstrations, Online: Association for Computational Linguistics, 2020), 287–93, https://doi.org/10.18653/v1/2020.acl-demos.32.

[154] "Tanbih Project," PRTA: A Tool for the Analysis of Propaganda Techniques in Texts, accessed April 21, 2022, https://www.tanbih.org/prta.

EUvsDisinfo. The target of the disinformation was determined by the language of the original article, which was also labeled by the news analysts at EUvsDisinfo.

The topics selected were “climate”, “coronavirus”, and “Nazi/Facist”. These topics were selected because they are likely to have little overlap between categories. Furthermore, each topic may rely on different argumentation analysis since climate and coronavirus disinformation is closely tied to science while Nazi disinformation is historical and highly political. The targets of disinformation selected for comparison were Arabic, French, and Russian. These languages are found in generally geographically distinct regions which have notable cultural differences and maintain vastly different relationships with the Kremlin. Additionally, there were many articles in EUvsDisinfo tagged with each of these three languages. We expect to be able to observe differences in disinformation characteristics by target.

4.3 Pipeline Results

Table 2 provides an overview of each selection of data as well as general results from the NLP pipeline including the average sentiment scores and the most common argument technique detected. Following Table 2, there is a more detailed visualizations and discussions of the overall results from the prototype pipeline.

Data Group	Group By	Number of Articles	Date Range of Articles (YY-MM-DD)	Most Common Keyword	Most Common Language	Average Sentiment Score	Top Two Most Common Argument Technique
Climate	Keyword / Topic	43	17-06-04 to 21-07-23	Conspiracy Theory	Russian	-0.0378	Loaded Language Doubt
COVID	Keyword / Topic	127	20-11-25 to 21-07-17	Vaccination	Russian	-0.2811	Loaded Language Doubt
Nazi	Keyword / Topic	133	20-03-02 to 21-07-30	WWII	Russian	-0.3422	Loaded Language Name Calling Labeling
Arabic	Language / Target	99	19-06-23 to 21-06-09	Conspiracy Theory	Arabic	-0.2260	Loaded Language Appeal to Fear of Prejudice
French	Language / Target	93	17-05-07 to 21-07-15	Crimea	French	-0.2301	Loaded Language Doubt
Russian	Language / Target	141	21-04-23 to 21-07-30	Anti-Russian	Russian	-0.2810	Loaded Language Doubt

Table 2 Summary of data

Visualizations of the sentiment distributions for each subset of data demonstrates pro-Kremlin disinformation's polarizing nature. In Figures 11-14, each data point represents a single, full article. Across topics and targets, as operationalized by keywords and language respectively, the full articles are highly emotive (strongly positive or negative).[155] This pattern of highly emotive content aligns with past observations of disinformation. EUvsDisinfo explains that pro-Kremlin disinformation

> will try to find those issues in our societies that garner most emotions around them, and it will try to fuel and amplify these emotions as far as possible – because an audience shaken by strong emotions will behave more irrationally and will be easier to manipulate.[156]

The results from the pipeline, which corroborate the findings of scholars, demonstrate that the protype pipeline can detect important characteristics of disinformation.

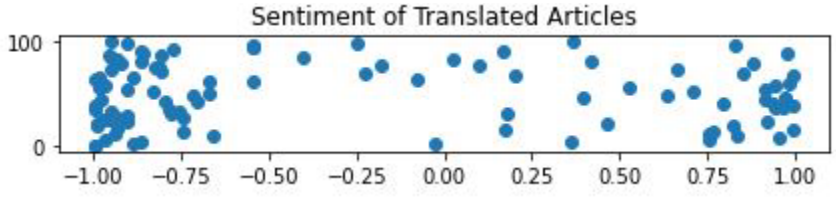

Figure 11 Arabic language text sentiment distribution

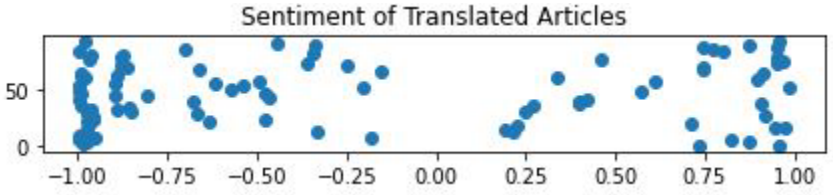

Figure 12 French language text sentiment distribution

[155] For brevity, figures of all topics and languages are not included but a polarized pattern was consistent for all segments of the EUvsDisinfo analyzed.
[156] "The Strategy and Tactics of the Pro-Kremlin Disinformation Campaign," EU vs DISINFORMATION, June 27, 2018, https://euvsdisinfo.eu/the-strategy-and-tactics-of-the-pro-kremlin-disinformation-campaign/.

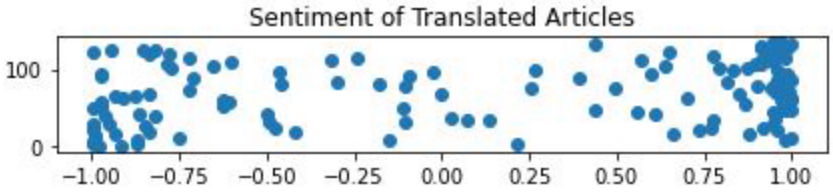

Figure 13 COVID-19 related text sentiment distribution

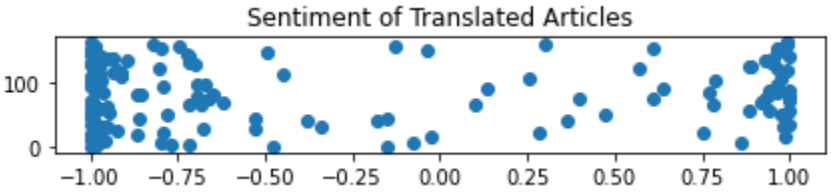

Figure 14 Nazi related text sentiment distribution

The protype pipeline also analyzed the argumentation style of disinformation across the same target and topical categories. The visualizations in Figures 15 and 16 compare proportion of each argument technique found in each subset of the data.[157]

[157] The "Proportion of Labeled Spans" is calculated by dividing the number of spans labeled with a specific technique by the total number of spans labeled with one of any technique. It is important to note that the majority of each text was labeled with "O", signifying that no propaganda technique was detected. The large amount of text unlabeled is not unsurprising. Often, disinformation texts integrate authentic journalism with falsehoods to gain readership and increase the overall credibility of the text and the source while still seeding disinformation.

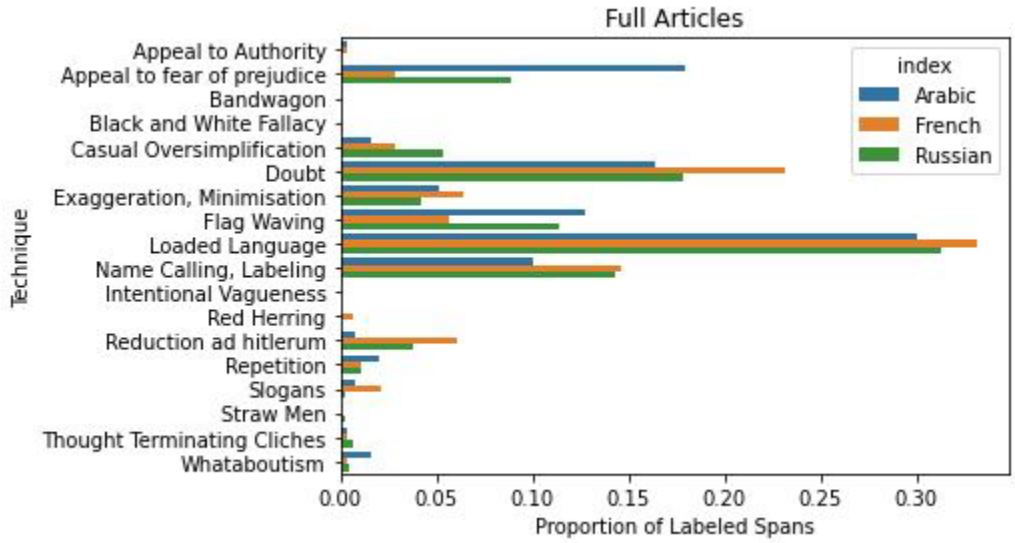

Figure 15 Proportion of text labeled with spans by target

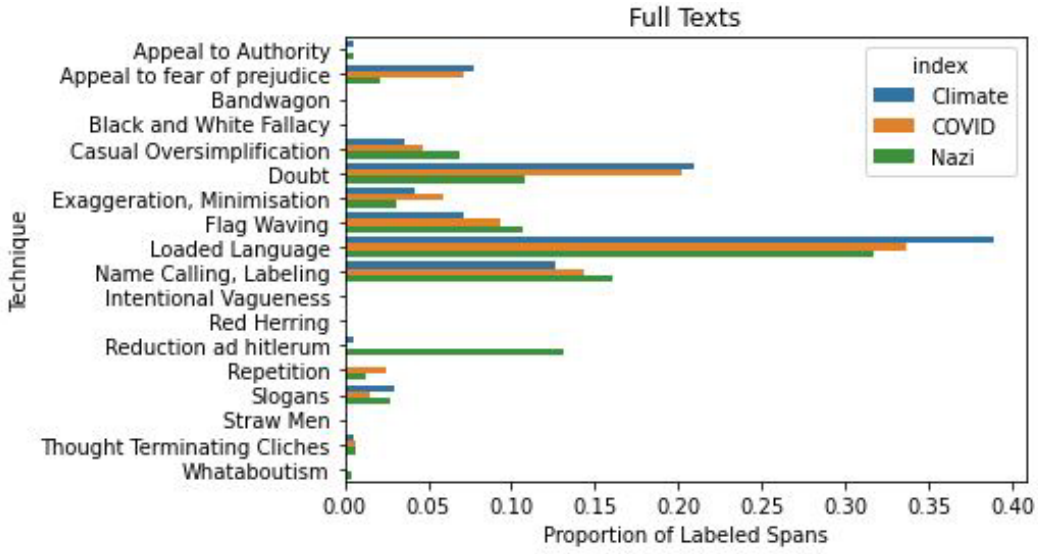

Figure 16 Proportion of text labeled with spans by topic

As was the case for the sentiment analysis results, the prototype pipeline was able to detect logical and explainable variations in pro-Kremlin disinformation. EUvsDisinfo reports that

> The campaign has different tactical aims and objectives for different audiences. It can present conspiracy theories to the audience that is ready to consume such conspiracies. It will play on pro-Russian and anti-Western feelings in one society, and exploit local national minority issues or anti-German/pan-Slavic emotions in another. It will fuel hysteria and polarisation through aggressively anti-refugee messaging or pro-refugee

> messaging (ditto anti-LGBT and pro-LGBT, or other divisive questions), to persuade both sides that those on the other side are an existential threat.[158]

In Figures 15 and 16, the tactic of exploiting different fears through argumentation technique can be seen through the appeal to fear of prejudice, which is significantly more common for disinformation targeted at Arabic speakers in the EU, a minority population. On the other hand, name calling and labeling is more prevalent in texts targeted at majority populations (French and Russian). It is highly possible that the disinformation both makes Arabic speakers fear prejudice and blatantly uses name calling against them in texts targeted at Russian and French populations.

The protype pipeline also detected variations in argumentation technique based on the topic of the disinformation, regardless of target audience. Again, these results are intuitive. For example, coronavirus and climate disinformation are heavily reliant on doubt, likely in an effort to convince the audience that the causes of climate change are unknown or unproven or the health effects of COVID-19 are unverified. Conversely, Nazi disinformation relies less on doubt and is perhaps more often trying to persuade audiences that there is, without a doubt, a neo-Nazi government in Ukraine. Instead, Nazi disinformation relies on reduction ad hitlerum, which associates actions and events with a widely hated group, such as Nazis, to inspire hatred or opposing actions.[159]

Overall, the prototype pipeline can characterize disinformation by sentiment and argumentative technique. The results align with current intuition for disinformation tactics and strategies. Realistic results suggest that pipelines like the prototype constructed here can help enable standardized characterization of disinformation. These pipelines, in combination with human oversight, can help establish legal precedent for what constitutes illegal manipulation of

[158] "The Strategy and Tactics of the Pro-Kremlin Disinformation Campaign."
[159] Giovanni Da San Martino et al., "Fine-Grained Analysis of Propaganda in News Articles," *ArXiv:1910.02517 [Cs]*, October 6, 2019, 4, http://arxiv.org/abs/1910.02517.

information. Future pipelines may expand the range of argumentation techniques labeled, detect manipulated multimedia or deepfakes, and identify microtargeting and racism in texts. It is important to note that although these pipelines offer the advantage of speed and consistency, they have notable limitations which are summarized in Section 6.

5 Potential Policy Pieces

The looming threat of information manipulation poses policy challenges because broadly defining what constitutes information manipulation, propaganda, disinformation, or fake news is difficult. Software based pipelines, such as the prototype constructed in Section 4, can enable a quilt-work policy approach by offering standardized characterizations of disinformation based on the qualities found in examples of information manipulation rather than its effects. This can allow the international community to slowly build up legal precedent for what information manipulation activities are unacceptable. The following section proposes a few potential qualities of information manipulation that could be starting points for the quilt-work approach. These starting points are 1) formally banning states from sponsoring and using fake accounts, bots, and trolls on social media platforms 2) narrowly defining at risk groups who may be targeted by or victims of racist and xenophobic content to improve the additional protocol to the *Budapest Convention* 3) banning states from targeting different identity groups with conflicting information. These starting points for the quilt-work policy approach are not exhaustive. There are many aspects of information manipulation that can be acutely characterized and slowly regulated to improve information manipulation policy.

Before detailing these starting points, it is important to mention the policy challenges posed by the use of ML enabled disinformation characterization pipelines such as the prototype in Section 4. These pipelines must be as transparent as possible to generate international consensus. In a similar way, they must be trained on an agreed upon pool of data which will necessarily include examples of state-sponsored disinformation from many states, not just Russia. Finally, the objectives of these pipelines and how they will be employed in combination with human experts and the judicial system must be considered and agreed upon. Although there are barriers to implementing technology for international legislation, it is not insurmountable.

The once novel technology of satellite imagery has been used for cases in the International Criminal Court in the *Srebrenica Trials* of the early 2000's.[160]

First, the international community should consider officially expanding the definition of information manipulation to include any information that comes from state-sponsored fake accounts, bots, or trolls. This is a feasible first step as social media platforms are currently detecting and removing fake accounts. A public private partnership as well as international treaties formally acknowledging that fake accounts are unacceptable could help improve the overall information environment.

Next, the international community should define several specific, at-risk social identity groups which cannot be targeted by racist or xenophobic content by state-sponsored media. This may help to refine the additional protocol added to the *Budapest Convention* explained in Section 2. A more refined protocol may encourage signatories to sign on and improve the information environment. Pipelines can be trained to search specifically for text that is racist, incites fear, or uses reduction ad hitlerum and flag illegal state-sponsored content.

Finally, pipelines can be developed to compare information between different state-sponsored media sites. If conflicting information is presented to different populations, operationalized by language as was the case with the prototype pipeline, the state should be investigated for information manipulation. Presenting different narratives to different populations is an indicator of manipulation to cause polarization or distress. For example, Russian operatives were detected in social network groups on both sides of the Black Lives Mater protests seeding

[160] "Satellite Imagery as Evidence for International Crimes | Coalition for the International Criminal Court," accessed May 4, 2022, https://www.coalitionfortheicc.org/news/20150423/satellite-imagery-evidence-international-crimes.

discontent.[161] Activity such as this can be narrowly defined and detected by the quality of the activity rather than the estimated impacts, which are hard to quantify.

[161] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird, "Acting the Part: Examining Information Operations within #BlackLivesMatter Discourse," *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW (2018).

6 Limitations and Future Work

6.1 Limitations

This thesis has several important limitations that must be acknowledged. Disinformation is difficult to study because actors attempt to conceal information manipulation actions. As such, the qualitative analysis comparing Russian information manipulation is limited to publicly available sources. Furthermore, the coding for the Media Manipulation Case Book was not verified by an additional analyst due to time and resource constraints. As is the case with any hand labeling, personal biases affect the final coded result.

Quantitively, the data used in this project is limited to that provided by EUvsDisinfo. The full texts of disinformation were automatically scraped using the URL provided. Articles of disinformation which were at one point available online have since been removed and were not able to be scraped. Furthermore, URL links sometimes led to articles different than that indicated by the EUvsDisinfo description. Although careful attention was paid during the data scraping and cleaning process, it is possible that some of the articles cataloged by EUvsDisinfo were mistaken for other articles due to the ephemeral nature of the internet and disinformation.

With regards to the pipeline constructed in this thesis, auto-translation of texts creates noise which may impact the accuracy of downstream tasks such as sentiment analysis and argumentation analysis. Additionally, the models used for these downstream tasks are not perfect and may fail to accurately categorize texts or label spans. However, the construction of this pipeline still demonstrates a useful methodology and avenue of research.

6.2 Future Work

There are many future research directions enabled by this work including building more extensive and robust pipelines to map aggregated disinformation text directly to disinformation frameworks such as DISARM or the Media Manipulation Case Book. To build these more robust

pipelines researchers must continue to collect and label disinformation data from a breadth of sources. A variety of different languages, topics, and text types must be cataloged to improve and expand disinformation studies. Rich and diverse examples of disinformation will be critical to training better models. Additionally, the underlying structure of these models should be continuously adapted with advancements in machine learning and natural language processing. Finally, models and pipelines should be constructed to analyze disinformation in the original language. This is an important step to improving the overall accuracy of analysis.

As more robust tools are constructed, the large-scale analysis of disinformation should continue to increase in scope to consider other actors and targets. With pipelines that can precisely characterize disinformation, researchers can examine variations in disinformation tactics, techniques, and procedures. This could enable better pre-bunking, detection, and removal. Furthermore, policy makers will be empowered to apply the quilt-work policy approach suggested in this thesis.

Machine learning pipelines may also be applied to tangential areas such as disinformation debunking procedures. The disinformation problem will likely never be fully eradicated, and it is not always possible to pre-bunk or block disinformation before it reaches the individual. Therefore, it is important that those countering disinformation are using the best possible technique to be as effective as possible. Pipelines, like that developed in this thesis, could be built to analyze the quality and standardization of disinformation summaries and disproofs, such as those available on EUvsDisinfo to improve efficacy.

7 Appendix

7.1 Detailed Explanation of Media Manipulation Coding

7.1.1 Russo-Georgia War 2008

Region: The primary region targeted by the information manipulation was **Georgia**. However, Russia state media also targeted **the international community** broadly, attempting to justify their acts of war and diminish support for Georgia. Finally, those in **Russia** were exposed to attempts to increase public support for the invasion.

Date: Although there is evidence of long-term planning, it is most appropriate to consider the dates of the campaign to be **August 8, 2008**, to align with the physical conflict and the establishment of the StopGeorgia.ru domain. Analysis ends with the end of the conflict on **August 16, 2008**, when the cease-fire was signed.

Strategy: StopGeorgia.ru and the defacement of government websites to further cyber-attacks and textbook examples of the **distributed amplification** strategy. **Reputation management** was conducted at an international scale as the Kremlin consolidated their narrative to claim that they were not responsible for the conflict and that their actions were justified. Finally, **targeted harassment** of important government sites and financial institutions was conducted by Russian actors to create chaos.

Tactics: Swarming is the most prominent tactic used during the Russo-Georgian War since much of the work of defacing websites was outsourced to individuals via the StopGeorgia.ru blog. In tandem with swarming, **bots** were employed to amplify the effects of DDoS attacks on information services. The graphic used to deface the government websites is closest to a **misinfographic**. Although it did not convey specific meaning or information, its consistent appearance and thoughtful planning is significant. Messages regarding the war often repeated themselves, which is coded as **copypasta.**

Network Terrain: Blogs, specifically StopGeorgia.ru, were used as an avenue to recruit help to further the information operations. Russian **state-controlled media outlets** attempted to reach Georgians with alternative narratives regarding the conflict, even though Georgian officials quickly blocked these channels. Similarly, **media outlets** such as CNN were attacked to stop the international flow of information used to broadcast Russia's claims regarding the war to the international community to stifle eagerness to aid. Finally, the Kremlin depended on **websites**, such as those created to further cyber-attacks and the important government sites that were attacked.

Vulnerabilities: The information manipulation directly coincided with the ongoing war and therefore exploited the **active crisis** to further the impact of operations. Additionally, it is highly probable that Russian operatives took advantage of **lax security practices** mostly because the general novelty of the campaigns and the early lack of security practices embedded in early internet use. Sensitive ethnic ties and cultural issues were used as justification for the invasion and are coded as **prejudice and wedge issues.**

Attribution: Network factions, recruited through websites, were a large part of accomplishing information manipulation activities during the conflict. More importantly, evidence suggests that most activities were planned and supported by **state actors**.

Targets: The government of Georgia, represented by the ruling **political party**, seems to be the primary target of the attacks. A secondary target was the **individuals** of Georgia, who were cut off from news sources either as a defense measure against propaganda or because of Russian cyber and physical attacks on information services. **Activists and social identity groups** who opposed the invasion were targeted and accused of committing atrocities against ethnic Russians.

Observable Outcomes: Psychological harassment of Georgians due to a lack of information and the interruption of typical services during the already chaotic kinetic war. **Media exposure** of Russian journalists sent to make false reports from Georgia helped build the Russian narrative abroad. The dearth of information in the crisis **muddied the waters**, especially for the civilian population. Despite efforts to operate secretly, Georgian officials were well-aware of propaganda and cyber-attacks justifying the coding of **recognition by the target**.

Mitigation: Georgian officials responded aggressively to information manipulation with **media blackout. Critical press**, such as official statements from Georgian leadership justifying the blockage of Russian media sites, was also an important mitigation strategy. Following the conflict there were notable **research and investigation** efforts such as that conducted by the US CCU.

7.1.2 Annexation of Crimea 2014

Region: Ukraine was the primary target of Russian disinformation to generate pro-Russian sentiment and drive a wedge between Ukraine and NATO. **Russia** also targeted its own population to justify its actions and downplay violence. Finally, the **international community** was subject to the blatant lies of President Vladimir Putin and obfuscation of events such as the downing of flight MH-17.

Date: Although there are claims of psychological warfare and information manipulation long before the unmarked invasion and annexation of Crimea, this analysis will focus on the activities from **February 27, 2014 through the signing of the vote on March 16, 2014 and continuing until the signing of the Minsk Protocol in September 5, 2014.** This is a notably longer period of substantial information manipulation activities demonstrating Russian commitment to taking permanent control of the information environment.

Strategy: To generate support for President Putin during the on-going crisis and to sway voters to act in favor of the annexation of Crimea, **astroturfing** was leveraged to create the appearance of popular support. A large part of the Russian operations was focused on **reputation management** to deny any illegal involvement and justify actions. **Meme-wars** surfaced as a mechanism for carrying malicious information disguised as

entertainment. Finally, **targeted harassment** of journalists was used to censor opposition voices.

Tactics: Bots and **trolls** were enabled by social media and leveraged throughout the crisis to amplify Russian narratives. Both bots and trolls were found to repeat text which is coded as **copypasta. Recontextualized media,** such as footage of pro-Russian protests, was used to create multi-media news stories that misrepresented actual events. **Evidence collage** was heavily utilized to deny involvement in incidents such as the initial invasion and the downing of MH-17. Finally, social media created the appropriate space and audience for **memes.**

Network Terrain: State controlled media, such as RT, was a primary source of information manipulation through social media and **websites.** Similarly, other **media outlets** spread Russian narratives, sometimes unknowingly. Russian state-controlled media's presence on social media sites such as **YouTube, Facebook, and Twitter** were important to the overall information campaign.

Vulnerabilities: Active crisis, generated by the invasion, protests, and subsequent violence, was leveraged to create and disseminate false and misleading information. Furthermore, over the long campaign, **breaking news events** such as large protests or the downing of MH-17 were used to sew confusion. The referendum which formalized the annexation of Crimea is coded as an **election period. Inconsistent social media regulations** to prevent disinformation was an easily exploited by Russian operatives. Long historical and ethnic ties allowed Russians to use **wedge issues.** A narrative of antisemitism was pervasive in justifying Russian aggression in Ukraine and is coded as **prejudice.** There was generally a **lax security environment** that was largely unprepared to deal with information manipulation contributed to the inability of the international community to pre-bunk.

Attribution: Russia, was the **state actor** responsible for a large portion of the information manipulation. They employed **trolls** on social media outlets. By seeding conspiracy theories, they employed **conspiracists** to further propagate disinformation. Similarly, those on social media in highly **partisan** online communities likely funded by the Kremlin were responsible for spreading Russian falsehoods.

Targets: Activist groups who opposed Russian invasion and occupation were targeted and even labeled as Nazis. In the same wa**y, social identity groups**, such as Jews were targeted with information meant to convince them that the Ukrainian government was antisemitic. **Individual** voters were targeted at large to sway the election and suppress anti-Russian protests. In general, the post-revolutionary Ukrainian government **(political party)** was subject of slander to destroy the populations trust.

Observable Outcomes: Conflicting narratives and confusion was certainly an achieved end of Russian information manipulation (**harassment**). Especially when the first invasion of Ukraine was denied by Russian officials, they succeeded in casting doubt among individuals and the international community, **muddying the waters.** Russian

stories were propagated through the information ecosystem, which is coded as **media exposure.** Ukraine and the international environment **recognized** manipulation efforts as evident by the **blocking of Russian news sites.** However, the referendum passed which is coded as **political adoption.**

Mitigation: The mitigation of the information manipulation was largely conducted through **media blackout** by Russia and Ukraine. Following the referendum, there was extensive **research and investigation**, much of which was used to write this report. The results of investigation generated **critical press**.

7.1.3 Invasion of Ukraine 2022

Region: The primary regions affected include **Ukraine** and **Russia,** as information manipulation seeks to win over the hearts and minds of local populations. Furthermore, this conflict saw potential coordination with **China** to distribute inauthentic material in generally isolated internet environments. The **international community** was also targeted through increasingly global communication channels in a wide array of languages.

Date: The ongoing conflict began **February 23, 2022.**

Strategy: There are many strategies employed by Russian operatives to execute effective information manipulation. In efforts to amplify content, there were reports of **astroturfing, distributed amplification** on social media, and **gaming the algorithms** with emotive and attention grabbing content to increase spread. Some of this content is **memes** and multimedia. Russian operatives clamped down on media freedom in Russia for the purposes of **reputation management**. Journalists and individual websites were reportedly **targeted and harassed** by hackers associated with the Kremlin, although this was less common than in 2008.

Tactics: Recontextualized media and cheap fakes were planned to be used to make it appear as though Ukrainians attacked Russians first in a staged video. Russia also reportedly paid for **advertising**, though many ads were quickly blocked by Google. The personalized nature of social media allows for fake **testimonial** content. **Bots** are still used by the Kremlin to amplify content, like **memes**, reach and spread. The Z symbol has become a **viral emblem/slogan** to demonstrate pro-Russian support.

Network Terrain: Despite EU bans, Russia continues to leverage **state-controlled media** sites to reach those who live in locations where bans are not in effect. **Traditional media outlets,** who may pick up well planted and disguised disinformation are also responsible for the spread of malicious content. Popular platforms heavily relied upon to spread content include **YouTube, Facebook, Twitter, Telegram, Websites, Reddit,** and **TikTok.**

Vulnerabilities: The ongoing war is an **active crisis** which makes the information environment less secure. Russian forces effectively leverage **inconsistent regulatory enforcement** by using government official's personal accounts to evade government account bans on sites such as Twitter. **Prejudice** against Ukrainians and ethnic minorities is used to fuel pro-war sentiment.

Attribution: These cases analyze Russia's actions taken as a **state actor**. The Russian state leveraged **influencers** on platforms like TikTok. Similarly, **trolls** attempt to amplify content by commenting on and reposting content online.

Targets: Activist groups in Russia and Ukraine who may wish to aid the war effort were targeted with disinformation to reduce support. Social media allows for micro-targeting of **individuals** based on algorithms. The Ukrainian government, especially President Zelensky (a **politician**) and those who are a part of and support his **political party**, were targets of manipulated information. Similarly, **social identity groups** such as ethnic Russians in Ukraine were targeted to generate pro-Russian sentiment.

Observable Outcomes: Especially in the case of the 2022 conflict, targeted nations and the international community anticipated and recognized information manipulation activities conducted by the Russian state. Despite significant efforts to pre-bunk, manipulated information has **muddied the waters**, ironically, there are even reports of President Putin's advisors sharing less than accurate information about the progress of the conflict.[162] **Media exposure** of pro-Kremlin disinformation narratives has occurred throughout the conflict both as an intentional effort to pre-bunk or de-bunk and in media sites like China. General confusion caused by the information manipulation is considered as **harassment**.

Mitigation: Ukraine and the EU executed significant **media blackout** against state-media sites although social media appears to be harder to black out. So far, researchers and journalists already appear to be conducting **research and investigation** of the scale and types of information manipulation to mitigate effects. **Content removal** has been conducted by companies like Google and Twitter. Social media companies continue to implement **blocking, flagging, and labeling.** EUvsDisinfo and other fact checking sites are **debunking.** Global news media sites are using **critical press** to call out the Kremlin.

[162] Julian E. Barnes, Lara Jakes, and John Ismay, "U.S. Intelligence Suggests That Putin's Advisers Misinformed Him on Ukraine.," *The New York Times*, March 30, 2022, sec. World, https://www.nytimes.com/2022/03/30/world/europe/putin-advisers-ukraine.html.

7.2 Recommendations for EUvsDisinfo

As previously described, the website and database maintained by EUvsDisinfo are user friendly for small-scale use cases. However, the website does not provide easy access to the complete body of data. For access to all the data and metadata associated with the catalog of pro-Kremlin disinformation, an API was provided upon request. The information for the API was provided in a link to a Google Doc which provides the API endpoint, the database schema, example queries, and links to documentation regarding the underlying systems which support the API.

The API is hosted by Sanity.io, a flexible content platform that allows for collaboration and data driven work.[163] The API relies upon the Graph-Relational Object Queries (GROQ) language. By writing queries in this language, those with access to the API can retrieve the specified data from the database. Since the database is hosted on platform which allows for real time collaboration, the queries return the most up to date content. The query language for retrieving disinformation articles, GROQ, is powerful language and is well documented by the information published through Sanity.

The structure of the data is provided by the schema, which is currently available in a bulleted list format in Google Doc. The schema contains fields for more detailed data about the article of disinformation than is available to the public on the EUvsDisinfo website. For example, the schema dedicates fields to preserving the disinformation in the original language, the time stamps of disinformation within YouTube videos in the dataset, and the domain of the publisher. The schema has also reserved fields for metrics such as the Buzzumo Number of Shares and the

[163] "The Unified Content Platform - Sanity.Io," accessed December 9, 2021, https://www.sanity.io/.

Alexa Rank, both of which could be used to trace the reach of disinformation but are still under development according to EUvsDisinfo.

Since the API is still under development, EUvsDisinfo also provided a snapshot of the data in a .csv file. The .csv file contains notably less information than that which is available through the API but is much more easily accessible since queries of the .csv can be effectively performed in programming languages such as Python,[164] which offer functionality to automatically parse, manipulate, and visualize data from .csv format.

Making the API more accessible for academic research would expand the avenues of research for disinformation to include the consistent study and monitoring of disinformation TPPs within and beyond the EU. With this knowledge, policy makers may be able to generate a fuller consensus on the unlawful use of disinformation. My specific recommendations are:

1) improve the usability of the API by creating a quick-start guide

2) improve the consistency within the database

3) create a 'For Researchers' part of the EUvsDisinfo website

4) prioritize the expansion of the database schema.

The first two of these tasks can be implemented immediately, and the third task can be accomplished in the near future, but the final task will take more time and funding to accomplish. All three should be high-priority items. I discuss each of them below.

1) **Improve the usability of the EUvsDisinfo API by creating quick-start guide**

   The API for EUvsDisinfo can broaden the community of researchers who use of the data curated by EUvsDisinfo. However, there are currently no detailed explanations of the

[164] I have conducted significant computational work in Python referencing the .csv version of the data.

database schema and the API advertised for researchers. EUvsDisinfo should immediately begin the creation of clear and succinct documentation specific to the EUvsDisinfo API.

Sections within the quick-start guide should include: a detailed and transparent description of the data collection, a graphical representation of the database schema, and example queries and responses in the GROQ query language. The guide should also link to outside documentation for the Sanity platform and the GROQ query language.

This quick-start guide will allow researchers to understand the data more quickly. Additionally, it will help researchers efficiently learn how to use the GROQ query language and the Sanity platform for EUvsDisinfo data.

2) **Improve the consistency within the EUvsDisinfo database**

Research already conducted using the API has revealed there are some inconsistencies within how the data is labeled by keywords. Based on what has been expressed by the developers at EUvsDisinfo, the keywords to label an article are selected by the news analysts. The analysts can choose to select a keyword from an existing bank or add a new keyword. The bank of keywords currently contains 619 keywords. Cases of disinformation may be labeled with multiple, relevant keywords. Because new analysts have a large degree of freedom, keywords within the bank of keywords often overlap. For example, both “WWII” and “World War II” are different keywords in the database.

Ensuring that keywords are consistent will enable better query results and downstream analysis of disinformation. Other fields of the schema, which may suffer from a similar problem, should be examined to confirm the database is as consistent as possible.

3) Create a 'For Researchers' portion of the EUvsDisinfo website

In the near future, EUvsDisinfo should create a portion of their public facing website targeted at data scientists and computer scientists who wish to perform large scale analysis of disinformation cases. The current website is extremely user friendly but is more clearly directed at citizens, journalists, and policy makers. There are many ongoing research projects of disinformation that use computational techniques and that could benefit from the EUvsDisinfo database. For example, Graphika[165] uses AI and analytical techniques for disinformation research and writes detailed reports about disinformation campaigns.

Therefore, on the "For Researchers" portion of the website, EUvsDisinfo should advertise a beta version of the API, publish any relevant computational work related to the database, and provide contact information for researchers to request API access and the quick-start guide.

Expanding the community of researchers who use of the API can generate feedback on the database schema and the API. This feedback can be leveraged to iteratively improve the EUvsDisinfo database and API. Iterative improvement of the API is optimal because

[165] "Graphika," accessed December 9, 2021, https://graphika.com/.

it allows for EUvsDisinfo and the API to respond to the evolving threats of disinformation.

4) **Prioritize the expansion the database schema for TTP**

Currently, the data available via the API contains no information regarding the TTP of the case of disinformation. The creation of a field for TTP could immensely further researchers' ability to study the underlying literary mechanisms of disinformation. This is not an easy task, and it will require time and research to implement and curate. Therefore, an immediate policy recommendation is to increase funding for EUvsDisinfo, as was seen in 2018[166] but dedicate funds to researching TTP. This falls well within the scope of the EU's stated objective of "improving the capabilities of Union institutions to detect, analyse and expose disinformation."[167]

The EUvsDisinfo initiative is well suited to lead the way on this task for several reasons. First, they already have a large dataset of disinformation which can be labeled for techniques. Second, EUvsDisinfo already performs qualitative analysis of disinformation TTP through their weekly reviews and has analysts capable of discerning TTP. Third, EUvsDisinfo data can easily be shared to other nations and initiatives not only through the API but also through partnerships with other initiatives in the Action Plan against Disinformation such as the Rapid Alert System.

[166] "'To Challenge Russia's Ongoing Disinformation Campaigns.'"
[167] High Representative of the Union for Foreign Affairs and Security Policy, "Action Plan against Disinformation" December 5, 2018, pg. 5 available from https://eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf

If successfully implemented, this could expand the use of consistent labeling schemes to describe disinformation among academic researchers. Frameworks which could be leveraged by EUvsDisinfo to label the TTP of their data include the Media Manipulation Case Book[168] or the AMITT Framework.[169] More consistent terminology and frameworks used to describe disinformation can improve the quality of disinformation research and expand the ability of policy makers to make disinformation policy supported by the content and methods of disinformation.

[168] "About Us."
[169] *AMITT Disinformation Tactics, Techniques and Processes (TTP) Framework*, Jupyter Notebook (2020; repr., Cognitive Security Collaborative, 2021), https://github.com/cogsec-collaborative/AMITT.

8 Works Cited

"A 5-Minute Guide to Understanding Ukraine's Euromaidan Protests." Accessed April 20, 2022. https://www.opensocietyfoundations.org/explainers/understanding-ukraines-euromaidan-protests.

Poynter. "A Guide to Anti-Misinformation Actions around the World." Accessed May 8, 2021. https://www.poynter.org/ifcn/anti-misinformation-actions/.

The Economist. "A Scripted War," August 16, 2008. The Economist Historical Archive.

EU vs DISINFORMATION. "About." Accessed May 9, 2021. https://euvsdisinfo.eu/about/.

Media Manipulation Casebook. "About Us," March 17, 2020. https://mediamanipulation.org/about-us.

archive.ph. "Action Plan on Strategic Communication," November 23, 2016. http://archive.ph/iaGkd.

Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems, Pub. L. No. 189 (2001).

Ahmed, Alim Al Ayub, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. "Detecting Fake News Using Machine Learning : A Systematic Literature Review." *ArXiv:2102.04458 [Cs]*, February 8, 2021. http://arxiv.org/abs/2102.04458.

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-4010.

Allison, Roy. "Russia Resurgent? Moscow's Campaign to 'Coerce Georgia to Peace.'" *International Affairs (Royal Institute of International Affairs 1944-)* 84, no. 6 (2008): 1145–71.

Al-Rawi, Ahmed, and Anis Rahman. "Manufacturing Rage: The Russian Internet Research Agency's Political Astroturfing on Social Media." *First Monday*, August 16, 2020. https://doi.org/10.5210/fm.v25i9.10801.

AMITT Disinformation Tactics, Techniques and Processes (TTP) Framework. Jupyter Notebook. 2020. Reprint, Cognitive Security Collaborative, 2021. https://github.com/cogsec-collaborative/AMITT.

Arif, Ahmer, Leo Graiden Stewart, and Kate Starbird. "Acting the Part: Examining Information Operations within #BlackLivesMatter Discourse." *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW (2018).

"Audit Preview: EU Action Plan against Disinformation," March 2020. https://www.eca.europa.eu/en/Pages/DocItem.aspx?did=53299.

Baade, Björnstjern. "Fake News and International Law." *European Journal of International Law* 29, no. 4 (December 31, 2018): 1357–76. https://doi.org/10.1093/ejil/chy071.

Barnes, Julian E. "U.S. Exposes What It Says Is Russian Effort to Fabricate Pretext for Invasion." *The New York Times*, February 3, 2022, sec. U.S. https://www.nytimes.com/2022/02/03/us/politics/russia-ukraine-invasion-pretext.html.

Barnes, Julian E., Lara Jakes, and John Ismay. "U.S. Intelligence Suggests That Putin's Advisers Misinformed Him on Ukraine." *The New York Times*, March 30, 2022, sec. World. https://www.nytimes.com/2022/03/30/world/europe/putin-advisers-ukraine.html.

Beehner, Lionel, Liam Collins, Steve Ferenzi, Robert Person, and Aaron Brantly. “Analyzing the Russian Way of War,” 2008, 98.
Biersack, John, and Shannon O’Lear. “The Geopolitics of Russia’s Annexation of Crimea: Narratives, Identity, Silences, and Energy.” *Eurasian Geography and Economics* 55, no. 3 (May 4, 2014): 247–69. https://doi.org/10.1080/15387216.2014.985241.
Bradshaw, Samantha, Hannah Bailey, and Phillip Howard. “Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation.” Oxford, UK: Programme on Democracy & Technology. Accessed April 21, 2022. demtech.oii.ox.ac.uk.
Bumgarner, John, and Scott Borg. “Overview by the US-CCU of the Cyber Campaign against Georgia in August of 2008.” US Cyber Consequences Unit, 2009.
Case Concerning Military and Paramilitary Activities In and Against Nicaragua (Nicaragua v. United States of America) (International Court of Justice (ICJ) June 27, 1986).
EU vs DISINFORMATION. “Change of Terminology in the EUvsDisinfo Database,” January 24, 2018. https://euvsdisinfo.eu/change-of-terminology-in-the-euvsdisinfo-database/.
Chayka, Kyle. “Ukraine Becomes the World’s ‘First TikTok War.’” *The New Yorker*, March 3, 2022. https://www.newyorker.com/culture/infinite-scroll/watching-the-worlds-first-tiktok-war.
“Code of Practice on Disinformation | Shaping Europe’s Digital Future.” Accessed November 15, 2021. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.
Colom-Piella, Guillem. “Cyber Activities in the Grey Zone: An Overview of the Russian and Chinese Approaches.” *STRATEGIES XXI International Scientific Conference The Complex and Dynamic Nature of the Security Environment*, November 5, 2020, 189–98.
Da San Martino, Giovanni, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. “Prta: A System to Support the Analysis of Propaganda Techniques in the News.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 287–93. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-demos.32.
Deibert, Ronald J., Rafal Rohozinski, and Masashi Crete-Nishihata. “Cyclones in Cyberspace: Information Shaping and Denial in the 2008 Russia–Georgia War.” *Security Dialogue* 43, no. 1 (2012): 3–24.
“DISARM Foundation.” Accessed April 21, 2022. https://www.disarm.foundation/.
EU vs DISINFORMATION. “DISINFO DATABASE.” Accessed April 21, 2022. https://euvsdisinfo.eu/disinformation-cases/.
EU vs DISINFORMATION. “Disinformation Review.” Accessed December 9, 2021. https://euvsdisinfo.eu/disinfo-review/.
Downey, Elizabeth A. “A Historical Survey of the International Regulation of Propaganda,” n.d., 21.
“EU Imposes Sanctions on State-Owned Outlets RT/Russia Today and Sputnik’s Broadcasting in the EU.” Accessed March 4, 2022. https://www.consilium.europa.eu/en/press/press-releases/2022/03/02/eu-imposes-sanctions-on-state-owned-outlets-rt-russia-today-and-sputnik-s-broadcasting-in-the-eu/.
“European Council Conclusions, 19-20 March 2015.” Accessed December 10, 2021. https://www.consilium.europa.eu/en/press/press-releases/2015/03/20/conclusions-european-council/.
“European Union and the United Nations - Wikipedia.” Accessed December 9, 2021. https://en.wikipedia.org/wiki/European_Union_and_the_United_Nations.

Treaty Office. “Full List.” Accessed December 9, 2021. https://www.coe.int/en/web/conventions/full-list.
Reuters. “Georgia Cuts Access to Russian Websites, TV News,” August 19, 2008, sec. Internet News. https://www.reuters.com/article/us-georgia-ossetia-media-idUSLJ36223120080819.
“Georgia Internet Users.” Accessed April 16, 2022. https://www.internetlivestats.com/internet-users/georgia/.
Giles, Keir. “Russian Information Warfare.” In *The World Information War*, edited by Timothy Clack and Robert Johnson, 1st ed., 139–61. Abingdon, Oxon ; New York, NY : Routledge, [2021] | Series: Routledge advances in defence studies: Routledge, 2021. https://doi.org/10.4324/9781003046905-12.
Golovchenko, Yevgeniy. “Measuring the Scope of Pro-Kremlin Disinformation on Twitter.” *Humanities and Social Sciences Communications* 7, no. 1 (December 11, 2020): 1–11. https://doi.org/10.1057/s41599-020-00659-9.
Granholm, Niklas, Johannes Malminen, and Gudrun Persson. “Ramifications of Russian Aggression Towards Ukraine,” n.d., 94.
“Graphika.” Accessed December 9, 2021. https://graphika.com/.
Han, SuHun. “Googletrans Documentation,” n.d., 29.
Helmus, Todd C. *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe*. Research Report (Rand Corporation), RR-2237-OSD. Santa Monica, Calif: RAND Corporation, 2018.
Higgins, Eliot. *We Are Bellingcat: An Intelligence Agency for the People*. Bloomsbury Press, 2021.
High Representative of the Union for Foreign Affairs and Security Policy. “Joint Communication to the European Parliment, the European Council, the Council, The European Economic and Social Committee, and the Commitee of Regions: Action Plan against Disinformation.” Brussels, Belgium: European Commission, May 12, 2018. https://eeas.europa.eu/sites/default/files/action_plan_against_disinformation.pdf.
BBC News. “How Kremlin Accounts Manipulate Twitter,” March 19, 2022, sec. Technology. https://www.bbc.com/news/technology-60790821.
Hutto, C., and Eric Gilbert. “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.” *Proceedings of the International AAAI Conference on Web and Social Media* 8, no. 1 (May 16, 2014): 216–25.
Iasiello, Emilio. “Russia’s Improved Information Operations: From Georgia to Crimea.” *The US Army War College Quarterly: Parameters* 47, no. 2 (June 1, 2017). https://press.armywarcollege.edu/parameters/vol47/iss2/7.
“International Convention Concerning the Use of Broadcasting in the Cause of Peace - Wikipedia.” Accessed December 9, 2021. https://en.wikipedia.org/wiki/International_Convention_Concerning_the_Use_of_Broadcasting_in_the_Cause_of_Peace.
International Convention concerning the Use of Broadcasting in the Cause of Peace Broadcasting, Pub. L. No. 4319, 186 303 (1936). https://treaties.un.org/doc/Publication/UNTS/LON/Volume%20186/v186.pdf.
International Covenant on Civil and Political Rights, 2200A § (1966). https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx.

Jack, Caroline. “Lexicon of Lies.” Data & Society. Data & Society Research Institute, August 9, 2017. https://datasociety.net/library/lexicon-of-lies/.
Jamnejad, Maziar, and Michael Wood. “The Principle of Non-Intervention.” *Leiden Journal of International Law* 22, no. 2 (June 2009): 345–81. https://doi.org/10.1017/S0922156509005858.
Lawrence, John, and Chris Reed. “Argument Mining: A Survey.” *Computational Linguistics* 45, no. 4 (January 1, 2020): 765–818. https://doi.org/10.1162/coli_a_00364.
Markoff, John. “Before the Gunfire, Cyberattacks.” *The New York Times*, August 12, 2008, sec. Technology. https://www.nytimes.com/2008/08/13/technology/13cyber.html.
Martino, Giovanni Da San, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. “Fine-Grained Analysis of Propaganda in News Articles.” *ArXiv:1910.02517 [Cs]*, October 6, 2019. http://arxiv.org/abs/1910.02517.
Microsoft Digital Security Unit. “An Overview of Russia’s Cyberattack Activity in Ukraine,” April 27, 2022.
Nakashima, Ellen. “Inside a Russian Disinformation Campaign in Ukraine in 2014.” *Washington Post*, December 25, 2017, sec. National Security. https://www.washingtonpost.com/world/national-security/inside-a-russian-disinformation-campaign-in-ukraine-in-2014/2017/12/25/f55b0408-e71d-11e7-ab50-621fe0588340_story.html.
EU vs DISINFORMATION. “News and Analysis.” Accessed May 9, 2021. https://euvsdisinfo.eu/news/.
Pamment, James. “The EU’s Role in Fighting Disinformation: Taking Back the Initiative.” Carnegie Endowment for International Peace. Accessed November 15, 2021. https://carnegieendowment.org/2020/07/15/eu-s-role-in-fighting-disinformation-taking-back-initiative-pub-82286.
Pomerantsev, Peter, and Michael Weiss. “How the Kremlin Weaponizes Information, Culture and Money,” n.d., 44.
“Project Grey Goose Phase II Report: The Evolving State of Cyber Warfare.” Greylogic, March 20, 2009.
Pruitt, Sarah. “How a Five-Day War With Georgia Allowed Russia to Reassert Its Military Might.” HISTORY. Accessed April 28, 2022. https://www.history.com/news/russia-georgia-war-military-nato.
United States Department of State. “Report: RT and Sputnik’s Role in Russia’s Disinformation and Propaganda Ecosystem.” Accessed March 4, 2022. https://www.state.gov/report-rt-and-sputniks-role-in-russias-disinformation-and-propaganda-ecosystem/.
Revaz, Topuria. “Russia’s Weapon of Words in Numbers. Evolution of Russian Assertive (Dis)Information Actions: Comparative Analysis of the Cases of Russo-Georgian War 2008 & Annexation of Crimea 2014 .” *Ante Portas*, 2020, 35.
Rid, Thomas. *Active Measures: The Secret History of Disinformation and Poltical Warfare*. First Edition. New York: Farrar, Straus and Grioux, 2020.
Rotaru, Vasile. “‘Mimicking’ the West? Russia’s Legitimization Discourse from Georgia War to the Annexation of Crimea.” *Communist and Post-Communist Studies* 52, no. 4 (October 19, 2019): 311–21. https://doi.org/10.1016/j.postcomstud.2019.10.001.
Twitter. “RT (@RT_com) / Twitter.” Accessed April 20, 2022. https://twitter.com/RT_com.
Human Rights Watch. “Russia: Halt Orders to Block Online Media,” March 23, 2014. https://www.hrw.org/news/2014/03/23/russia-halt-orders-block-online-media.

Reuters. “Russia Lifts Ban on Telegram Messaging App after Failing to Block It,” June 18, 2020, sec. Technology News. https://www.reuters.com/article/us-russia-telegram-ban-idUSKBN23P2FT.

Canadian Global Affairs Institute. “Russian Cyber-Operations in Ukraine and the Implications for NATO.” Accessed May 6, 2022. https://www.cgai.ca/russian_cyber_operations_in_ukraine_and_the_implications_for_nato.

BBC News. “Russian-Majority Areas Watch Moscow’s Post-Crimea Moves,” March 26, 2014, sec. Europe. https://www.bbc.com/news/world-europe-26713975.

“Russo-Georgian War.” In *Wikipedia*, April 14, 2022. https://en.wikipedia.org/w/index.php?title=Russo-Georgian_War&oldid=1082738325.

Safronova, Valeriya, Neil MacFarquhar, and Adam Satariano. “Where Russians Turn for Uncensored News on Ukraine.” *The New York Times*, April 16, 2022, sec. World. https://www.nytimes.com/2022/04/16/world/europe/russian-propaganda-telegram-ukraine.html.

“Satellite Imagery as Evidence for International Crimes | Coalition for the International Criminal Court.” Accessed May 4, 2022. https://www.coalitionfortheicc.org/news/20150423/satellite-imagery-evidence-international-crimes.

Schreck, Carl. “From ‘Not Us’ To ‘Why Hide It?’: How Russia Denied Its Crimea Invasion, Then Admitted It.” *Radio Free Europe/Radio Liberty*, 17:01:50Z, sec. Russia. https://www.rferl.org/a/from-not-us-to-why-hide-it-how-russia-denied-its-crimea-invasion-then-admitted-it/29791806.html.

Shaw, Mabel. “Guides: International and Foreign Cyberspace Law Research Guide: Tallinn Manual & Primary Law Applicable to Cyber Conflicts.” Accessed December 9, 2021. https://guides.ll.georgetown.edu/c.php?g=363530&p=4821482.

Washington Post. “Social Networks and Social Media in Ukrainian ‘Euromaidan’ Protests.” Accessed April 17, 2022. https://www.washingtonpost.com/news/monkey-cage/wp/2014/01/02/social-networks-and-social-media-in-ukrainian-euromaidan-protests-2/.

Statista. “Social Networks for News in Ukraine 2021.” Accessed May 4, 2022. https://www.statista.com/statistics/1029018/social-networks-for-news-in-ukraine/.

Sokol, Sam. “Russian Disinformation Distorted Reality in Ukraine. Americans Should Take Note.” *Foreign Policy* (blog). Accessed April 20, 2022. https://foreignpolicy.com/2019/08/02/russian-disinformation-distorted-reality-in-ukraine-americans-should-take-note-putin-mueller-elections-antisemitism/.

SOMA Disinfobservatory. “SOMA Officially in the European Commission’s Plan to Tackle Disinformation,” October 4, 2019. https://www.disinfobservatory.org/soma-officially-in-the-european-commissions-plan-to-tackle-disinformation/.

EU vs DISINFORMATION. “STUDIES AND REPORTS.” Accessed December 9, 2021. https://euvsdisinfo.eu/reading-list/.

Sullivan, Becky. “Russia’s at War with Ukraine. Here’s How We Got Here.” *NPR*, February 24, 2022, sec. World. https://www.npr.org/2022/02/12/1080205477/history-ukraine-russia.

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

"Tanbih Project." PRTA: A Tool for the Analysis of Propaganda Techniques in Texts. Accessed April 21, 2022. https://www.tanbih.org/prta.
Tay, Xuan W. "Reconstructing The Principle of Non-Intervention and Non-Interference – Electoral Disinformation, Nicaragua, and the Quilt-Work Approach," n.d., 47.
Tenove, Chris. "Protecting Democracy from Disinformation: Normative Threats and Policy Responses." *The International Journal of Press/Politics* 25, no. 3 (July 1, 2020): 517–37. https://doi.org/10.1177/1940161220918740.
Media Manipulation Casebook. "The Media Manipulation Case Book: The Code Book," October 15, 2020. https://mediamanipulation.org/code-book.
EU vs DISINFORMATION. "The Strategy and Tactics of the Pro-Kremlin Disinformation Campaign," June 27, 2018. https://euvsdisinfo.eu/the-strategy-and-tactics-of-the-pro-kremlin-disinformation-campaign/.
"The Unified Content Platform - Sanity.Io." Accessed December 9, 2021. https://www.sanity.io/.
Fortune. "TikTok's Algorithm Shows Users Fake News on Ukraine War." Accessed May 4, 2022. https://fortune.com/2022/03/21/tiktok-misinformation-ukraine/.
Reuters. "Timeline: Political Crisis in Ukraine and Russia's Occupation of Crimea," March 8, 2014, sec. Emerging Markets. https://www.reuters.com/article/us-ukraine-crisis-timeline-idUSBREA270PO20140308.
EU vs DISINFORMATION. "'To Challenge Russia's Ongoing Disinformation Campaigns': The Story of EUvsDisinfo," April 22, 2020. https://euvsdisinfo.eu/to-challenge-russias-ongoing-disinformation-campaigns-the-story-of-euvsdisinfo/.
Council on Foreign Relations. "Tracking Cyber Operations and Actors in the Russia-Ukraine War." Accessed May 4, 2022. https://www.cfr.org/blog/tracking-cyber-operations-and-actors-russia-ukraine-war.
Trained Models & Pipelines. "Trained Models & Pipelines · SpaCy Models Documentation." Accessed April 14, 2022. https://spacy.io/models.
Troianovski, Anton. "Russia Takes Censorship to New Extremes, Stifling War Coverage." *The New York Times*, March 4, 2022, sec. World. https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html.
"Ukraine - The Crisis in Crimea and Eastern Ukraine | Britannica." Accessed April 17, 2022. https://www.britannica.com/place/Ukraine/The-crisis-in-Crimea-and-eastern-Ukraine.
"Ukraine Internet Users." Accessed April 17, 2022. https://www.internetlivestats.com/internet-users/ukraine/.
Human Rights Watch. "Ukraine: Police Attacked Dozens of Journalists, Medics," January 30, 2014. https://www.hrw.org/news/2014/01/30/ukraine-police-attacked-dozens-journalists-medics.
United Nations. Universal Declaration of Human Rights (1948). https://www.un.org/sites/un2.un.org/files/udhr.pdf.
Volkova, Svitlana, and Jin Yea Jang. "Misleading or Falsification: Inferring Deceptive Strategies and Types in Online News and Social Media." In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 575–83. Lyon, France: ACM Press, 2018. https://doi.org/10.1145/3184558.3188728.
Nieman Lab. "Why Telegram — despite Being Rife with Russian Disinformation — Became the Go-to App for Ukrainians." Accessed May 4, 2022.

https://www.niemanlab.org/2022/03/why-telegram-despite-being-rife-with-russian-disinformation-became-the-go-to-app-for-ukrainians/.
Wiggins, Bradley E. “Crimea River: Directionality in Memes from the Russia–Ukraine Conflict,” 2016, 35.
Wong, Edward, and Julian E. Barnes. “China Asked Russia to Delay Ukraine War Until After Olympics, U.S. Officials Say.” *The New York Times*, March 2, 2022, sec. U.S. https://www.nytimes.com/2022/03/02/us/politics/russia-ukraine-china.html.
Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.” *ArXiv:1609.08144 [Cs]*, October 8, 2016. http://arxiv.org/abs/1609.08144.
Yuan, Li. “How China Embraces Russian Propaganda and Its Version of the War.” *The New York Times*, March 4, 2022, sec. Business. https://www.nytimes.com/2022/03/04/business/china-russia-ukraine-disinformation.html.
Zhang, Lei, Shuai Wang, and Bing Liu. “Deep Learning for Sentiment Analysis: A Survey.” *WIREs Data Mining and Knowledge Discovery* 8, no. 4 (2018): e1253. https://doi.org/10.1002/widm.1253.