Examining the Internet Research Agency's Exploitation of Cognitive Biases Through its Disinformation Campaign Targeted at the US 2016 Presidential Election

July 2020 Submission

University of Glasgow Student ID: 2393530C

Dublin City University Student ID: 18114334

Charles University Student ID: 23238301

Presented in partial fulfilment of the requirements for the Degree of:
International Master in Security, Intelligence and Strategic Studies (IMSISS)

Word Count: 21,831
Supervisor: Dr. Jakub Tesař
Date of Submission: July 27, 2020

Table of Contents

Introduction

In October 2019, the US Senate Select Committee on Intelligence released the second volume of its report: 'On Russian Active Measures Campaigns and Interference in the 2016 US Election'.[1] Where the initial report focused broadly on the Intelligence Community's findings confirming Russian election interference, the second found that the Russian government and its subsidiary, the Internet Research Agency (IRA), engaged in a multi-faceted campaign aimed to 'undermine public faith in the US democratic process, denigrate [candidate] Secretary Clinton, and harm her electability and potential presidency.'[2] The committee found that the Russian effort was not solely intended to influence the results of the US election.[3] The campaign was, in fact, part of a greater effort on the part of the Kremlin to sow discord in American society and undermine faith in its governmental institutions. In carrying out what the New York Times labelled the 'Pearl Harbor of the social media age', the Russian election-meddling effort employed an online campaign designed to gain the attention and influence the opinions of the American public.[4] In the year before the election, the IRA had roughly 400 employees working shifts twenty-four hours a day and posting content across virtually every social media platform.[5] Reports conflict on the efficacy of the online disinformation effort, but an Oxford Internet Study found that the 17 million tweets posted during the election period led Russian automated bot accounts to attain online network positions of 'measurable influence,' which placed them in a strong posture to seek to manipulate the views of US online users.[6] Another study found that every 25,000 retweets of IRA tweets predicted a 1 percent increase in candidate Donald Trump's polling numbers. [7]

---

[1] UNITED STATES COMMISSION ON SECURITY AND COOPERATION IN EUROPE, "THE SCOURGE OF RUSSIAN DISINFORMATION," 2017, https://www.technologyreview.com/s/604084/russian-disinformation-technology/; US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION" (U.S. Senate, 2019), https://www.warner.senate.gov/public/_cache/files/0/d/0dc0e6fe-4d52-49b0-9e92-a15224a74a29/C2ABC2CD38BA3C5207D7FA5352D53EC2.report-volume2.pdf.

[2] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."

[3] Christopher A. Bail et al., "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017," *Proceedings of the National Academy of Sciences*, November 25, 2019, https://doi.org/10.1073/pnas.1906420116.

[4] Bail et al.

[5] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."

[6] UNITED STATES COMMISSION ON SECURITY AND COOPERATION IN EUROPE, "THE SCOURGE OF RUSSIAN DISINFORMATION."

[7] Damian J Ruck et al., "Internet Research Agency Twitter Activity Predicted 2016 U.S. Election Polls," *First Monday* 24, no. 7 (2019), https://doi.org/10.5210/fm.v24i7.10107.

The mass popularity of social media, particularly as a news source, has afforded the Kremlin and its supporting agencies unprecedented direct access into the thoughts and conversations of American citizens. What previously required elaborate physical efforts to foster rifts in US society, such as drafting forged letters to the African American community purporting to be from the Ku Klux Klan to engage in information warfare, the Kremlin can now weaponise information soldiers to inject themselves directly into the socio-political conversations of millions of Americans.[8] Further enhancing the potency of its online initiative, the IRA was able to take advantage of hashtags and algorithmic effects that allowed IRA employees to identify and direct disinformation towards groups utilising a method known as micro-targeting.[9] Kremlin automated 'bots', accounts masquerading as Americans known as 'sock puppets', and agitators known as 'trolls' would inject themselves into online discourse and attempt to influence opinions of the American voting population. The Kremlin's efforts were not uniquely targeted at the US. The French 2017 presidential election, the 2016 Brexit vote, among several other Western democratic processes were targeted by similar agile disinformation efforts directed by the Kremlin.[10]

Information warfare is not a new element of Russian strategy. The concept of reflexive control, which involves influencing the adversary's decision making by altering factors intended to manipulate his or her perceptions, has been characterised in academic research dating as far back as the 1960.[11] Russia initially acknowledged its comparative strategic inferiority in 1999 when Russia's Minister of Defence admitted that Moscow could not compete militarily with NATO and the West.[12] Consequently, Russia has sought to undermine the West asymmetrically by nipping and pulling at the threads that underlie Western democratic cohesion. Namely, electoral processes, media integrity, pluralistic social discourse, and confidence in the credibility of government institutions. When confronted

[8] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."
[9] Philip N Howard, John Kelly, and Camille François, "The IRA, Social Media and Political Polarization in the United States , 2012-2018," *Computational Propaganda Research Project*, 2018.
[10] W. Lance Bennett and Steven Livingston, "The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions," *European Journal of Communication* 33, no. 2 (2018): 122–39, https://doi.org/10.1177/0267323118760317.
[11] Keir Giles, "Handbook of Russian Information Warfare," *NATO Defence College* 9, no. November (2016): 1–90.
[12] Andrew Stuttaford, "Information War and Other Deceptions," *National Review*, April 12, 2015, https://www.nationalreview.com/corner/information-war-and-other-deceptions/.

about its behaviour, the IRA-controlled accounts justified, dismissed, normalised, and redirected blame.[13]

Additional key factors contributing to the potency of the Kremlin's information warfare operation are the three vulnerabilities that render the West a promising target for Kremlin-style online propaganda, which are best characterised as: media, algorithmic, and cognitive vulnerability.[14] Western news media maintains a pluralistic ethic and relies heavily on outlets' reputations founded on a legacy of publishing fair and accurate news. This presents an opportunity for Kremlin media propagators because they can publish quickly and distribute widely stories that fit the narrative that Russia wants its adversary to believe with comparably minimal concern for the accuracy of the content. A Pew Research survey found that in 2016, the year of the election, 62 percent of Americans consumed their news on social media.[15] This number has been steadily climbing for years, as 2016 marked the first year that social media overtook print news. This trend poses a key challenge for shielding against the affliction of online disinformation.

The contents of social media websites are algorithmically defined. This means that the platform's concern over the accuracy of a user's newsfeed comes secondary to its motive to keep the user engaged. The algorithm, therefore, will continue to serve users with content it assesses that the user would like to see. Algorithms can be exploited by motivated actors with an understanding of how the platform prioritises content. Indeed, the IRA would employ bots to amplify the perceived importance of a particular topic and micro-target by relying on hashtags to serve that information to its desired userbase.[16] Algorithmic gaming is especially useful because of human online cognitive vulnerability to false information. Schmidta et al (2016) found that online users are 'cognitively lazy' and tend to limit their consumption to a finite grouping of pages, while only pursuing information that confirms their already held views, which has the side effect of feeding into the influence of confirmation bias.[17] The IRA

[13] Rory Cormac and Richard J. Aldrich, "Grey Is the New Black: Covert Action and Implausible Deniability," *International Affairs*, 2018, https://doi.org/10.1093/ia/iiy067.
[14] Theo Brinkel, *Netherlands Annual Review of Military Studies 2017: Winning Without Killing:The Strategic and Operational Utility of Non-Kinetic Capabilities in Crises*, 2017, http://ebookcentral.proquest.com/lib/mindef/detail.action?docID=4915496; Howard, Kelly, and François, "The IRA, Social Media and Political Polarization in the United States , 2012-2018"; Rand Waltzman, "The Weaponization of Information: The Need for Cognitive Security," *The Weaponization of Information: The Need for Cognitive Security* (RAND Corporation, 2017), https://doi.org/10.7249/ct473.
[15] https://www.reuters.com/article/us-usa-internet-socialmedia/two-thirds-of-american-adults-get-news-from-social-media-survey-idUSKCN1BJ2A8
[16] Robert S Mueller, "Report On The Investigation Into Russian Interference In The 2016 Presidential Election," vol. I and II, 2019.
[17] Christina Nemr and William Gangware, "'Weapons of Mass Distraction'?," *Park Advisors*, 2019, https://doi.org/10.2307/j.ctv2n7qxg.13.

took advantage of this by winning over online groups through sock puppet accounts that posed as legitimate American accounts and tailored posts and comments to the views of its target audience. Once embedded in these online communities, these accounts could use their influence to spread disinformation as well as non-falsifiable but provocative memes and photos meant to influence users' perceptions, such as an image of both candidates appearing outraged designed to persuade potentially disaffected voters to boycott the election. See Table 1 for reference[18] It is unlikely that Russian disinformation will end in the near term as the Senate report found that IRA social media activity actually increased after the election and as recently as 2020, Russia was cited for engaging in disinformation relating to the COVID-19 pandemic.[19]

Table 1

Source: Sproul (2019)[20]

To secure electoral systems and the health of the pluralistic discourse laying at the foundation of liberal societies, the US and the rest of the West will have to develop policies

[18] Spencer Sproul, "Fake News ! Russian Disinformation Targets American Cognitive Biases Through Diverse Mediums Fake News ! Russian Disinformation Targets American Cognitive Biases Through Diverse Mediums," 2019.
[19] Sergey Sukhankin, "Covid-19 As a Tool of Information Confrontation: Russia's Approach," *The School of Public Policy Publications* 13, no. 3 (2020): 1–10.
[20] Sproul, "Fake News ! Russian Disinformation Targets American Cognitive Biases Through Diverse Mediums Fake News ! Russian Disinformation Targets American Cognitive Biases Through Diverse Mediums."

and behaviours to shield themselves against the impacts of online disinformation. A robust response will require a clear understanding of how the disinformation campaign worked, what drove its potency, and enacting policy designed to secure against these vulnerabilities. The primary question this paper must answer is: *was the effort coherent, meaning did it follow a clear directive that may allow us to characterise the scope and nature of the effort?* Policy will need to focus on the vulnerability of social media platforms and the means by which state actors exploit them to spread disinformation. The other essential question, informed by reflexive control doctrine, is: *how did the Kremlin employ cognitive exploitation to infiltrate American online communities?* The Kremlin's specific understanding of American users' digital cognition that drove the effort is yet unknown, but one theory is that it specifically targeted cognitive biases and heuristics. Individuals consume online media differently from the physical form and are able to interact with the news they consume through comments, likes, and shares unlike ever before and there is a sizable body of literature outlining how cognitive biases influence online media consumption.

Cognitive bias is defined as when: 'human cognition reliably produces representations that are systematically distorted compared to some aspect of objective reality', which fits neatly with the Soviet Union's definition of reflexive control that appears to continue to inform Russian strategic doctrine today.[21] Heuristics are mental shortcuts used to assess a scenario based on incomplete information in complex situations. Heuristics can also be useful tools that allow individuals to render suboptimal but often adequate decisions when times and resources are constrained.[22] If research could identify that the IRA may have exploited users' cognitive biases through its disinformation campaign, the resulting information could prove revelatory to policy intended to secure individuals and platforms from future attempts to pursue similar efforts. *Therefore, this paper will assess if the Russian disinformation effort during the 2016 US presidential election targeted cognitive biases and heuristics, and if so, which may have been targeted?* For the purposes of an adversary, biases and heuristics present an appealing target because in a large sample they are both predictable and entail incomplete conscious thought and thus render individuals vulnerable to adversaries' efforts to

[21] Martie G. Haselton, Daniel Nettle, and Paul W. Andrews, "The Evolution of Cognitive Bias," *The Handbook of Evolutionary Psychology*, 2015, 724–46, https://doi.org/10.1002/9780470939376.ch25.; Reflexive Control: the practice of predetermining an adversary's decision in Russia's favour, by altering key factors in the adversary's perception of the world. Giles, "Handbook of Russian information warfare."
[22] Gerd Gigerenzer and Wolfgang Gaissmaier, "Decision Making: Nonrational Theories," *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* 5 (2015): 3304–3309, https://doi.org/10.1016/B978-0-08-097086-8.26017-0.

manipulate inputs to intervene and misdirect decision making without the target even recognising. As a hypothetical, if an adversary were aware that in a complex decision-making environment, an individual or group would be more likely to accept a suggestion if that suggestion provoked vivid imagery, then the adversary could stack the deck by vigorously sending such inputs to the recipient group.[23] The adversary's choice of biases would therefore rely on what content the platform rewards by boosting the posts' recognition and which biases are most likely to influence cognition on the platform.

To evaluate if biases were targeted, the enquiry will first need to identify which cognitive biases might have been exploited. There are no direct references on the Kremlin's part to targeting specific cognitive biases and a search for direct references to biases within Twitter's IRA dataset returned negative results.[24] Therefore, the paper selects three biases for analysis based on known IRA tactics and the most likely biases the group would have targeted. What is known about the nature of the Kremlin's 2016 disinformation effort is that it employed micro-targeting and sought to embed itself in online communities, it produced a high volume of posts, and exploited tense fissures in the US social fabric.[25] This suggests that the biases that the Kremlin may have targeted are confirmation bias: 'a tendency to search for, interpret, focus on and remember information in a way that confirms one's preconceptions'; the availability heuristic: 'a tendency to judge the frequency or likelihood of an event by the ease with which relevant information comes to the mind;' and vividness bias: 'the tendency of decision makers to gravitate towards salient and visually stimulating alternatives because they attract more attention and are easier to recall'.[26]

Lacking evidence of clear directives to exploit cognitive bias on the part of the Kremlin, this paper will employ conceptualisation and operationalisation to rely on the theory and definitions surrounding each of the three biases and then assess how the IRA might have exploited these cognitive errors. The conceptualisation will posit: if Russia were to exploit each bias, how would that be conducted and what might be the effort's dominant features?

---

[23] Rebecca M. Todd et al., "Psychophysical and Neural Evidence for Emotion-Enhanced Perceptual Vividness," *Journal of Neuroscience* 32, no. 33 (2012): 11201–12, https://doi.org/10.1523/JNEUROSCI.0155-12.2012.
[24] Vijaya Gadde and Yoel Roth, "Enabling Further Research of Information Operations on Twitter," Twitter.com, 2018, https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html.
[25] Christopher Paul and Miriam Matthews, "The Russian 'Firehose of Falsehood' Propaganda Model: Why It Might Work and Options to Counter It," *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*, 2017, https://doi.org/10.7249/pe198.
[26] Han Bouwmeester, "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory'" (T.M.C. Asser Press, The Hague, 2017), 125–53, https://doi.org/10.1007/978-94-6265-189-0_8; Fred D. Davis et al., "Information Systems and Neuroscience," *Gmunden Retreat on NeuroIS 2016*, 2017, https://doi.org/10.1007/978-3-319-41402-7.

These features will then be analysed quantitatively to identify the degree to which the effort manifested these characteristics. The dominant trait of confirmation bias is a tendency to seek information in alignment with one's previously held views. Therefore, the enquiry will assess to what degree the Kremlin's effort exploited the digital infrastructure to ensure that users were provided with narratives that support their preconceived thought patterns. Vividness bias is defined by a preference granted to thoughts and ideas that are most easy to recall due to their visceral characteristics. This paper will thus rely on a sentiment analysis to identify if IRA tweets invoked higher degrees of emotion than the broader American Twitter population. The availability heuristic occurs when an individual's perception is influenced by thoughts that easily come to mind due to their comparably greater presence than other ideas. Following this definition, the paper will analyse if the IRA produced tweets in higher volumes than the American user population to identify if the IRA accounts sought to game users' perceptions through the availability heuristic. The paper will evaluate the case study of the 2014 Russian annexation of Crimea as a more local case to Russia and assess what parallels and evidence it offers to aid in evaluating the US 2016 election operation. It will then evaluate the likelihood that confirmation bias was exploited based on the literature relating to IRA network infiltration and exploitation of retweet networks, which are communities of like-minded users that engage and share ideologically homogenous content. Next, it will analyse the literature on vividness bias and emotion and assess a randomised sample of IRA tweets to evaluate their emotional content through a sentiment analysis using Bessi & Ferrara (2016) as a control. It will follow by analysing the literature on the availability heuristic and assess the volume of IRA tweets relative to human users also relying on Bessi & Ferrara's analysis as a control.[27]

The results section will evaluate the data versus the control and assess the likelihood that each bias was exploited based on the results. If the paper confirms the author's hypothesis that the IRA did in fact target the three examined biases, the findings would serve as strong evidence that Russia has been leveraging the digital medium to exploit cognitive processes and undermine the social cohesion of its adversaries. These findings would call for further assessment of solutions specifically tailored to the cognitive element of the disinformation landscape. The paper will rely on such findings to identify specific policy elements that might serve to secure democracies against this threat going forward. It will

[27] Alessandro Bessi and Emilio Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election," *First Monday* 21, no. 11 (2016): 1–15.

refer to Finland and Italy's models for disinformation resilience building, assess opportunities for public-private cooperation, and identify previous efforts to build cognitive resilience that might prove informative to future efforts to secure the public against the evolution of this threat.[28] The evidence suggests that disinformation presents a mounting challenge to the integrity of Western democratic processes and due to its efficacy and low cost, will continue to impair the democracies of the West until they can develop robust wide-reaching policy positioned to combat disinformation's corrosive impacts.

Literature Review

Impact

The literature provides substantial evidence that modern Russian information warfare targets decision-making processes that closely align with the definition of cognitive biases. To this point, however, few papers have focused on Russia's exploitation of cognitive biases by name.[29] Accordingly, no research focused specifically on Kremlin efforts to exploit the cognitive biases of US users during the 2016 US presidential election. To narrow this gap, it is essential to interrogate the scope and impact of the Kremlin's effort during the period leading up to the election. The Kremlin's effort sent waves throughout the West. It was multipronged, involved hacking materials of prominent members of the Clinton campaign, and featured a large-scale social media effort designed to sow discord and undermine public confidence in US governmental institutions.[30]

The disinformation campaign was also robust. One study found that the IRA produced over 57,000 posts on Twitter, 2,400 on Facebook, and 2,600 on Instagram.[31] Furthermore, the US Senate found that IRA accounts achieved 'positions of measurable influence' and were able to leverage that placement to influence large-scale online conversations concerning the election.[32] Though it is difficult to assess the effort's actual impact on voting behaviour, another study found that every 25,000 retweets of IRA content served as a reliable predictor

---

[28] Daniel Fried and Alina Polyakova, "DEMOCRATIC DEFENSE AGAINST DISINFORMATION" (Washington, D.C., 2018), https://www.atlanticcouncil.org/wp-content/uploads/2018/03/Democratic_Defense_Against_Disinformation_FINAL.pdf.
[29] Georgii Pocheptsov, "COGNITIVE ATTACKS IN RUSSIAN HYBRID WARFARE," *Information & Security: An International Journal* 41 (2018): 37–43, https://doi.org/10.11610/isij.4103.
[30] Brinkel, *Netherlands Annual Review of Military Studies 2017: Winning Without Killing:The Strategic and Operational Utility of Non-Kinetic Capabilities in Crises*.
[31] Bail et al., "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017."
[32] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."

for a 1 percent increase in candidate Trump's polling numbers.[33] Other studies disputed the IRA's impact on political opinion. One of which employed six distinctive measures to try to identify IRA accounts' impact on political views and found no evidence that they measurably influenced the American users' political opinions. They found that users most likely to interact with IRA accounts were involved in networks with strong ideological congruity with their own views, were heavily interested in politics, were active Twitter users, and already leaned towards their respective ideological extremes.[34]

Zannettou (2019) found through a random sampling of IRA accounts that the profiles tended to adopt multiple online identities while building followers in order to increase their network influence.[35] IRA employees employed 'sock puppet' accounts intended to appear as legitimate users to deceive their online targets, paid trolls, automated bot accounts, and networks of bots known as botnets to swarm topics. The IRA deployed botnets with the goal of increasing the perceived importance of a particular conversation, a tactic referred to as 'computational propaganda'.[36] Through these accounts and micro-targeted advertisements, IRA members exploited the filter bubbles—ideologically homogenous online communities—and directed messages toward the specific online communities they sought to influence.[37] Rather than dictate a discussion, IRA bots sought to amplify the relevance of a topic through repeated posting. Till's (2020) report explained that over the course of a twelve-hour shift, IRA employees were required to comment a minimum of fifty times on new articles, oversee six Facebook accounts, engage in two conversations in news-focused groups, and post fifty total tweets from ten accounts.[38] At its core, the objective of the operation was to sow division amongst the US population and to direct voters against voting for candidate Clinton.[39] This evidence supports the notion that the IRA's effort had distinct qualities that

---

[33] By Damian Ruck, "Russian Twitter Propaganda Predicted 2016 US Election Polls," *The Conversation*, 2019, 1–8.
[34] Bail et al., "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017."
[35] Savvas Zannettou et al., "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web," *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 2019, 218–26, https://doi.org/10.1145/3308560.3316495.
[36] Howard, Kelly, and François, "The IRA, Social Media and Political Polarization in the United States , 2012-2018."
[37] Leo G Stewart, Ahmer Arif, and Kate Starbird, "Examining Trolls and Polarization with a Retweet Network," *Proceedings of WSDM Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*, 2018, 6, https://doi.org/https://doi.org/10.475/123_4.
[38] Christopher Till, "Propaganda through 'Reflexive Control' and the Mediated Construction of Reality," *New Media and Society*, 2020, 1–17, https://doi.org/10.1177/1461444820902446.
[39] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."

may have been driven by its overall strategy. This author hypothesizes that the deliberate deluge of information was intended to overwhelm US users with volume and compel them into forming their understanding based on the excessive content volume rather than the accuracy of the information enclosed within the posts.

Russian Strategy

To better understand the effort, it is important to contextualise the strategy within the literature assessing Russian strategic doctrine. Russia views NATO as an existential threat to its security and is aware that it cannot compete with NATO's military apparatus through conventional means.[40] Therefore, Moscow's view is that to maintain its security, it must challenge NATO members through an asymmetric approach. The precise origins of the Russian hybrid warfare doctrine are disputed. But in 2013, Chief of Staff of the Russian armed forces, Valeri Gerasimov unveiled in an article titled *'The Value of Science in Prediction'* the modern Russian strategic doctrine vis a vis the West and highlighted its emphasis on non-linear warfare with the ultimate goal of degrading the enemy's societal morale.[41] hybrid warfare includes military tactics but is distinguished by its emphasis on broadening the scope of the battlefield to the psychological domain through propaganda, media manipulation, and cyber-attacks.[42] Offering further evidence of the Russian move to a hybrid approach, in 2017, Russian Defence Minister Sergey Shoigu publicly acknowledged the 'information Army' within the Russian armed forces and proclaimed that due to modern developments, its efforts would serve as a far more effective tool than Russian historical propaganda efforts. In the same year, Gerasimov acknowledged that Russian non-military warfare was being employed at quadruple the rate of military efforts.[43]

The propaganda and disinformation that comprise the Kremlin's modern hybrid warfare strategy appear to be deeply embedded in Russian strategic doctrine and informed by the legacy of 'reflexive control'. Reflexive control meaning: 'predetermining an adversary's decision in Russia's favour, by altering key factors in the adversary's perception of the

[40] Harri Mikkola, "The Geostrategic Arctic: Hard Security in the High North," *Finnish Institute OfInternational Affairs*, no. April (2019), https://www.fiia.fi/wp-content/uploads/2019/04/bp259_geostrategic_arctic.pdf; Giles, "Handbook of Russian Information Warfare."
[41] Brinkel, *Netherlands Annual Review of Military Studies 2017: Winning Without Killing:The Strategic and Operational Utility of Non-Kinetic Capabilities in Crises*.
[42] Brinkel.
[43] Waltzman, "Weaponization Inf. Need Cogn. Secur."

world' can be traced back to the Soviet Union in the 1960s.[44] Giles explains that reflexive control is an essential component of asymmetric warfare as it defines its intention to influence behaviour through coercive action, closely resembling the intention behind Gerasimov's stated non-linear warfare that was evident in the IRA's effort to influence the US 2016 presidential election.[45] Retired Russian Major General Ionov, an information warfare specialist, highlights that reflexive control delivers a traditional Soviet-styled informational and psychological attack on the adversary. Additionally, he argues that by understanding the individual or group's personality and psychological makeup, reflexive control can be tailored to target the enemy's cognitive vulnerabilities.[46] Through this strategy, Kremlin information operatives endeavour to influence the adversary's decision-making such that it unwittingly arrives at decisions in the Kremlin's best interest. Retired US Lieutenant Colonel and Soviet military expert Timothy Thomas maintains that reflexive control theory, though rooted in historical Soviet strategic doctrine, is still undergoing refinement and remains influential to Russian strategy today.[47] Sufficient research has not yet analysed the role of reflexive control doctrine as applied to its modern disinformation efforts.

Global Scale

Russia's modern information warfare has been applied in several cases beyond the US. Indeed, the Kremlin and its subsidiary organisations engaged in coordinated operations intended to interfere with institutional processes and societal integrity during the 2014 annexation of Crimea in Ukraine, the 2016 Brexit vote, the 2017 French election, as well as the World Anti-Doping Agency after it exposed the Russian state-sponsored performance enhancing drug program.[48] The precise implementation of these operations varied in each case, but common features included, the employment of unofficial government-friendly institutions- such as APT28- or Fancy Bear, elaborate campaigns across social media platforms, and an overall objective to degrade the credibility and support of democratic institutional bodies.[49] Mounting evidence is pointing to the notion that the Kremlin is

[44] ML Jaitner, "Applying Principles of Reflexive Control in Information and Cyber Operations," Journal of Information Warfare, 2016, https://doi.org/10.1002/9781118381533.
[45] Giles, "Handbook of Russian Information Warfare."
[46] Timothy L. Thomas, "Russia's Reflexive Control Theory and the Military," *International Journal of Phytoremediation* 17, no. 2 (2004): 237–56, https://doi.org/10.1080/13518040490450529.
[47] Bouwmeester, "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory.'"
[48] Casey Cannon, "Russia's Employment of Covert Action against Western Democracies from 2013 to Present" (Glasgow, UK, 2018).
[49] Cannon.

continuing to engage in disinformation efforts and likely will persist so long as it believes that it can successfully influence and undermine its adversaries from within.

Vulnerabilities – Civic and Media

Several studies highlighted the factors that serve to explain why the Kremlin remains deeply engaged in directing its information warfare strategy towards the West. Online disinformation is vastly cheaper than conventional warfare, presents a low risk of casualties, and can be difficult to attribute.[50] The Kremlin also pursues the strategy due to a host of vulnerabilities unique to Western pluralism and the digitally integrated societies comprising the West. In his article, 'the Resilient Mind-Set and Deterrence', Theo Brinkel identifies increasing levels of civic disengagement and social capital as drivers leading the West to a sense of diminished shared commitment to collective prosperity and liberty and a loss of faith in the leaders and institutions required to uphold those values.[51] He argues that this trend is fuelled by waning influence of national political institutions due to the influence of supranational organisations such as the EU and IMF, the increased political weight of the private sector, and a greater dependence on global markets that inhibit citizens' access to domestic capital. This, in Brinkel's view, has contributed to populist and more radical movements that share distrust for traditional political dynamics and the media. Incidentally, these high levels of institutional mistrust and the advent of social media as a bastion for non-traditional news sources have provided a fertile environment for the Kremlin's information warfare campaign.[52]

Considering these conditions, this paper will outline three primary vulnerabilities that this author hypothesises were strategically exploited by the Kremlin to maximise the potency of its disinformation effort. These vulnerabilities consist of the West's pluralistic media environment, the vulnerabilities of the social media networks to coordinated propaganda campaigns, and the cognitive vulnerability of Western social media users. Pomerantsev and Weiss (2014) note that freedom of information and the press represent pillars of Western democracy. A negative consequence of these values, however, is that the West's commitment to plurality presents an opportunity for malicious actors intent on exploiting those freedoms to subvert debate and exploit discourse. Vasily Gatov, a Russian media researcher with the

[50] Jennifer Kavanagh et al., "FIGHTING DISINFORMATION Building the Database of Web Tools," *RAND*, n.d., https://www.rand.org/content/dam/rand/pubs/research_reports/RR3000/RR3000/RAND_RR3000.pdf.
[51] Brinkel, *Netherlands Annual Review of Military Studies 2017: Winning Without Killing:The Strategic and Operational Utility of Non-Kinetic Capabilities in Crises*.
[52] Brinkel.

University of Southern California, asserted that the 21st century's greatest challenge will be against malicious actors intent on abusing democracies' freedom of information.[53]

A key enabler of the Kremlin's efforts to take advantage of free press is the public's mass migration to consuming its news online, and specifically on social media. As the introduction of this paper points out, in 2016 more individuals consumed news on social media than print for the first time. Communication technologies have become so pervasive that Till (2020) believes they have fundamentally shifted the social interactions shaping the public's conception of meaning.[54] One commonality enabling disinformation amongst social media and the Western pluralistic ethic is access. Openly Kremlin-friendly news outlets such as Sputnik and Russia Today (RT) spread propaganda and permeate US airwaves uninhibited relative to foreign outlets in Russia.[55] According to Bouwmeester, Western media outlets' commitment to balanced news at times saw them unwittingly sharing Russian narratives with their audiences.[56] The emphasis on providing a voice to both sides may have the consequence of lending credibility to an 'other side' of a debate that was only created as a result of a Russian propaganda effort. In one prominent example, an IRA sock puppet Twitter account known by the name 'Jenna Abrams' was found to have appeared in articles by over thirty outlets, including mainstream publications like the New York Times and CNN.[57]

Social media also allows for a broadening of the social conception of what constitutes journalists and news outlets. This means that independent journalists and pundits have unprecedented ability to proliferate headlines and false narratives absent editorial control.[58] Waltzman describes this new landscape as a democratisation of media influence, as any individual or group can position themselves through social media platforms as a news publisher. This presents a particular challenge because reputable news outlets must undergo fact-checking and editorial processes before publishing stories to maintain their credibility.[59] Russian propaganda faces dissimilar constraints and can distort or spread falsehoods about an event well before traditional outlets can publish a reviewed article or take. This

---

[53] Peter Pomerantsev and Michael Weiss, "The Menace of Unreality: How the Kremlin Weaponizes Information, Culture and Money A Special Report Presented by The Interpreter, a Project of the Institute of Modern Russia," 2014.
[54] Till, "Propaganda through 'Reflexive Control' and the Mediated Construction of Reality."
[55] Fried and Polyakova, "DEMOCRATIC DEFENSE AGAINST DISINFORMATION."
[56] Bouwmeester, "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory.'"
[57] Yiping Xia et al., "Disinformation, Performed: Self-Presentation of a Russian IRA Account on Twitter," *Information Communication and Society* 22, no. 11 (2019): 1646–64, https://doi.org/10.1080/1369118X.2019.1621921.
[58] Paul and Matthews, "Russ. 'Firehose Falsehood' Propag. Model Why It Might Work Options to Count. It."
[59] Waltzman, "Weaponization Inf. Need Cogn. Secur."

characteristically nimble Kremlin disinformation style offers its preferred narrative advantages over well-intentioned mediums in delivering users their first impression of a particular event. This has added cognitive influence due to the anchoring bias, which is the phenomenon when an individual's first impression has a persistent effect in colouring their future perception of a particular idea or event.[60] Following the fallout from the 2016 election, social media platforms took efforts to restrict and remove IRA accounts. But up until the election, the IRA was able to conduct a relatively unbridled disinformation campaign across Facebook, Twitter, Instagram, and YouTube.[61]

Vulnerabilities – Social Media Platforms

Waltzman confirms the historical basis for state-led disinformation initiatives but explains that the 21st century has provided unique opportunity for these efforts to proliferate across borders. Social media lowers the cost and offers disinformation virtually instant access to targeted audiences across the world.[62] This, coupled with individuals' willingness to increasingly rely on social media as their primary source of news, presents unprecedented opportunity for states to proliferate disinformation. Social media networks offer both access to millions of individuals and the ability to spread messages widely before they can be fact-checked.[63] In addition to platforms' provided access to vast numbers of people across borders, social media offers the ability to micro-target specific user groups. Micro-targeted messaging can be finely tailored such that the intended users are receiving refined messages based on their perceived socio-political views. These conditions are precisely what makes networks so enticing to advertising companies and equally so for disinformation campaigns.

A fundamental aspect of social media platforms' utility for both advertisers and disinformation efforts are the algorithms determining site content. Indeed, the networks' algorithms are designed to boost engagement by serving users the content they believe they would want to see.[64] Savvy 'cyber troops' understand how to manipulate these algorithms through illegal data harvesting and 'micro-profiling' to exploit and amplify the narratives

[60] Bouwmeester, "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory.'"
[61] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."
[62] Waltzman, "Weaponization Inf. Need Cogn. Secur."
[63] Massimo Flore et al., "Understanding Citizens' Vulnerabilities to Disinformation and Data-Driven Propaganda," 2019, https://doi.org/10.2760/919835 10.2760/153543.
[64] Flore et al.

they intend to deliver, thereby increasing their perceived importance.[65] Furthermore, users have an ability through social media platforms to engage directly with content and other users through likes, comments, and shares, and to curate their own media experience by self-selecting accounts and news producers that appeal to their beliefs.[66] This ability, paired with the platform's inclination to provide users with tailored content, often leads to the development of online communities centred around specific nodes or hashtags. These 'echo chambers,' which are most prominent on Twitter, allow users to engage almost exclusively with content and views that endorse their preconceived belief structures. IRA accounts that embed themselves within these retweet networks can steer the conversations by proliferating and amplifying false narratives intended to push users to the extremes within their ideological circles.[67] The intention behind these efforts is to increase the divide across political views and sow discord amongst US users by promoting polarised content and secessionist movements. These divides can be so strong that one study found that around one viral hashtag, only .65 percent of tweets were shared across supporters of either candidate.[68]

Vulnerabilities – Cognitive

The information consumption environment that has developed on social media platforms has allowed researchers to study the unique characteristics surrounding users' social media behaviours. The key difference between consuming news and social discourse on platforms like Facebook, Twitter, and Instagram is that the networks host a deluge of news content unmanageable for any individual user to mentally process. As of 2018, 6,000 tweets were posted every second and 300 million photos were posted to Facebook each day.[69] Mass communication research has argued that the inability to process the vast swaths of news and commentary online leads to public confusion over socially prominent discourse.[70] Researchers on this topic argue that individuals are incapable of mentally processing these

[65] Howard, Kelly, and François, "The IRA, Social Media and Political Polarization in the United States , 2012-2018."
[66] Walter Quattrociocchi, Antonio Scala, and Cass R. Sunstein, "Echo Chambers on Facebook," no. 2016 (n.d.), https://doi.org/10.1145/2740908.2745939.
[67] Mueller, "Report On The Investigation Into Russian Interference In The 2016 Presidential Election."
[68] Darren L. Linvill et al., "'THE RUSSIANS ARE HACKING MY BRAIN!' Investigating Russia's Internet Research Agency Twitter Tactics during the 2016 United States Presidential Campaign," *Computers in Human Behavior* 99 (2019): 292–300, https://doi.org/10.1016/j.chb.2019.05.027.
[69] Alexandra Dobrowolski, "'Weapons of Mass Distraction'?" Citizenship as a Regime," n.d., https://doi.org/10.2307/j.ctv2n7qxg.13.
[70] Rebeka F. Guarda, Marcia P. Ohlson, and Anderson V. Romanini, "Disinformation, Dystopia and Post-Reality in Social Media: A Semiotic-Cognitive Perspective," *Education for Information*, 2018, https://doi.org/10.3233/EFI-180209.

rates of volume and therefore readily discard what they deem irrelevant information. This balance between human attention and the volume of content challenges individuals' ability to quality assess their information inputs. When disinformation efforts queue in on which content is most likely to elude a user's discard process, then they can drastically improve the likelihood that their intended narrative is observed and absorbed. In the most impactful instances, these narratives circumvent the disposal process of large online populations within targeted groups and can go viral.[71] The referenced literature reveals the ties between social media behaviour and cognitive bias, but there remains a notable gap in papers' emphasis on cognitive exploitation within digital information operations.

Heuristics and Biases

Rational decision making can be modelled and clearly evaluated under conditions with known risks. When risks are not all known, the actor must employ 'nonrational' tools such as heuristics. Heuristics are 'mental shortcuts' that drive decision-processes under degrees of uncertainty. Heuristics can provide accurate and robust compasses under such conditions. They also, particularly under unmanageable information inputs due to quantity, may lead to sub-optimal outcomes through what are known as cognitive biases.[72] Biases are predictable in complex environments and if a given group like the IRA were to deliver an input intended to provoke the bias, they could potentially sway the target's viewpoints. There are numerous defined biases, but the biases addressed in this analysis are:

- *Confirmation bias:* a tendency to search for, interpret, focus on and remember information in a way that confirms one's preconceptions.[73]
- *Availability heuristic:* a tendency to judge the frequency or likelihood of an event by the ease with which relevant information comes to the mind.
- *Vividness bias:* the tendency of decision makers to gravitate towards salient and visually stimulating alternatives because they attract more attention and are easier to recall.[74]

[71] Dobrowolski, "'Weapons of Mass Distraction'?" Citizenship as a Regime."
[72] Gigerenzer and Gaissmaier, "Decision Making: Nonrational Theories."
[73] Bouwmeester, "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory.'"
[74] Davis et al., "Information Systems and Neuroscience."

All three of these biases appear to influence human cognition throughout digital media consumption. As mentioned in the description of network vulnerabilities, the algorithmic infrastructure driving social media platforms serves users with content and users they believe they want to see. This has the side effect of supporting like-minded user networks which, when proliferating erroneous information, risk leading the user to draw a false conclusion based on high levels of exposure to a false input through the availability heuristic. These networks, combined with the vast information accessible online, contribute to users' preference to seek out content and information that confirms their preconceived beliefs, which at times leads the user to form false conceptions as a result of confirmation bias.[75] Further enhancing this phenomenon is that in a political context, users within an online community tend to associate with common causes, which may lead to a sense of collective identity. This integration of online discourse and identity risks rendering contradictory information not simply unwelcome by the user but perceived as a potential assault on the receiving users' perceived identity.[76]

Because social media algorithms tailor content to the user, they increase the likelihood that their first impression of an emerging issue or topic may be influenced by an algorithmically prejudiced input. If the information provided by the input is false but confirms the user's preconceived beliefs, it risks leading to sub-optimal judgement through confirmation bias. Within a highly tailored social media network, in line with the 'if it bleeds it reads' adage, the most salient and emotionally charged content tends to draw the highest levels of attention and consideration, which can risk negatively influencing decision modelling through the vividness bias.[77] All four of these cognitive errors risk exploitation online through savvy network infiltration driven by an understanding of users' online mental processes. If an attacker finds success in manipulating users' perceptions through any one of these biases, he or she may have the ability to shape the perceptions of objective reality for a large user group, which can in turn influence their decision making.[78] The literature is absent specific examples of state actors deliberately exploiting cognitive bias through online disinformation. But Wang et al (2018) demonstrated in a controlled experiment by applying

[75] Dobrowolski, "'Weapons of Mass Distraction'?" Citizenship as a Regime."
[76] Marlene E. Turner and Anthony R. Pratkanis, "A Social Identity Maintenance Model of Groupthink," *Organizational Behavior and Human Decision Processes* 73, no. 2–3 (1998): 210–35, https://doi.org/10.1006/obhd.1998.2757.
[77] Erik C. Nisbet and Olga Kamenchuk, "The Psychology of State-Sponsored Disinformation Campaigns and Implications for Public Diplomacy," *The Hague Journal of Diplomacy* 14, no. 1–2 (2019): 65–82, https://doi.org/10.1163/1871191X-11411019.
[78] Till, "Propaganda through 'Reflexive Control' and the Mediated Construction of Reality."

the decoy effect that manipulating key conditions in an individual's perception can reliably influence that individual's decision-making and exploit his or her cognitive bias.[79] This paper will argue that by deliberately altering key conditions, a digital disinformation effort might be able to exploit a similar phenomenon.

Was the effort a coherent strategy?

To establish if the Kremlin's effort to influence the US 2016 election targeted cognitive biases, this paper must first assess the likelihood that the disinformation campaign followed a coherent strategy. A coherent strategy would mean that the effort and its component tactics followed a central objective and consisted of efforts intended to achieve that objective. If it did not, then it would be difficult to confidently draw conclusions concerning the intention behind the effort's component tactics. If this enquiry can determine that it did, then the tactics employed can be contextualised within the broader effort. Similar assessments have been performed in other information manipulation efforts. King et al (2014) performed an analysis to determine the likelihood that the Chinese domestic censorship apparatus followed a consistent approach concerning what content was permitted and removed.[80] The researchers determined that there were few consistent themes that the Chinese government restricted, but any identified attempts or references to organizing protest were censored.

This question can be difficult to answer in the context of the Russian election meddling campaign because of the vast scope of the effort and apparent variance in tactics and tools applied, but King et al's (2014) research demonstrated that a broad calibrated analysis of an information operation can produce clear takeaways.[81] In *the Tactics & Tropes of the Internet Research Agency*, Matney et al (2019) found numerous consistent features characterising the campaign including extensive operations targeting the African American communities, voter suppression efforts including 'text to vote' scams, and the targeting of candidates opposing Donald Trump's election including Marco Rubio, Ted Cruz and

[79] Zhen Wang et al., "Exploiting a Cognitive Bias Promotes Cooperation in Social Dilemma Experiments," *Nature Communications* 9, no. 1 (2018): 1–7, https://doi.org/10.1038/s41467-018-05259-5.
[80] Gary King, Jennifer Pan, and Margaret E. Roberts, "Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation," *Science* 345, no. 6199 (2014), https://doi.org/10.1126/science.1251722.
[81] King, Pan, and Roberts.

candidate Hillary Clinton.[82] Despite the seemingly disparate, and ostensibly discordant efforts, virtually all IRA actions can be traced to an intention to either influence the outcome of the election and/or sow social discord within the US social fabric. This intent was described in the first two parts of the senate reports on the Scourge of Russian Disinformation, Bail et al (2018), as well as the Mueller Report on the Investigation into Russian Interference (2019).[83]

Despite the variance across the IRA's tactics, the research makes a strong argument that the Kremlin's disinformation effort was coherent and informed by a keen understanding of social media platforms, cognition, and American politics. Waltzman argues that the IRA's success in influencing US users can be attributed to the speed and reach of the IRA efforts, as well as their attuned understanding of how to exploit fear and anxiety to influence Americans' cognition.[84] The Kremlin's grasp over American cognitive vulnerabilities is likely owed in great part to Russia's legacy of information warfare, and reflexive control in particular. According to Till (2020), reflexive control requires a profile and understanding of the moral and psychological dispositions guiding a group's behaviour.[85] This understanding can be informed by their upbringing, habits, and relationships. Till (2020) proceeds to explain that for reflexive control to be effective, the adversary must believe that they made the decision independently. For the Russian operator's influence to act as a 'reflex', the deceiver must emulate the enemy's behaviour and induce them to render a decision favourable to the Kremlin's objectives.[86]

The logically under-developed decision making leading to a suboptimal outcome that reflexive control is designed to trigger mirrors the definition of a cognitive bias. Due to the modern prescience of reflexive control doctrine in Russian strategy, the noted online vulnerabilities of US social media users, and the Kremlin's keen understanding and exploitation of US user behaviour, it stands to reason that its disinformation effort may have deliberately targeted cognitive biases. The selection of biases that this enquiry hypothesises

[82] Renee DiResta et al., "The Tactics & Tropes of the Internet Research Agency, New Knowledge," n.d., https://int.nyt.com/data/documenthelper/533-read-report-internet-research-agency/787lea6d5bafbf19/optimized/full.pdf#page=1.
[83] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION"; UNITED STATES COMMISSION ON SECURITY AND COOPERATION IN EUROPE, "THE SCOURGE OF RUSSIAN DISINFORMATION"; Mueller, "Report On The Investigation Into Russian Interference In The 2016 Presidential Election"; Bail et al., "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017."
[84] Waltzman, "Weaponization Inf. Need Cogn. Secur."
[85] Till, "Propaganda through 'Reflexive Control' and the Mediated Construction of Reality."
[86] Till.

were targeted was informed by the known tactics employed by the IRA. One key element informing this paper's selected is the IRA's use of micro-targeting, which involved the tailored messaging and exploitation of a specific online group. To accomplish this, the IRA developed accounts tailored to right-wing, left-wing, and African American online communities.[87] Within each group, IRA accounts altered their content in an effort to profile and propagate messaging suited to the given user group's beliefs and collective sense of identity.

To deepen the penetration of IRA accounts within these communities, the IRA repurposed and selected names similar to real-life groups. Groups included United Muslims of America, Cop Block, and Black Guns Matter.[88] Matney found that the content the IRA pushed was largely driven by current events but included references to popular culture and conspiracy theories. One study found that if the IRA were to have only operated within a single echo chamber, they would have significantly diminished their influence. By embedding their accounts within a host of online communities centred around various viewpoints, the IRA gained a foothold and the potential to influence vast many users' conversations across ideological boundaries.[89] The 2019 Senate Report identified that no other group was targeted on the scale of the African American community, which served as the target for the bulk of IRA ads, offers to work as paid activists, and the intended audience of five of the top ten most influential IRA Instagram accounts.[90] Based on this psychometric targeting and the literature informing the likelihood that the IRA targeted users' cognitive biases, this paper will subsequently review the potentially targeted biases and how they may have been manipulated based on the IRA tactics described in the literature. This is informed by the paper's methodology, which will employ a conceptualisation and operationalisation to assess the theory behind the bias, its relevance to Russian strategic doctrine, and how exploitation may have manifested in the context of the disinformation effort. The literature analysed in this paper shares common views about which tactics and objectives drove the

[87] Stewart, Arif, and Starbird, "Examining Trolls and Polarization with a Retweet Network."
[88] Robert Matney et al., "The Tactics & Tropes of the Internet Research Agency," *New Knowledge*, 2019, https://int.nyt.com/data/documenthelper/533-read-report-internet-research-agency/787lea6d5bafbf19/optimized/full.pdf#page=1.
[89] Darren L Linvill and Patrick L Warren, "Troll Factories : Manufacturing Specialized Disinformation on Twitter Troll Factories : Manufacturing Specialized Disinformation On," *Political Communication*, 2020, 1–21, https://doi.org/10.1080/10584609.2020.1718257.
[90] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."

IRA operation but have generally fallen short of explicitly describing the coherent strategy guiding the effort.

Ukraine Case Study

This paper previously noted that Russian disinformation efforts were carried out before and after the US 2016 election in the Ukraine, the United Kingdom, France, and elsewhere.[91] The Russian disinformation effort in Ukraine during its 2014 Annexation of Crimea serves as a strong case to analyse how Russia employed a similar strategy and tactics under different conditions. Indeed, former NATO Lieutenant Colonel Han Bouwmeester argues that Russia's engagement in Crimea served as a 'laboratory' through which Russia could test its modern application of reflexive control.[92] An independent researcher and expert in information warfare and communications, Georgii Pocheptsov, argued that Russia engaged in coordinated cognitive attacks on Ukrainians during the annexation to avoid additional use of force that would have negatively coloured the international community's perception of Russia's role in the conflict.[93] He explained that through cognitive manipulation, Russia sought to convince the world that the operation was a product of a voluntary reunion between Russia and Crimea. Moscow's operation was paired with a robust integration of various mediums, including television, internet, and radio in order to control the narrative reaching the public. Pocheptsov makes specific reference to the availability heuristic, explaining that all four Russian channels broadcasting to Ukraine maintained a consistent narrative in support of the Kremlin's involvement. The objective was to lead Ukrainian citizens to believe that the Russia-friendly perspective was the only viewpoint due to a high availability of Kremlin-friendly opinion and minimal alternatives. Bouwmeester observed that Russia's framing made frequent reference to potent and salient associations, such as 'The Great Russian Empire', 'The Great Patriotic War' and 'Nazi atrocities.' Based on Bouwmeester's assessment that the Kremlin used its Annexation of Crimea as an incubator for its modern employment of reflexive control, it appears that in doing so, the Kremlin sought to test the efficacy of a number of cognitive attacks tailored to exploiting the public's cognitive biases.

The Ukraine case study highlighted the Kremlin's modern commitment to information warfare and exploitation of cognitive processes. The literature clearly establishes that

[91] Cannon, "Russia's Employment of Covert Action against Western Democracies from 2013 to Present."
[92] Bouwmeester, "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory.'"
[93] Pocheptsov, "COGNITIVE ATTACKS IN RUSSIAN HYBRID WARFARE."

Russia's effort to influence the 2016 US presidential election was vast in scope, tailored to micro-targeting subgroups online, and designed to sow discord within the US. It demonstrated that Russia's modern iteration of information warfare has roots in 20th century reflexive control doctrine.[94] It identified that US media and online cognitive processes render its citizenry vulnerable to such exploitation efforts due to social media's emergent role as a primary news source. The literature also demonstrates that specific cognitive biases appear to be both vulnerable and potentially exploited by online disinformation. However, no study to this point has sewn together reflexive control doctrine, cognitive biases, and the US 2016 presidential election. An enquiry into this possibility would likely prove illustrative of the strategy behind other Kremlin disinformation efforts and would offer insight into how democracies can build resilience and combat these campaigns designed to undermine societal cohesion and electoral processes.

Methodology

The methodology will employ conceptualisation and operationalisation to identify how an exploitation of the biases might manifest. The literature and analysed dataset lacked direct references by Kremlin officials to targeting cognitive biases. Therefore, the theory surrounding the bias or heuristic will be identified and contextualised within Kremlin strategy to assess how it might have been targeted. Reflexive control theory suggests that Russia believes that it can influence the adversary's decisions by transmitting motives in order to prompt them to make the decision that serves Russian interests.[95] Traditionally, this is thought of as a battlefield approach that would lead the enemy to make a self-sabotaging decision in support of the Russian military's strategic objectives. As was evident in Russia's disinformation effort surrounding its 2014 annexation of Crimea, the Kremlin has repurposed reflexive control from the battlefield to the social media platforms through its online disinformation effort. As described in the Mueller Report (2019), the Kremlin's objective is to erode faith in the population's own government's ability to operate such that institutions become weakened under diminished public confidence.[96] In the Netherlands 2017 Annual Report of Military Studies, author Han Bouwmweester argues that reflexive control theory is

[94] Jaitner, "Applying Principles of Reflexive Control in Information and Cyber Operations."
[95] Giles, "Handbook of Russian Information Warfare."
[96] Mueller, "Report On The Investigation Into Russian Interference In The 2016 Presidential Election."

still undergoing refinement but remains a feature of Russian strategy today.[97] Considering the tailored nature of Russia's disinformation effort during the US 2016 presidential election, it stands to reason that the IRA likely employed reflexive control to influence the psychology informing American users' decision making by attempting to exploit specific cognitive biases. The enquiry will evaluate the characterisation of IRA strategy from the literature and assess the IRA's behaviour to hypothesise the specific biases targeted. It will then determine how that bias would be exploited based on the theory underlying its definition, determine a measurement mechanism, and perform a quantitative analysis based on previous research. The remainder of this section will describe each bias in the context of the IRA effort, identify how it will be measured, and identify the control dataset and how it was sampled.

Confirmation Bias

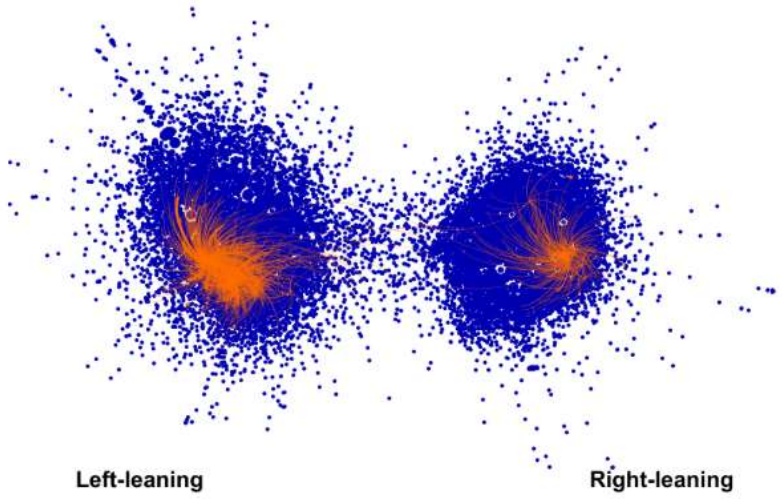
The first bias that will be assessed is confirmation bias. confirmation bias independent of Kremlin exploitation already holds a documented influence over users' online behaviour and, at the extremes, may be tied to a sense of collective identity within online groups. Due in great part to factors listed above concerning the overwhelming levels of information hosted on these platforms, users tend to seek out information that supports their preferred narratives. A consequence of this phenomenon is that fact-checking false information has an underwhelming impact on influencing these users' opinions and disinformation spread by trolls within echo chambers are more likely to gain traction.[98] Ray argues that the top five hashtags by use following the controversial Ferguson, Missouri shooting death of resident Michael Brown by a police officer divided users according to their sense of collective identity.[99] Online discourse in response to polarising events tends to split along ideological lines through hashtags and shared viewpoints. The difficulty in identifying if the IRA targeted confirmation bias is that the networks already facilitate this phenomenon independently. Disinformation targeted at a group within the networks may serve users with information aligned with their belief structures but not deliberately be intended to exploit confirmation bias.

The compelling evidence suggesting that the IRA may have targeted the mental vulnerability lies in its targeting tactics. The IRA weaponised the public's tendency to seek

[97] Bouwmeester, "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory.'"
[98] Flore et al., "Understanding Citizens' Vulnerabilities to Disinformation and Data-Driven Propaganda."
[99] Rashawn Ray et al., "Ferguson and the Death of Michael Brown on Twitter: #BlackLivesMatter, #TCOT, and the Evolution of Collective Identities," *Ethnic and Racial Studies* 40, no. 11 (2017): 1797–1813, https://doi.org/10.1080/01419870.2017.1335422.

out self-affirming information by directing content according to its perception of the groups' viewpoints. Indeed, a text analysis by Badawy (2019) revealed that conservative-facing trolls focused on topics such as refugees, terrorism, and Islam and liberal-facing trolls focused on school shootings and policing.[100] Fundamentally, it appears that it was the IRA's understanding of network effects that informed its targeted approach. Depicted in figure 2, Stewart et al (2018) used a network analysis to assess the degree to which IRA accounts infiltrated polarised networks.[101] They found that IRA accounts gained significant platforms within these conversations and used their positioning to accentuate division across political lines and fuel outrage in accordance with either side's viewpoints. While the literature lacks any clear references to the IRA's deliberate exploitation of confirmation bias, the IRA undoubtedly infiltrated US online networks with the intention of spreading false information according to users' preferred narratives and were observed influencing the views of many of these users through what Bessi & Ferrara (2016) labelled a 'cognitive inoculation'.[102]

Table 2:[103] Network visualization of IRA penetration into left and right-leaning retweet networks



Source: Stewart et al (2018), p. 5[104]

[100] Adam Badawy et al., "Characterizing the 2016 Russian IRA Influence Campaign," *ArXiv*, 2019, 1–12, https://doi.org/10.1007/s13278-019-0578-6.
[101] Stewart, Arif, and Starbird, "Examining Trolls and Polarization with a Retweet Network."
[102] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."
[103] Stewart, Arif, and Starbird, "Examining Trolls and Polarization with a Retweet Network."
[104] Ahmer Arif, Leo G. 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), https://doi.org/10.1145/3274289.

The Availability Heuristic

The availability heuristic often provides the individual an accurate estimation. Something that comes to mind easily may do so because it occurs more often. However, if the availability of the input is altered relative to the probability, as would be in the case of a manipulated effort to game a social network's hosted content, the individual would form an inaccurate perception of the issue.[105] On a social media platform, the availability heuristic might influence a user's judgement as a result of high degrees of exposure to an idea. On such a network, high degrees of exposure would be a consequence of a high volume of posts promoting a particular viewpoint. In an online setting, a user with repeated exposure to a piece of information may more likely accept that information as true regardless of its veracity due to the availability heuristic. Giles (2016) explains that reflexive control can deliver the most effective intervention during the early stages of a situation when the individual is still forming his or her understanding.[106] As Till (2020) explains, reflexive control requires the influence to occur without the target realizing that he or she is being manipulated.[107] In the context of the IRA effort, this would mean that a high degree of exposure to a sufficiently veiled IRA-fed false narrative might lead a US user to draw a conclusion or form a viewpoint intuitively without critically assessing it. The user's perception that this information is widely-present might then sub-consciously lead them to accept the input as true. Echo chambers contribute greatly to the influence of the availability heuristic because they increase the likelihood that a given user will be exposed to content consistent with the group's socio-political preferences.[108] Echo chambers therefore allow disinformation actors to feed false narratives through a cycle in which high volumes of false information are delivered a receptive community, which then amplifies the narrative, thereby increasing the availability of the information and the likelihood that users accept the input as fact.

RAND specialists, Paul and Matthews (2017), described the Russian approach to spreading messaging as, 'the firehose of falsehood' due to its emphasis on high volume and willingness to spread a narrative regardless of its veracity.[109] Considering that the Kremlin's information warfare seemed to be intended to exploit thinking patterns in a way that appears

[105] Gigerenzer and Gaissmaier, "Decision Making: Nonrational Theories."
[106] Giles, "Handbook of Russian Information Warfare."
[107] Till, "Propaganda through 'Reflexive Control' and the Mediated Construction of Reality."
[108] Petter Tornberg, "Echo Chambers and Viral Misinformation : Modeling Fake News as Complex Contagion," 2018, 1–21.
[109] Paul and Matthews, "Russ. 'Firehose Falsehood' Propag. Model Why It Might Work Options to Count. It."

to align with cognitive biases and heuristics, this paper hypothesises that one tactic that the Kremlin digital disinformation effort employed was to exploit the availability heuristic through its 'firehose of falsehood' messaging approach. Further evidence of this phenomenon is the Kremlin's employment of bots and botnets. The primary role of bots was to take an existing message and amplify its online presence.[110] A University of Southern California study found that roughly one fifth of political discourse surrounding the 2016 election may have been generated by automated bot accounts.[111] Therefore, to argue that the IRA was exploiting the availability heuristic, an analysis would need to demonstrate that IRA accounts were posting on social media at a notably higher volume than human accounts under comparable conditions. Several challenges confront this means of enquiry, including the difficulty in discerning human from IRA, as well as the fact that the vast majority of accounts associated with the IRA have been since removed from Twitter and would be inaccessible for data analysis.

The IRA Twitter Dataset

In 2018 Twitter released the most comprehensive dataset of IRA activity to date. The file includes 3,481 accounts attributed to the IRA and includes characteristics such as profile description, date of account creation, tweet content and date. According to the Twitter press release, the dataset dates to the earliest activity by these accounts in 2009 and includes 10 million tweets and two million gifs, videos, and Periscope broadcasts. This dataset was referenced in the 2019 Senate report, which asserts that the dataset contains all available tweets from IRA-linked accounts.[112] Twitter is often the preferred social media platform for research due to its open Application Programming Interface (API), which allows researchers limited but sizable access to query data relative to other platforms.[113] This enquiry will rely upon this dataset as it is the most complete file currently available attributed to the IRA and has been cited in several published academic papers.[114] The Twitter IRA dataset will allow

[110] Mueller, "Report On The Investigation Into Russian Interference In The 2016 Presidential Election."
[111] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."
[112] Gadde and Roth, "Enabling Further Research of Information Operations on Twitter."
[113] Stefan Steiglitz et al., "Do Social Bots (Still) Act Different to Humans? – Comparing Metrics of Social Bots with Those of Humans," *Social Computing and Social Media. Human Behavior* 10282 (2017): 379–395, https://doi.org/10.1007/978-3-319-58559-8.
[114] Charles Kriel and Alexa Pavliuc, "REVERSE ENGINEERING RUSSIAN INTERNET RESEARCH AGENCY TACTICS THROUGH NETWORK ANALYSIS" 6 (2019), https://doi.org/10.30966/2018.RIGA.6.; Linvill et al., "'THE RUSSIANS ARE HACKING MY BRAIN!' Investigating Russia's Internet Research Agency Twitter Tactics during the 2016 United States Presidential Campaign."

sufficient sampling for this paper to analyse the IRA's tweet volume as compared to the control analysed by Bessi & Ferrara (2016).[115]

Bessi & Ferrara (2016) provides a suitable control for an enquiry into the volume of IRA online content. The study sought to quantify the impact of social bots on the 2016 US Presidential election by collecting Twitter data from election related hashtags and keywords during the time period around the presidential debates, which represented a key period in the election cycle. The study focused on the geographic origin of the bots, their embeddedness within the election Twitter network community, their sentiment on a negative four to positive four scale, as well as the volume of tweets posted by bots vs. human accounts. Bots were detected through a Python-based software called *BotorNot* that identifies bots with an accuracy of 95 percent based on metadata, sophistication of profile, and tweet activity.[116] The longest observation window by the study was from September 26 – October 10th, 2016. Bessi & Ferrara (2016) identified the top 50,000 accounts form their dataset by volume of tweets and found 40,163 human users in that accounting for 10.3 million tweets in the above thirty-five-day time period accounting for 256 tweets per account. The study also identified 7,183 bot users producing 2,330,252 tweets, totalling 324 tweets per account.[117] Bessi & Ferrara (2016) will serve as the baseline comparison for this paper due to its specific analysis of tweet volume and finding that among the highest volume accounts, bot accounts produced 27 percent higher tweet volume than humans during significant election periods.

Sampling

This paper will rely upon the IRA dataset to compare tweets per account versus the human and bot rates identified in Bessi & Ferrara (2016). Specific account handles were removed for the purposes of anonymity. Bessi & Ferrara's Python query identified 2.78 million accounts. But because it would be impossible to analyse this large of a sample, they isolated the top 50,000 accounts by post volume, which accounted for 60 percent of the total tweets in their query.[118] The sample represents 1.8 percent of accounts in their total dataset. To provide comparable data to measure against Bessi & Ferrara (2016), this study will also isolate the top 1.8 percent of accounts in the IRA dataset by tweet volume. There are several challenges in isolating this subset. First, the IRA file is roughly 5.1 gigabytes, which exceeds

[115] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."
[116] Bessi and Ferrara.
[117] Bessi and Ferrara.
[118] Bessi and Ferrara.

the Random-Access Memory of any computing devices available to the researcher in this study. Another option is to compartmentalize the dataset into manageable file sizes. The challenge with this approach, however, is that it would be would be difficult to ensure that a sufficient sampling of high volume accounts are isolated from each file and remain confident that the selected accounts comprise 1.8 percent of the total due to the risk that duplicate accounts appear in several of the files. Therefore, the researcher employed a python script that queried the entire dataset and isolated the top accounts by volume, labelled as 'user ID,' along with the individual tweet identifier –'tweet ID—' and the date and time of the tweet. See Table 3 as an example row from the spreadsheet. These files were combined into a single spreadsheet hosting the fifty-nine accounts that represented the top Twitter accounts by volume in the dataset.

Table 3

User ID	Date/Time	tweet ID
2620614029	4/24/2017 15:10	8.5653E+17

Source: Subsection of IRA dataset. Example of collected data from top accounts by tweet volume.[119]

The research will evaluate the volume of tweets from the IRA accounts for the same time period as the Bessi & Ferrara: September 26 – October 10th, 2016. It will then divide the number of tweets by days and accounts to identify the key figure: tweets per account per day. Bessi & Ferrara (2016) found the following results for human and bot accounts within their subset for the above timeframe:

Table 4 – Human vs Bot tweet volume during significant election period

Human Accounts	tweets	Days	tweets per day	tweets per Account per Day
40,163	10,300,000	35	294285.71	7.33

Bot Accounts	tweets	Days	tweets per day	tweets per Account per Day
7,183	2,330,252	35	66578.63	9.27

[119] Gadde and Roth, "Enabling Further Research of Information Operations on Twitter."

Source: Bessi & Ferrara (2016), p. 5[120]

Their analysis showed that bot accounts during the run up to the election tweeted with a 22 percent higher frequency than humans among the top 50,000 accounts by volume. The key differences between Bessi & Ferrara's dataset and the IRA subset used in this study is that the former isolated bot accounts without identifying their origin. This means that the bots could be operated by any number of backers and could not be confidently labelled as products of the Russian operation. The IRA dataset is of course exclusively comprised of troll, bot, and sock puppet accounts and is absent organic human accounts. Because Bessi & Ferrara (2016) isolated the human and bots, this enquiry will be able to compare both the top volume IRA accounts vs. the top volume human and bot accounts. It is therefore the hypothesis of this author that the IRA dataset's tweet volume will reveal a higher rate of tweets than the human Twitter accounts discussing the 2016 presidential election in Bessi & Ferrara's (2016) dataset but likely lower than the bot accounts, as the IRA dataset is expected to also include human troll accounts that will likely tweet at a lower volume than the automated accounts.

Vividness Bias

The vividness bias influences users' understanding of a topic such that salient and stimulating inputs gain greater traction in a user's mind and may increase its influence over his or her thinking. Prior research has also tied intense emotion to vivid recall such that memories and details were easier to recount when paired with a strong emotional cue.[121] IRA disinformation seemed to be informed by users' fallibility to vividness, as it made frequent reference to highly controversial events and people and directed those references to the most polarised users. Indeed, the IRA would seize on developments related to controversial events like the Charlottesville, Virginia protests, terrorist groups such as ISIS, as well as controversy surrounding both candidates.[122] IRA accounts operating from the same computers would target the most extreme conservatives with content tailored to their perceived sensitivities such as immigration or race, as well as the most extreme liberals with content related to gun violence or race relations with the intention of inciting greater division by amplifying

[120] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."
[121] Todd et al., "Psychophysical and Neural Evidence for Emotion-Enhanced Perceptual Vividness."
[122] Zannettou et al., "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web."

controversy.[123] Direct reference to exploiting vividness bias, or any bias for that matter, on the part of the Kremlin was absent in the literature. But the IRA's efforts consistently made use of salient and emotionally charged subject matter ostensibly to manipulate the perceptions and drive division amongst the US online population.[124]

Based on this definition, an effort that sought to exploit vividness bias would likely supply its target with disinformation rife with material intended to provoke an extreme or emotional response. Giles (2016) explains that reflexive control may entail subverting the enemy's strengths by directing them towards an objective preferential to the Russians.[125] In the case of cognitive bias, this strength would be the individual's attention. By employing vivid imagery and content, the intent would be to influence a US user to form a viewpoint friendly to Kremlin objectives over the most rational decision, thereby exploiting the vividness bias. NATO's STRATCOM Centre of Excellence (COE) found that during Russia's 2014 annexation of Crimea, its disinformation relied heavily on salient and emotional imagery harkening back to potent historical references, such as 'The Great Russian Empire', 'The Great Patriotic War', 'Nazi atrocities', and 'The fascists in the Ukraine'.[126]

Similarly, throughout the US disinformation effort, the IRA employees targeted American societal rifts with content that was either controversial or struck at sensitivities related to topics like race and immigration. In doing so, IRA content sought to target the most extreme view on the political spectrum. As referenced in Matney et al, IRA accounts embedded themselves in polarised online communities to push groups further to the extremes within their political positioning. Therefore, Stewart et al found that Russia took advantage of the already existing polarity in the information sphere by attempting to push both sides further toward their respective extremes.

As previously noted, high levels of emotion have also been associated with more vivid cognitive association. Indeed, in Todd et al (2012), the authors found through a controlled experiment that when paired with an emotional cue, subjects were able to recall memories and details more vividly, thereby anchoring these impressions within their memory.[127] Based on the literature concerning reflexive control theory and the IRA's

---

[123] Stewart, Arif, and Starbird, "Examining Trolls and Polarization with a Retweet Network."
[124] Gordon Pennycook and David G. Rand, "Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning," *Cognition* 188 (July 1, 2019): 39–50, https://doi.org/10.1016/j.cognition.2018.06.011.
[125] Giles, "Handbook of Russian Information Warfare."
[126] Brinkel, *Netherlands Annual Review of Military Studies 2017: Winning Without Killing:The Strategic and Operational Utility of Non-Kinetic Capabilities in Crises*.
[127] Todd et al., "Psychophysical and Neural Evidence for Emotion-Enhanced Perceptual Vividness."

proclivity to target visceral events and content, the author of this paper hypothesises that the IRA may have targeted vividness bias within its disinformation effort. In a novel assessment, this paper will evaluate the sentiment of IRA tweets to determine the likelihood that it deliberately targeted vividness bias. To accomplish this, the paper will select and evaluate a sampling of 1,215 tweets from the Twitter IRA dataset to meet the 1,000-sampling threshold and assess the content through a Sentiment Analysis using a software known as SentiStrength. SentiStrength was specifically designed to analyse social media data and its application to this method of analysis is described below:

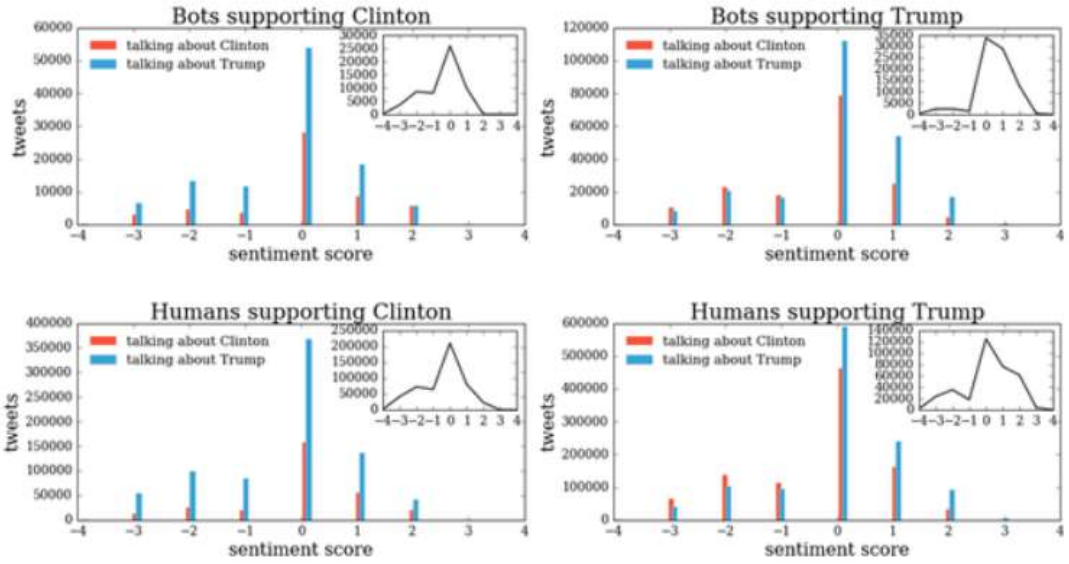
> This design choice provides some desirable advantages: first, it is optimized to annotate short, informal texts, like tweets, that contain abbreviations, slang, and other non-orthodox language features; second, SentiStrength employs additional linguistic rules for negations, amplifications, booster words, emoticons, spelling corrections, etc. Applications of SentiStrength to social media data found it particularly effective at capturing positive and negative emotions with, respectively, 60.6 percent and 72.8 percent accuracy (Thelwall, 2013). We tested it extensively and also used it in prior studies to validate the effect of sentiment on the diffusion of information in social media' (Ferrara and Yang, 2015).
>
> The algorithm assigns to each tweet t a positive S+(t) and negative S-(t) polarity score, both ranging between 1 (neutral) and 5 (strongly positive/negative). Starting from the polarity scores, we capture the sentiment of each tweet t with one single measure, the sentiment score S(t), defined as the difference between positive and negative sentiment scores: S(t) = S+(t) ? S?(t). The above defined score ranges between 4 and +4. The former score indicates an extremely negative tweet, and occurs when S+(t)=1 and S?(t)=5. Vice versa, the latter identifies an extremely positive tweet labelled with S+(t)=5 and S?(t)=1. In the case S+(t)=S?(t) — positive and negative sentiment scores for a tweet t are the same — the polarity S(t)=0 of tweet t is considered as neutral (note that neutral class represent the majority, by construction, since it contains all tweets that have equal number of positive and negative words, as well as all tweets with no sentiment labelled terms)'[128]

The positive and negative sentiment within the tweet will be added to determine an aggregate score for each tweet. This paper hypothesises that sentiment will favour extremely positive and extremely negative tweets forming a heavy-tailed distribution.

[128] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."

Before working with this dataset, it is necessary to reformat the data to remove any indiscernible tweets. Retweets were removed as well to focus only on original IRA tweets. Reformatting involved a fairly simple process that filtered out all non-English tweets. The resulting file held 1,215 tweets. Bessi & Ferrara (2016) performed a sentiment analysis of the same dataset used in the control for the availability heuristic section. Their results were discriminated for humans and bots and again based on which of the two candidates they favoured. Neither the sentiment analyses observing the bots nor the humans resulted in a heavy-tailed distribution and, in fact, the humans' distribution resulted in higher emotional load tilted toward the extremes than the automated accounts. As noted in the availability section, there are distinct differences between the Bessi & Ferrara (2016) dataset and the dataset employed in this study.[129] Foremost, their study included the top 50,000 accounts by tweet volume and distinguished humans from bots. This paper will not aggregate the IRA accounts versus human accounts and has granted no representative preference to higher volume output accounts. Still, their results present largely normal distributions and suggest that it may be unlikely for an aggregate sentiment analysis employed in this study to result in a Heavy-Tailed distribution. The selected tweets in this enquiry will be fed into SentiStrength to return a list of unique values for each tweet and will be visualized through a pivot table and bar graph depicting quantities of tweets for each sentiment value.

Table 5 – Human vs Bot tweet sentiment aggregated by supported candidate



[129] Bessi and Ferrara.

Source: Bessi & Ferrara (2016), p.8[130]

Results

Availability Heuristic

The analysis evaluated the 59 highest volume Twitter accounts, which alone accounted for 1.8 percent of the total accounts and produced 2,024,847 tweets, roughly 20 percent of the dataset's total.[131] The number of accounts was substantially smaller than the 40,163 human users and 7,183 bot users identified for tweet volume in Bessi & Ferrara (2016) and therefore tweets were evaluated on a per account/per day basis over the 35-day span. The results are displayed below relative to the results identified in Bessi & Ferrara. During this period, the fifty-nine IRA accounts produced 78,251 tweets, 2,235.74 tweets per day, and 37.89 tweets per account per day.

Table 6

This paper's analysis of IRA account frequency

IRA accounts	tweets	Days	tweets per day	tweets per account per day
59	78,251	35	2,235.74	37.89

Bessi & Ferrara (2016)

Human accounts	tweets	Days	tweets per day	tweets per account per day
40,163	10,300,000	35	294285.71	7.33
Bot accounts	tweets	Days	tweets per day	tweets per account per day
7,183	2,330,252	35	66578.63	9.27

Source: The author's analysis of the top 1.8 percent accounts by volume compared to Bessi & Ferrara's top accounts.[132]

The results in Table 6 reveal a stark disparity between the output of the IRA accounts and both the human and bot accounts isolated in Bessi & Ferrara (2016).[133] The human accounts

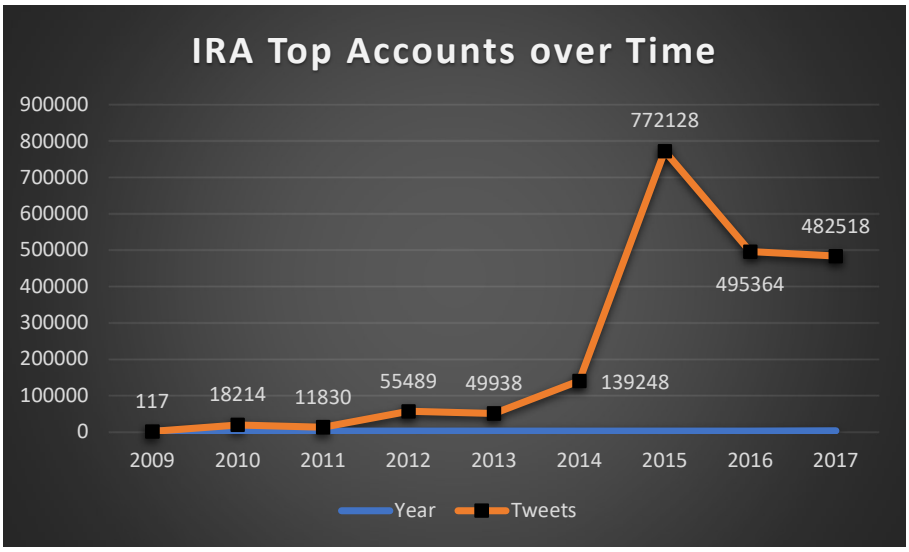
[130] Badawy et al., "Characterizing the 2016 Russian IRA Influence Campaign."
[131] Gadde and Roth, "Enabling Further Research of Information Operations on Twitter."
[132] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."; Gadde and Roth, "Enabling Further Research of Information Operations on Twitter."
[133] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."

in their subset tweeted roughly at a fifth of the rate of the accounts in this paper's dataset and the bot accounts at a quarter of the rate. The author of this paper hypothesised that the IRA dataset would exceed the output of Bessi & Ferrara's (2016) humans but fall below the number of tweets that the bot accounts produced. This turned out to be the case for the human accounts, but the IRA tweet output dwarfed the numbers of both humans and bots in Bessi & Ferrara.

There are several plausible explanations for the disparity. The IRA accounts were attached to a known coordinated effort to influence the election. As referenced in the introduction, the IRA accounts were staffed by over 400 employees working around the clock to publish content across social media platforms, in addition to the bots the bot accounts the organisation employed. Ostensibly, this effort greatly exceeded the output of non-IRA human and bots due to its intentions to steer the Twitter political conversation. Another key component is the relevance of the pre-election period (September 26 – October 10th, 2016) to each dataset. Indeed, qualitatively, it is well documented that the IRA placed a great importance on the period leading up to the election as the objective of the effort was in part to steer the election's outcome. Table 7 confirms this observation, demonstrating that 2016 was among the IRA's highest volume periods. Though the bots depicted in Bessi & Ferrara's dataset were tweeting about election-related topics, these bots could have been employed through any number of initiatives, many of which may have intersected with the election but not been fundamentally driven by election influence.[134] For example, if a bot had been employed by a company's marketing effort to drive digital presence over the course of years, that bot may have touched on election related content without that period holding greater significance than other time periods.

Table 7: Histogram depicting activity of top IRA accounts from onset to removal.

[134] Bessi and Ferrara.



Source: Author's analysis of top 1.8 percent of IRA Twitter accounts by tweet volume.[135]

Another possible explanation as to why the IRA output greatly exceeded Bessi & Ferrara's (2016) data is that the twenty-three election related keywords they used to query Twitter's API were insufficient and missed sizable portions of the Twitter conversation. Although, this is unlikely as the keywords emphasised both candidates, the third-party candidates, and applied keywords specific to the debates that occurred within the time frame. The data seems to confirm that the IRA did leverage high output as a key component of its disinformation effort.[136] It appears that the Kremlin held a keen understanding of Twitter network infrastructure and how to leverage tweet volume to exploit that infrastructure, a subject that will be explored in greater depth in the Discussion section of this paper. Furthermore, considering that reflexive control doctrine endeavours to exploit the adversary's cognition without their recognition, it remains plausible that the IRA recognised that high posting volume would increase the algorithmic significance of the Twitter content and influence the American user's perception of the accuracy of a false narrative based on the content's artificially increased availability.

Vividness Bias

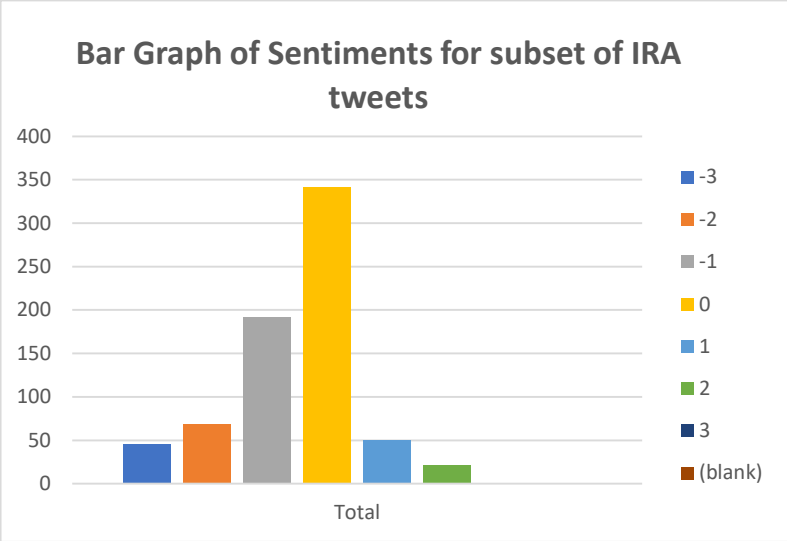
[135] Gadde and Roth, "Enabling Further Research of Information Operations on Twitter."
[136] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."

The sentiment analysis performed on the subset of IRA tweets returned aggregate values from -3, representing the most negative, and 3, representing the most positive. Results are depicted in a value table and bar chart in Table 8.

Table 8

Count of Sentiment Values

-3	-2	-1	0	1	2	3
46	69	191	342	50	22	1



Source: Author's analysis of SentiStrength derived tweet sentiment of IRA subset.[137]

The sample size in this paper's sentiment analysis is significantly smaller than in Bessi & Ferrara (2016) and is not aggregated by bot/human and Clinton/Trump. Similar to all results in Bessi & Ferrara (2016) represented in Table 5, negative tweets were more common than positive, and the bulk of the distribution favoured an aggregate sentiment score of zero. One issue with the comparison was that Bessi & Ferrara (2016) did not provide the data

[137] Bessi and Ferrara; Gadde and Roth, "Enabling Further Research of Information Operations on Twitter."

underlying their graphs, so this author was unable to compare the two results using a chi-squared analysis.[138] This confounding factor, however, did not inhibit the author's ability to evaluate this study's results relative to the hypothesis, which was rejected by the results. The results show no substantive variances in emotional load relative to the human subsection identified in the control study. Similar variances in sampling exist with the vividness bias subset as the availability heuristic analysis relative to the control. The sampling is significantly smaller than the control and is not isolated based on high volume. These results reject the hypothesis that the analysis would result in a heavy-tailed distribution. This outcome could be a result of the totalling process the software undergoes. A statement intended to be negative may be weighted with positively scored words and may not precisely reflect the overall intention of the tweet. One notable aspect of the results is a notable favouring of negative tweets over positive.

As noted in the availability heuristic section, the IRA effort was nuanced, and accounts were appropriated for varying functions. The overarching effort of certain sock puppet accounts within online networks may have been to exploit emotional and vivid content, but a sentiment analysis of the individual tweets may have failed to capture that overall objective.[139] Furthermore, emotional extremes are only one aspect of the vividness bias. A strongly worded tweet describing a controversial event, such as a police officer involved shooting, a protest, or a statement about either candidate could have provoked a vivid image without carrying a high emotional load, positive or negative. Exploiting vividness bias may prove a difficult effort to measure and may be better suited to a content analysis of the sorts of topics at which IRA accounts directed their emphasis. Based on Stewart et al's (2018) analysis, it appears that IRA employees closely followed American politics and micro-targeted groups by emphasising content towards highly controversial topics that were likely to provoke a visceral reaction.[140] Below in Table 9 is a sampling of visceral tweets from this study's dataset that received generally neutral scores through the SentiStrength software annotated with explanations. Notably, all four of the listed tweets are at least loosely based in fact.

Table 9

[138] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."
[139] Zannettou et al., "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web."
[140] Stewart, Arif, and Starbird, "Examining Trolls and Polarization with a Retweet Network."

Annotated Sampled tweets from the Dataset:

Fake Liberal Democrat says schools should be able to suggest PROSTITUTION as a career to pupils https://t.co/QSHxO5INeX #Fake
-1

'Fake' was a common term popularized by candidate Trump to refer to critical media sources as well as opponents. The tweet is ostensibly targeted to a UK audience and references a UK politician named Dennis Parsons' suggestion that academic institutions should promote prostitution as a viable career option.[141] Though the US Democratic party is not referred to as 'Liberal Democrats,' the tweet could be interpreted as a reference to the US party paired with an adjective. The intention of the tweet would be to play up on users' preconceived beliefs that the left-wing holds 'immoral' political positions that might lead them to promote institutionalizing the sex industry. This tweet registered as only slightly negative through SentiStrength.

George Soros who sent instructions via #HillarysEmails on how to organise riots - is now Hillary's top donor at $13M https://t.co/GrGsxy4x9F
0

George Soros is a billionaire investor and notable democratic donor.[142] He is frequently accused by far-right outlets of financing violent demonstration agitators. Therefore, this tweet, likely targeted at the IRA's right-leaning audience is presumably intended to insight a visceral response as it notes that an alleged democratic agitator is now the principal donor for candidate Clinton, yet this tweet registers as neutral sentiment.

This pic was dug up by Hillary's campaign in 2007 questioning where Obama was born! #HRCOriginalBirther #Birtherism https://t.co/aYayinogTF
0

[141] "Cheltenham Lib Dem Chair Quits after Sex Worker Debate," *BBC*, September 19, 2016, https://www.bbc.com/news/uk-england-gloucestershire-37404894.
[142] Deborah D'Souza, "Top 10 Contributors to the Clinton Campaign," in *Investopedia*, 2019, https://www.investopedia.com/articles/investing/033116/top-10-corporate-contributors-clinton-campaign.asp.

This tweet refers to the 'birther' conspiracy theory that then President Barack Obama was born outside of the US and ineligible to be president that candidate Trump had referenced and eventually dismissed.[143] The tweet is arguing that candidate Clinton investigated this same question during her first run for president. Clinton served as one of President Obama's Secretaries of State and counted Obama as a key endorsement. If she were to have legitimately initiated an enquiry questioning his eligibility to become president, it likely would have become a sizable scandal.

Donald Trump Jr. likens Syrian refugees to poisoned skittles https://t.co/7U1WGe4CAF
-2

This tweet holds the most polarised score of the selection and represents the only tweet likely targeted towards the IRA's left-leaning audience. The tweet references a legitimate tweet by candidate Trump's son that shared a meme used to justify his father's proposed policy to ban admission into the US from numerous Muslim-majority countries.[144] Although the tweet scored as more negative than the others, the intention of sharing the tweet was likely to present the candidate's son, who was deeply involved in the campaign, as callous and dehumanizing immigrants by comparing them to a candy.

Source: Author's analysis of tweets from subset of IRA Twitter dataset.[145]

This qualitative review is not meant to challenge the legitimacy nor accuracy of the SentiStrength software, which is well documented. It is instead intended to propose that a sentiment analysis may have failed to capture the IRA's tweets' intention to provoke an emotional or visceral response. The examples provided in the qualitative analysis demonstrate that a phrase may be intended to provoke an emotional response without carrying heavy emotional load in its phrasing. This demonstrates the importance of pairing an analysis with a sentiment analysis software with a qualitative review that analyses the word choice, overall message, and socio-political context of the tweet. Following this logic, the Discussion section will rely on additional literature characterising the IRA's effort to analyse

[143] Kyle Cheney, "No, Clinton Didn't Start the Birther Thing. This Guy Did.," *POLITICO*, September 16, 2016, https://www.politico.com/story/2016/09/birther-movement-founder-trump-clinton-228304.
[144] Christine Hauser, "Donald Trump Jr. Compares Syrian Refugees to Skittles That 'Would Kill You,'" *The New York Times*, September 20, 2016, https://www.nytimes.com/2016/09/21/us/politics/donald-trump-jr-faces-backlash-after-comparing-syrian-refugees-to-skittles-that-can-kill.html.
[145] Gadde and Roth, "Enabling Further Research of Information Operations on Twitter."

how it selected its targets, infiltrated networks, and what subject matters it favoured to maximise the effort's potency.

Discussion

Following mixed results in the quantitative analysis relative to the hypotheses for each bias, this paper will assess the results in the context of the IRA effort and the likelihood that biases were targeted based on reflexive control doctrine. This section will therefore reflect on the likelihood that the effort followed a coherent strategy, identify which tactics and characteristics of the effort were implemented to meet the objective of the coherent strategy, and assess the results' implications relative to the overall strategy. It will focus principally on Matney et al (2019) and Paul And Matthews' (2017) 'firehose of falsehood' characterisation of the effort.[146] The IRA effort employed volume and swarmed the most controversial topics to maximise its accounts' presence and virality. The paper will re-examine how these features demonstrated that the Kremlin implemented a coherent strategy and tie these determinations to the biases and theory by examining how the IRA's high volume approach fit into reflexive control doctrine and the 'firehose of falsehood' characterisation, re-evaluate the likelihood that the effort exploited confirmation bias, as well as how IRA behaviour may suggest that it sought to exploit vividness bias absent evidence of strong emotion within IRA tweets. Matney et al (2019) describe the substantial overall operation.

> The scale of their operation was unprecedented — they reached 126 million people on Facebook, at least 20 million users on Instagram, 1.4 million users on Twitter, and uploaded over 1,000 videos to YouTube. As Department of Justice indictments have recently revealed, this manipulation of American political discourse had a budget that exceeded $25 million USD and continued well into 2018. IRA documents indicate the 2017 operational budget alone was $12.2 million US dollars.[147]

The effort also boosted around 73 million engagements on Twitter. Clearly, the scope and reach of the effort was substantial. However, a large operation can simply be designed to maximise impact. To determine if the effort specifically targeted the vividness bias, confirmation bias, and that its high volume approach, confirmed in the results of this paper,

[146] Paul and Matthews, "Russ. 'Firehose Falsehood' Propag. Model Why It Might Work Options to Count. It"; Matney et al., "The Tactics & Tropes of the Internet Research Agency."
[147] Matney et al., "The Tactics & Tropes of the Internet Research Agency."

demonstrates that the IRA sought to exploit the availability heuristic, the paper must successfully analyse and characterise the broader goal and strategy defining the effort. This requires reflecting on whether the effort executed a coherent strategy.

If one were to argue that the IRA's disinformation effort did not follow a coherent strategy, he or she would likely assert that the effort was broad and multi-faceted and that an analysis could draw any number of connections to the IRA's tactics but none of which constitute a clear overarching framework. It is true that Russia targeted many different groups and published immense volumes of social media content. Matney et al (2019) also makes clear that the IRA effort was run like a 'sophisticated marketing agency'.[148] It posted large swaths of content that neither included false information nor appeared directly tailored for any specific voter bloc. However, research taking a broad view of the IRA effort has proposed several consistent features that suggest that the orchestrated effort followed clear encompassing guidelines. These traits include the substantial financial investment that saw a full-time staff posting across platforms in large volumes, a keen understanding and intention to exploit the algorithmic infrastructure across platforms, a degree of effort to disguise the accounts and make them appear as if they were US based, and a clear preference to support the election of candidate Trump.[149]

Paul & Matthews (2017) argued that the IRA effort was fast, repetitive, and uncommitted to consistency. They note that this approach runs at odds with traditional communication influence operations published by government and defence sources, which highlight the value of honesty, credibility, and consistency in narratives.[150] Perhaps by implementing an approach founded in these ideals, the Russian effort would have found greater success and offered clearer metrics for the efficacy of the effort. But this is unlikely and such an approach would run counter to Russian objectives and communication methods. As referenced in the introduction of this paper, Cormac and Aldrich note that the Russian response to public accusations of wrongdoing is to justify, dismiss, normalise, and redirect blame. This approach is designed to obfuscate culpability even when the evidence against Russia is overwhelming, as was the case in the downing of Malaysian flight MH17 over Ukrainian airspace as well as Russia's 2018 poisoning of former MI6 spy, Sergei Skripal, and his daughter Yulia. Cormac and Aldrich described this tactic as 'implausible deniability'.[151] Such brazen dismissal and

[148] DiResta et al., "The Tactics & Tropes of the Internet Research Agency, New Knowledge."
[149] Paul and Matthews, "Russ. 'Firehose Falsehood' Propag. Model Why It Might Work Options to Count. It."
[150] Paul and Matthews.
[151] Cormac and Aldrich, "Grey Is the New Black: Covert Action and Implausible Deniability."

disregard for consistency is enabled by the social media environment that has permitted alternative and non-verifiable sources to flood the feeds of users, many eager to consume content from sourcing regardless of credibility that supports their prior socio-political views. In this sense, the Russian disregard for consistency paradoxically remains one of the most prominent consistent features of its systematic effort.[152]

To further make sense of the IRA's grand strategy, it may serve to re-imagine the effort by starting with the structure and capabilities of the social media platforms and how they could serve the effort's overall objectives, rather than how each micro-tactic on the platform aligned with Russian geopolitical views or its overall election objectives. Through this lens, the infrastructure would drive the tactics such that they maximised the IRA's presence and access to chosen online communities to amplify discord and sway the outcome of the election. This was enabled by Russia's keen awareness of the platforms' network infrastructure. As a result of this understanding, the IRA could identify and target specific online communities to engage with and ensure they were exposed to IRA content.[153] With groups identified, the IRA could employ keywords and hashtags that the target community would already be engaging in or drawn to. The key to influencing these groups, however, was the topic selection. The topic selection of IRA tweets best illustrates how the IRA employed micro-targeting to influence the thoughts and views of its target groups.

The written work on reflexive control reveals that Russia has been well aware that one cannot influence the adversary simply by telling or showing them what you hope to make them think. The doctrine, which has influenced and appears to continue to influence Russian strategy, seeks to manipulate the adversary's decision-making process such that they render decisions in the Kremlin's interest and without knowing they had done so.[154] If reflexive control informed the IRA's disinformation effort, it stands to reason that it sought to manipulate US Twitter users' views by exploiting specific biases to which the population was susceptible as a consequence of the vulnerabilities—media, algorithmic, and cognitive--. The biases analysed in this paper were selected because they appeared to be targeted based on the IRA's known tactics. Matney et al (2019) provides a strong description of how these tactics were employed and to which groups.[155] This explanation has allowed this research to

[152] Paul and Matthews, "Russ. 'Firehose Falsehood' Propag. Model Why It Might Work Options to Count. It."
[153] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."
[154] Giles, "Handbook of Russian Information Warfare."
[155] Matney et al., "The Tactics & Tropes of the Internet Research Agency."

examine those efforts to review how the tactics may have aligned with the selected biases and connect that analysis to the data in this paper's results.

Matney explains that only about 6 percent of IRA tweets mentioned either political candidate by name. In fact, they found that less than 10 percent of the IRA's original posts were about the election, though the group did place great emphasis on the pre-election period and on engaging with other accounts discussing the election.[156] These numbers appear puzzling considering that one of the key characteristics of the broader effort was an overall goal of influencing the election's result. However, following with the IRA's understanding of how users' cognition processes on social networks, Matney et al (2019) explain that much of the IRA's efforts were a function of gaining status and trust within the networks. By posting ostensibly innocuous content and tailoring that content to the community's views, the IRA's accounts could appear as an organic voice within the broader conversation and more covertly influence the target community's views of current events and pertinent socio-political issues. At different stages of the IRA accounts' lifespan, the accounts published content and engaged in ways that sought to confirm the target community's previously held beliefs, posted and engaged in notably high volumes, and focused on highly contentious subject matters that were likely to provoke visceral reactions by the human users. The remainder of this section will focus on Matney et al's (2019) analysis of these tactics and how their findings tie into the three biases explored in this study.

The first bias evaluated in this study was the confirmation bias. Matney et al explains that the 'The themes selected by the IRA were deployed to create and reinforce tribalism within each targeted community; in a majority of the posts created on a given Page or account, the IRA simply reinforced in-group camaraderie'.[157] The term 'tribalism' is meant to describe a shared in-group identity that sees members in the group sharing a greater sense of community and those outside viewed to some degree as the other. Identity and beliefs are deeply intertwined in this sense such that beliefs within a group are often informally assigned and adopted amongst members. As Matney et al (2019) describe, the IRA's posts would be designed to share information endorsing the group's beliefs or share negative information stoking the sense of otherness directed towards those outside of the tribe and particularly viewed as at odds with the community's beliefs.[158] Notably, the sizable majority of posts had a corresponding target group. See Table 10 with examples from Matney et al (2019) of

[156] Matney et al.
[157] Matney et al.
[158] Flore et al., "Understanding Citizens' Vulnerabilities to Disinformation and Data-Driven Propaganda."

frequent posting themes and this author's analysis of the corresponding groups to which the posts sought to appeal. This paper was unable to identify an adequate quantitative means to assess the likelihood that the IRA tweets targeted users' confirmation bias. But the literature shares a confident assessment that IRA accounts focused on infiltrating groups and exploiting in-group belief structures. Referring to figure 2, it is abundantly clear that IRA accounts found footholds at the centre of retweet networks. Based on its topics of emphasis and success in infiltrating these networks, it is quite likely that the IRA managed to do so by deliberately exploiting the cognition influencing the users' behaviour by tailoring content to appeal to the users' already held beliefs through the confirmation bias.

Table 9

Posting Subject	**Corresponding Group**
Black culture, community, Black Lives Matter Organisation/Movement	African American community and left leaning
Blue Lives Matter	Law Enforcement and right leaning
Southern Culture (Confederate history)	Subset of Southern community, right-leaning
Christian	Christian community, right-leaning
'Red Pill'	Anti-feminist; men's rights advocates
Pro-Bernie Sanders and Jill Stein	Progressive left/environmentalists; disaffected democrats

Source: Matney et al (2019), p. 11[159]

Within each social theme, IRA posts endeavoured to erode the middle and push users to the extremes within their respective political leanings. The second bias evaluated in this section will be the vividness bias. The Sentiment Analysis in this paper did not yield a heavy-tailed result that would reveal a preference for either side of the emotional extremes. A qualitative review of the dataset and methodology, however, suggest that a Sentiment Analysis may not have offered an effective means to measure the likelihood that the IRA effort sought to exploit the vividness bias. Vividness bias involves granting cognitive preference to more salient facts and ideas. Emotion could represent a component of vividness, as noted in the Todd et al (2012) study, but a sentiment analysis would be unlikely

[159] Matney et al., "The Tactics & Tropes of the Internet Research Agency."

to capture the salience of the topic, vivid imagery, and contextual clues that may render a post particularly salient.[160] A strong example of the IRA's employment of salient Tweet content was in its disinformation effort directed at the African American community. The IRA's objective in targeting the African American community appeared to be to both cultivate and stoke within the community a sense that it should fear for its safety and that it was being in some sense discarded by American society. One way it promoted this sense was by covertly paying African Americans citizens to teach self-defence, likely in an effort to encourage the view that such classes would be necessary for ensuring their safety and well-being.[161] The two following tweets from Matney et al (2019) offer examples of strongly salient statements that fail to reach the tails on either end of a sentiment analysis (featured below followed by their corresponding aggregate score).

- White people invent tools for killing, this Black child is inventing a tool for saving lives (-1)
- St. Louis mother wants answers after 'hideous' photo of officer posing with her dead son surfaces (-2)[162]

Both above tweets were directed toward the African American community and carry heavy meaning and imagery. The tweets serve to provoke the sense of tribalism and otherness described in the confirmation bias paragraph of this section. They also both refer to death and a clear sense of callousness directed towards death of a member of one racial group by another. The second tweet also happened to gain the greatest engagement of any IRA Twitter account posing as a member of the African American community.[163] A qualitative review of the entire IRA dataset will also find many tweets discussing mundane news or other seemingly innocuous content. This highlights the nuance characterising the effort. The IRA accounts sought to infiltrate networks and become prominent voices within those networks. This meant that posting a barrage of highly salient content may have proven ineffective to growing a grassroots following. It is likely, therefore, that IRA accounts used largely inoffensive content directed towards target groups at the onset and *then* used their positioning

[160] Todd et al., "Psychophysical and Neural Evidence for Emotion-Enhanced Perceptual Vividness."
[161] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."
[162] Matney et al., "The Tactics & Tropes of the Internet Research Agency."
[163] Matney et al.

and algorithmically trusted standing to spread the more vivid content intended to influence and sow discord across online communities. This would have meant that early tweets may have not sought to provoke the vividness bias but tweets disseminated after status had been attained might direct a greater focus towards influencing the given community's views through vivid and provocative posting.

Paul & Matthews (2017) explain that when all other factors are equal, messages observed in greater volume tend to be more persuasive. Additionally, the larger the variety of sources contributing to the volume, the more persuasive the messaging tends to be. The reasons they provide are that the volume can distract from other dissenting messages, it can consume the target's attention, the more varied the channels the wider the scope of their potential audience, and the multiplicity of sourcing allows the information to appear more credible.[164] They add that particularly when the target's interest is low, the message will carry more weight if it is endorsed by a large number of messengers. The IRA campaign seemed to be informed by this wisdom as it employed efforts intended to achieve nearly all of these aims by posting in high volumes, across many accounts, and on many platforms. The fact that the IRA prioritized volume was confirmed in the quantitative analysis within this paper during the period just prior to the election. Accordingly, Paul & Matthews (2017) endorse this paper's review of the availability heuristic such that high-volume posting was likely intended to persuade users of the disinformation and non-falsifiable narratives that IRA accounts sought to promote. Based on the current prescience of reflexive control doctrine to Russian information warfare strategy and the IRA's tactics, it stands to reason that the IRA may have targeted the three biases. This point will require further assessment still, but the evidence found in this paper suggests that the IRA understood how to manipulate users' cognition online and used that knowledge to drive a tactical approach that targeted these biases to weak US societal cohesion and influence the outcome of the election.

Conclusion

This paper sought to deepen the collective understanding of the cognitive underpinnings of the modern Russian digital disinformation strategy. The effort targeted at

[164] Paul and Matthews, "Russ. 'Firehose Falsehood' Propag. Model Why It Might Work Options to Count. It."

the US 2016 presidential election has been explored extensively by government, the media, and the academic community alike. For democracies to better secure themselves against this threat, they must first understand precisely what made it unique and allowed it to reach such a broad audience on such a large scale. Much has been learned through these inquiries concerning the scope and intensity of the IRA and its daily operations in its attempt to manipulate American opinion and sway the outcome of the election. The IRA operated across social media platforms and played the part of users across many different groups to infiltrate and influence the conversations of real Americans online. While it remains to be seen exactly how effective the Kremlin's disinformation effort was at realizing its goals, the operation's footprint was strongly felt, and IRA accounts infested the cores of networks hosting American online socio-political conversations. This research identified three factors that allowed the Kremlin to achieve this unprecedented influence against an adversary's populace. The IRA exploited US *civic and media* vulnerabilities to promote distrust towards the media and other social groups, users *cognitive* fallibility to tailored disinformation on social media networks, and *algorithmic* infrastructure on social networks that rewards mechanisms like high volume posting and serves users with content that confirms their preconceived beliefs. The IRA was able to exploit these vulnerabilities collectively to maximise its impact.

To understand if the IRA intended to exploit cognitive processes, this paper needed to first assess if the effort was guided by a coherent strategy. This would allow the researcher to contextualise and ascribe meaning to the litany of tactics and behaviours comprising the operation. The bulk of the literature as well as the far-reaching Senate paper seemed to confirm that the IRA held the primary objectives of sowing discord throughout the US population and influencing the election in favour of candidate Trump.[165] Therefore, the various tools including micro-targeting, the use of trolls, sock puppets, bots, and botnets across platforms were deployed in order to serve that goal. The paper lacked specific references to cognitive biases or mental processes on the part of the IRA or Kremlin decision-makers. However, Russian reflexive control doctrine proved deeply insightful for explaining how previous efforts of Russian disinformation might be repurposed for the digital environment. reflexive control is fundamentally designed to influence the adversary's decision-making without their awareness, much like cognitive bias. information warfare expert and retired Lieutenant Colonel Timothy Thomas notes that reflexive control remains

[165] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."

influential to Russian strategy and asymmetric warfare today.[166] The literature explains both that the IRA's efforts may have sought to exploit cognitive processes online and that social media users are vulnerable to cognitive bias. This insight informed this paper's efforts to identify if the IRA endeavoured to exploit cognitive biases and if so, which ones. The biases were chosen based on analysis of IRA tactics, which emphasised high volumes of tweets, network infiltration, and discourse emphasising the most controversial social subjects. Consequently, the paper hypothesised that high volume could be employed to exploit the availability heuristic, network infiltration could have been used to exploit the confirmation bias, and the emphasis on the most contentious social issues could have been targeted at the vividness bias.

The paper then sought to perform an assessment of each bias to identify if and to what degree each may have been exploited by the IRA effort. The author performed a qualitative analysis of the confirmation bias to assess how the IRA may have stoked confirmation bias within retweet networks. The key factors suggesting that the IRA may have pursued this bias was the IRA's concerted effort to infiltrate online communities and flood them with content endorsing the group's consensus views. There are other explanations for why the IRA may have pursued such an approach absent exploiting the confirmation bias including that they simply intended to gain group recognition to later sway the group's viewpoints, as noted in Matney et al (2019), and saw endorsing the collective viewpoints as a means to do so. However, the IRA's commitment to infiltrating online social groups and retweet networks rather than targeting the entire US online population suggests that it sought to exploit collective reasoning and belief.[167] This research delivered a unique examination of the connection between the bias and the broader IRA efforts. At the least, the evidence found in this paper suggests that further research should examine if IRA efforts were targeted at this bias.

The next cognitive process was the availability heuristic. The author assessed that if a group were to exploit the availability heuristic, of would endeavour to flood users with high volumes of a single narrative stemming from a variety of different sources on the given platform. This paper identified a suitable control dataset from Bessi & Ferrara (2016), who had identified a bot and human population and assessed each's volume during the pre-

[166] Bouwmeester, "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory.'"
[167] Matney et al., "The Tactics & Tropes of the Internet Research Agency."

election period.[168] If the IRA were deliberately employing volume to game the user base's cognition, it would be likely that it would post at a higher volume than each. The top 1.8 percent of tweets by volume from the IRA's dataset was compared versus the control and was found to tweet at a significantly higher volume. This confirmed the Senate's findings that IRA employees were publishing social media posts at high volumes to meet quotas that were set in accordance with the IRA's directive.[169] It is difficult to distinguish between the IRA specifically targeting the availability heuristic and posting at high volumes solely for the purpose of improving their accounts' notorieties by exploiting Twitter's algorithm. However, even if it had been primarily targeting the algorithm, such behaviour suggests that the IRA was aware that volume would increase the prominence of its posts and consequent likelihood that a given individual would be exposed to those posts.[170]

The final bias this paper evaluated was the vividness bias. The bias and the research surrounding its influence suggest that salient thoughts come more easily to mind and may influence the individual's perception that the salient memory may have been more accurate. This research relied on a previous study by Todd et al (2012) that explained that intense emotion increased perceptual vividness.[171] Therefore, this author determined that one way to measure if the effort had targeted the vividness bias, the IRA tweet content would have maintained an unusual degree of emotional load relative to the control. Once again, Bessi & Ferrara (2016) was selected as this paper's control as both analyses carried out a sentiment analysis using the software SentiStrength to analyse the sentiment scores for each's tweet datasets.[172] The results of this paper's analysis revealed no fat-tailed distribution favouring the emotional extremes on either side, which rejected the author's hypothesis.

To better contextualise these results the author performed a qualitative review of a subset of tweets form the analysed dataset to identify if there were other means of publishing salient content that may have eluded the sentiment analysis software. The author identified several tweets that referred to highly contentious and salient topics that returned relatively neutral values. In the Discussion section the author examined Matney et al's (2019) work to identify how this observation may have fit within the context of IRA tactics. This review revealed that in many cases, the IRA sought to create division by highlighting the most

[168] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."
[169] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."
[170] Paul and Matthews, "Russ. 'Firehose Falsehood' Propag. Model Why It Might Work Options to Count. It."
[171] Todd et al., "Psychophysical and Neural Evidence for Emotion-Enhanced Perceptual Vividness."
[172] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."

contentious issues in order to provoke a visceral response and in the case of the African American community, to fuel a sense that members should fear for their safety. Matney et al (2019) included a 'roster of themes' favouring social issues at which the IRA most consistently directed its content.[173] These themes all related to socio-political groups within the US and focused on the most polarising topics featured in conversations within these groups. This observation drove an important point about the IRA's efforts. The IRA appeared to drive division across groups and push users within groups to the extremes of their collective views. By understanding this objective, one can deduce that the IRA did seek to provoke a response from the user base by targeting highly visceral content that emphasised the group's identity and emotion through content rather than emotionally charged phrasing. This further suggests that the IRA may have sought to exploit user groups' cognition by favouring vivid content and subject matter.[174] In accordance with these findings, the paper will end by evaluating how the evidence of the IRA's efforts to manipulate US users' cognition can be employed to inform strategic solutions designed to combat their efforts and empower users to recognise and discard online disinformation.

Solutions

Russian digital disinformation has had a pernicious impact on Western democracies over the last few years. Moscow's longstanding *dezinformatsiya* (disinformation) approach has been reimagined for the digital age and has yet to be met by the West with an effective concerted response that would shield its voting populations from the scourge of this digital brand of disinformation.[175] The literature lacks any confident assessments that the Russian effort had a meaningful influence on the election's outcome. But if it had, this cost-effective largely unhindered class of information warfare would have yielded Moscow a greater geostrategic victory than it could have realised by conventional means. Furthermore, even absent clear metrics indicating that the operation successfully manipulated the election outcome, the doubt amongst the American media, political class, and voting public concerning the election can on its own right create an insidious impact on the general population's confidence in its own institutions and electoral processes.

[173] Matney et al., "The Tactics & Tropes of the Internet Research Agency."
[174] Flore et al., "Understanding Citizens' Vulnerabilities to Disinformation and Data-Driven Propaganda."
[175] Paul Baines and Nigel Jones, "Influence and Interference in Foreign Elections," *RUSI Journal* 163, no. 1 (2018): 12–19, https://doi.org/10.1080/03071847.2018.1446723.

It is also clear that this new brand of Russian information warfare did not limit its targets to simply the US, nor did it stop targeting US voters after the election had been decided. Indeed, after the Kremlin's preferred candidate was elected, the IRA began proliferating posts protesting the Electoral College and calling for Trump's impeachment.[176] As the introduction states, Russia targeted other Western states, as well as Ukraine with a focus geared toward elections and major global events involving Russia, such as the downing of the Malaysian flight MH17 and the World Anti-Doping Agency's finding that Russia had maintained a state-sponsored athletic doping program.[177] While the degree of success Russia achieved in these cases in achieving its goals remains uncertain, its continued pursuit of online disinformation may speak to its own estimation of the program's efficacy. Due to its low cost, the challenge of attribution, and incomparable ease of access relative to conventional means of warfare, it is likely that other states and motivated actors will join those already engaged by pursuing geostrategic objectives through online disinformation.[178] As it stands, the West appears woefully unprepared to meet the potential scope of this challenge.

Current efforts to combat disinformation range across Western nations. Social media platforms have taken measures to adapt to the threat and all accounts within the IRA dataset have been purged from the platform. But Russia will undoubtedly continue to innovate and pursue alternative means to pursue its objectives on the platforms. The US has targeted policy specifically to deal with 'deepfake' technology, which involve forged recordings manipulated video that are weaponised to deceive audiences.[179] Broad investigations, including those cited in this report, assessed the scope and impact of disinformation and the State Department's Global Engagement Centre (GEC), which is tasked with combatting global disinformation, has requested an extra $138 million in budgetary considerations to improve its capabilities to meet the growing challenge.[180] Europe, for its part, has pursued a number of initiatives including a voluntary Code of Practice on Disinformation, which it announced is the:

[176] US Senate, "RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION."
[177] Cannon, "Russia's Employment of Covert Action against Western Democracies from 2013 to Present."
[178] Bennett and Livingston, "The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions."
[179] "What Are the Policy Options Available for Countering Disinformation?," *THE CIPHER BRIEF*, May 2020, https://www.thecipherbrief.com/column_article/what-are-the-policy-options-available-for-countering-disinformation.
[180] "What Are the Policy Options Available for Countering Disinformation?"

> first worldwide self-regulatory set of standards to fight disinformation voluntarily signed by platforms, leading social networks, advertisers and advertising industry in October 2018.[181] Signatories are Facebook, Twitter, Mozilla, Google and associations and members of the advertising industry. Microsoft subscribed to the Code of Practice in May 2019. TikTok joined the code in June 2020.[182]

The European Union (EU) also assembled a Digital Media Observatory designed to provide a central organising structure for fact-checkers, academics, and other stakeholders to engage collaboratively on the challenge of disinformation.[183] Lastly, Canada directed its focus to legislation prohibiting foreign paid advertisements in the pre-election period and banning false statements about candidates during the campaign.[184]

Despite these efforts, An optimal response to addressing this problem will likely require a whole-of-society intra-Western approach that features legislation enforcing the social media platforms' responsibility, public initiatives that lead and support research concerning the evolution of online disinformation, media literacy education of both public and private citizens, and information and intelligence sharing policy that facilitate situational awareness and consequent policy adaptation across the pluralistic societies in the West.[185] Legal limitations could hamper some of these efforts, as free speech is paramount to liberal societies and they do not want to risk descending into the same restricted information space hosted within their adversaries. But policy could rely on recent precedent, such as 2016 EU and US voluntary code compelling social media platforms to combat hate speech.[186] Further efforts could see researchers and the platforms themselves facilitating software developed to warn users of potentially false or misleading speech as a quality control measure. Even the most effective screening mechanisms, however, will not be capable of restricting all disinformation across platforms. This demonstrates the necessity of media literacy promotion. Such efforts should facilitate public-private partnerships based on the Finnish model that found success in educating citizens and public officials on how to assess the quality of online

[181] Media Convergence and Social Media (Unit I.4), "Tackling Online Disinformation," n.d., https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation.
[182] Media Convergence and Social Media (Unit I.4).
[183] Media Convergence and Social Media (Unit I.4).
[184] Tariq Ahmad, "Government Responses to Disinformation on Social Media Platforms: Canada," 2019, https://www.loc.gov/law/help/social-media-disinformation/canada.php.
[185] Sean Monaghan, "Countering Hybrid Warfare" 8, no. 2 (2019): 82–99; Fried and Polyakova, "DEMOCRATIC DEFENSE AGAINST DISINFORMATION."
[186] Ellen P. Goodman and Lyndsey Wajert, "The Honest Ads Act Won't End Social Media Disinformation, but It's a Start," *SSRN Electronic Journal*, vol. 1, 2017, https://doi.org/10.2139/ssrn.3064451.

information and discern false and misleading content from good-faith journalism.[187] These programs were implemented in the adolescent education system and demonstrated clear efficacy in citizens' ability to assess the quality of online media. A bill in the US called the 'Digital Citizenship and Media Literacy Act' was introduced in 2019 to provide funds to K-12 schools to develop media literacy programs.[188] But this bill remains unpassed.

The findings in this paper support the reasoning guiding states' emphasis on critical media literacy. Disinformation is fundamentally concerned with influencing beliefs and thoughts to drive the preferred outcomes of the perpetrator. At present, the public appears increasingly driven to consume its news on social media platforms. Critical thinking curricula specifically designed to empower the next generation to view news content hosted on these platforms with healthy scepticism and a critical eye will help to ensure that the false messaging's targets present a less receptive posture. An obvious shortcoming of these policies is that they are directed towards the school-aged population and will not directly impact the adult population that will presumably continue to access its news on digital platforms. This remains particularly concerning as humans perform just 4 percent better than chance at identifying false information communicated via.[189] The findings in this current report assessed that not only do state-led efforts like the IRA's seek to meddle in elections through tailored disinformation, but they also are informed by users' online cognitive behaviour. The vulnerabilities that were presented in the introduction of this report: media, cognitive, and algorithmic are unlikely to disappear in the near-term. In the interim, it is likely that organisations will continue to endeavour to exploit cognitive processes, such as the confirmation bias, availability heuristic, vividness bias, among others. These efforts target the Wests' thoughts and decisions that will ultimately determine the outcomes of future elections and continued viability of their democracies. Securing the West against such threats therefore will likely require a collective effort to implement measures that specifically emphasise the cognitive element of such campaigns to support the public's mental independence from foreign interference.

---

[187] Sean (Ed.) Monaghan, Patrick Cullen, and Njord Wegge, "MCDC Countering Hybrid Warfare Project: Countering Hybrid Warfare," no. March (2019): 92, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/784299/concepts_mcdc_countering_hybrid_warfare.pdf.

[188] Senator Amy Klobuchar, "S.2240 - Digital Citizenship and Media Literacy Act," Pub. L. No. S.2240 (2019), https://www.congress.gov/bill/116th-congress/senate-bill/2240/text.

[189] Niall J. Conroy, Victoria L. Rubin, and Yimin Chen, "Automatic Deception Detection: Methods for Finding Fake News," *Proceedings of the Association for Information Science and Technology* 52, no. 1 (2015): 1–4, https://doi.org/10.1002/pra2.2015.145052010082.

In light of this concern, a compelling proposal was published by the RAND Corporation's Rand Waltzman in testimony before the US Senate Armed Services Committee. In his statement, he cited a new lens through which to view disinformation known as 'Cognitive Security.'

> Cognitive security (COGSEC) is a new field that focuses on this evolving frontier, suggesting that in the future, researchers, governments, social platforms, and private actors will be engaged in a continual arms race to influence—and protect from influence—large groups of users online. Although COGSEC emerges from social engineering and discussions of social deception in the computer security space, it differs in a number of important respects. First, whereas the focus in computer security is on the influence of a few individuals, COGSEC focuses on the exploitation of cognitive biases in large public groups. Second, while computer security focuses on deception as a means of compromising computer systems, COGSEC focuses on social influence as an end unto itself. Finally, COGSEC emphasises formality and quantitative measurement, as distinct from the more qualitative discussions of social engineering in computer security.[190]

He proposes that states form non-profit, nongovernmental, international centre dedicated to cognitive security. The centre would source contributions from the public sector, academia, think tanks, and private organisations across borders. The centre would convene experts to develop policy and strategies, to create practical technology goals to meet those strategies, devise best practices from contributors from all communities, and conduct independent research driving these functions.[191] This proposal recognises the destructive impact that cognitive attacks can have on societal resilience, particularly with the advent of digital connectivity. Such an organisation could rely on demonstrated means to add rigor and resilience to users' online behaviour, such as the Intelligence Community spawned Structured Analytic Techniques, and employ a deep roster of resources and expertise to develop strategies and tools to meet the ever-metastasizing challenge of information warfare.[192] Waltzman's proposed centre would not be the only means to meet such a challenge. But an effective response would recognise that the continued threat of disinformation is neither uniquely a Russian problem nor an online problem. To adequately address the challenge, the ways by which users engage with a digital sphere that now sits at the centre of media

[190] Waltzman, "Weaponization Inf. Need Cogn. Secur."
[191] Waltzman.
[192] Randolph H. Pherson and Penelope Mort Ranta, "Cognitive Bias, Fake News, and Structured Analytic Techniques" (Globalytica, LLC, 2018).

interpersonal communication and how the world shares and absorbs knowledge must be addressed. Such a strategy will require a collective multinational effort that transcends institutional boundaries to ensure that the future of digital discourse remains compatible with the liberal models that have allowed social media to flourish.

Future Research

As a result of the research highlighted in this paper, this author proposes several avenues for potential future research. This paper found promising results indicating that the IRA may have sought to exploit specific cognitive biases. The research confirmed previous analysis demonstrating that the IRA employed high volume posting, emphasised contentious content, and infiltrated networks on platforms in an effort to pose as organic users. One key additional question that would serve to bridge remaining gaps in the analysis would be to assess how tactics were employed relative to the IRA accounts' positions within networks. This paper hypothesises that IRA accounts may have favoured innocuous but relatable content to their target group before attaining strong network position and then increasing their disinformation output once prominent positions were achieved. Bessi & Ferrara (2016) referred to this phenomenon as 'Bots embeddedness' and measured the condition through a k-core decomposition technique that measured how embedded a bot was within a network.[193] A similar analysis could be performed starting with most prominent accounts, measuring their score over time, and performing a content analysis of tweets to assess if there was any correlation between account position and tweet content.

This paper's analysis of reflexive control doctrine and its persistent role in Russian strategy seems to provide convincing evidence that the Kremlin might have sought to exploit cognitive bias in its 2016 information operation targeted at the presidential election. This paper selected vividness bias, confirmation bias, and the availability heuristic as they appeared to be cognitive processes likely targeted by the effort. However, there remain other biases that also reveal the potential to have been exploited, including cognitive dissonance and anchoring. Cognitive dissonance is 'the psychologically uncomfortable state induced by the co-presence of inconsistent attitudes, beliefs, or behaviours'.[194] Cognitive dissonance can be related to confirmation bias in this context as online networks of users may share common

[193] Bessi and Ferrara, "Social Bots Distort The 2016 U.S. Presidental Election."
[194] Jay J. Van Bavel and Andrea Pereira, "The Partisan Brain: An Identity-Based Model of Political Belief," *Trends in Cognitive Sciences* 22, no. 3 (2018): 213–24, https://doi.org/10.1016/j.tics.2018.01.004.

views, which means they would likely resist similar ideas and information. Based on the IRA's efforts to infiltrate networks and exploit users' common beliefs, it stands to reason that it may have similarly attempted to exploit their cognitive dissonance. The anchoring bias is 'a tendency to make judgments or decisions coloured by previously and often first presented information' (Bouwmweester). In line with this paper's emphasis on the IRA's high-volume output in an effort to exploit the availability heuristic, volume could also have served to attempt to represent the user's first impression of a thought or idea. Other literature has noted that Kremlin-affiliated outlets posting false stories have an agility advantage in that they are not subject to the fact-checking and editorial process that would constrain a traditional news platform.[195] Such an analysis might examine prominent false narratives about real events promoted by the IRA and compare the volume of accounts posting in the early period of a story and then track the output as time progressed. Inevitably as time passes, a story becomes less relevant and would receive lower social media recognition regardless. However, if such an enquiry identified a swarm-like posting in the early stages of a false story, it might suggest that the organisation placed an unusual priority on ensuring that it served as users' first impression concerning the particular instance.

Lastly, further research should be dedicated to Waltzman's proposal to develop a cognitive security centre. Analysis should focus on comparable case studies that might serve to inform the viability of such an organisation.[196] It could also identify practices that may show promise for broader application. Much has been written and developed regarding tools and education for individuals seeking to secure themselves against fake news. But envisioning a solution as a geopolitical issue requiring a mass collective effort to address this threat would prove essential to developing effective policy designed to protect states from the modern threat of adversarial cognitive interference. Even the most effective program would need to consider the research that suggests that users are 'cognitively lazy' when it comes to their online behaviour and solutions would therefore need to account for the risk of a low compliance rate.[197] This concern supports Waltzman's view that such an institution would need to include experts, technology, and tools across public and private divisions, across industry, and disciplines. Such an intricate problem will require comparable deeply considered nuanced solutions.

---

[195] Paul and Matthews, "Russ. 'Firehose Falsehood' Propag. Model Why It Might Work Options to Count. It."
[196] Waltzman, "Weaponization Inf. Need Cogn. Secur."
[197] Pennycook and Rand, "Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning."

References

Ahmad, Tariq. "Government Responses to Disinformation on Social Media Platforms: Canada," 2019. https://www.loc.gov/law/help/social-media-disinformation/canada.php.

Arif, Ahmer, Leo G. 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). https://doi.org/10.1145/3274289.

Badawy, Adam, Aseel Addawood, Kristina Lerman, and Emilio Ferrara. "Characterizing the 2016 Russian IRA Influence Campaign." *ArXiv*, 2019, 1–12. https://doi.org/10.1007/s13278-019-0578-6.

Bail, Christopher A., Brian Guay, Emily Maloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017." *Proceedings of the National Academy of Sciences*, November 25, 2019. https://doi.org/10.1073/pnas.1906420116.

Baines, Paul, and Nigel Jones. "Influence and Interference in Foreign Elections." *RUSI Journal* 163, no. 1 (2018): 12–19. https://doi.org/10.1080/03071847.2018.1446723.

Bavel, Jay J. Van, and Andrea Pereira. "The Partisan Brain: An Identity-Based Model of Political Belief." *Trends in Cognitive Sciences* 22, no. 3 (2018): 213–24. https://doi.org/10.1016/j.tics.2018.01.004.

Bennett, W. Lance, and Steven Livingston. "The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions." *European Journal of Communication* 33, no. 2 (2018): 122–39. https://doi.org/10.1177/0267323118760317.

Bessi, Alessandro, and Emilio Ferrara. "Social Bots Distort The 2016 U.S. Presidental Election." *First Monday* 21, no. 11 (2016): 1–15.

Bouwmeester, Han. "Lo and Behold: Let the Truth Be Told—Russian Deception Warfare in Crimea and Ukraine and the Return of 'Maskirovka' and 'Reflexive Control Theory,'" 125–53. T.M.C. Asser Press, The Hague, 2017. https://doi.org/10.1007/978-94-6265-189-0_8.

Brinkel, Theo. *Netherlands Annual Review of Military Studies 2017: Winning Without Killing:The Strategic and Operational Utility of Non-Kinetic Capabilities in Crises*, 2017. http://ebookcentral.proquest.com/lib/mindef/detail.action?docID=4915496.

Cannon, Casey. "Russia's Employment of Covert Action against Western Democracies from 2013 to Present." Glasgow, UK, 2018.

"Cheltenham Lib Dem Chair Quits after Sex Worker Debate." *BBC*, September 19, 2016. https://www.bbc.com/news/uk-england-gloucestershire-37404894.

Cheney, Kyle. "No, Clinton Didn't Start the Birther Thing. This Guy Did." *POLITICO*, September 16, 2016. https://www.politico.com/story/2016/09/birther-movement-founder-trump-clinton-228304.

Conroy, Niall J., Victoria L. Rubin, and Yimin Chen. "Automatic Deception Detection: Methods for Finding Fake News." *Proceedings of the Association for Information Science and Technology* 52, no. 1 (2015): 1–4. https://doi.org/10.1002/pra2.2015.145052010082.

Cormac, Rory, and Richard J. Aldrich. "Grey Is the New Black: Covert Action and Implausible Deniability." *International Affairs*, 2018. https://doi.org/10.1093/ia/iiy067.

D'Souza, Deborah. "Top 10 Contributors to the Clinton Campaign." In *Investopedia*, 2019. https://www.investopedia.com/articles/investing/033116/top-10-corporate-contributors-clinton-campaign.asp.

Davis, Fred D., René Riedl, Jan vom Brocke, Pierre Majorique Léger, and Adriane B. Randolph. "Information Systems and Neuroscience." *Gmunden Retreat on NeuroIS 2016*, 2017. https://doi.org/10.1007/978-3-319-41402-7.

DiResta, Renee, Dr Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. "The Tactics & Tropes of the Internet Research Agency, New Knowledge," n.d. https://int.nyt.com/data/documenthelper/533-read-report-internet-research-agency/787lea6d5bafbf19/optimized/full.pdf#page=1.

Dobrowolski, Alexandra. "'Weapons of Mass Distraction'?" Citizenship as a Regime," n.d. https://doi.org/10.2307/j.ctv2n7qxg.13.

Flore, Massimo, Alexandra Balahur-Dobrescu, Aldo Podavini, and Marco Verile. "Understanding Citizens' Vulnerabilities to Disinformation and Data-Driven Propaganda," 2019. https://doi.org/10.2760/919835 10.2760/153543.

Fried, Daniel, and Alina Polyakova. "DEMOCRATIC DEFENSE AGAINST DISINFORMATION." Washington, D.C., 2018. https://www.atlanticcouncil.org/wp-content/uploads/2018/03/Democratic_Defense_Against_Disinformation_FINAL.pdf.

Gadde, Vijaya, and Yoel Roth. "Enabling Further Research of Information Operations on Twitter." Twitter.com, 2018. https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html.

Gigerenzer, Gerd, and Wolfgang Gaissmaier. “Decision Making: Nonrational Theories.” *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* 5 (2015): 3304–3309. https://doi.org/10.1016/B978-0-08-097086-8.26017-0.

Giles, Keir. “Handbook of Russian Information Warfare.” *NATO Defence College* 9, no. November (2016): 1–90.

Goodman, Ellen P., and Lyndsey Wajert. “The Honest Ads Act Won’t End Social Media Disinformation, but It’s a Start.” *SSRN Electronic Journal*. Vol. 1, 2017. https://doi.org/10.2139/ssrn.3064451.

Guarda, Rebeka F., Marcia P. Ohlson, and Anderson V. Romanini. “Disinformation, Dystopia and Post-Reality in Social Media: A Semiotic-Cognitive Perspective.” *Education for Information*, 2018. https://doi.org/10.3233/EFI-180209.

Haselton, Martie G., Daniel Nettle, and Paul W. Andrews. “The Evolution of Cognitive Bias.” *The Handbook of Evolutionary Psychology*, 2015, 724–46. https://doi.org/10.1002/9780470939376.ch25.

Hauser, Christine. “Donald Trump Jr. Compares Syrian Refugees to Skittles That ‘Would Kill You.’” *The New York Times*, September 20, 2016. https://www.nytimes.com/2016/09/21/us/politics/donald-trump-jr-faces-backlash-after-comparing-syrian-refugees-to-skittles-that-can-kill.html.

Howard, Philip N, John Kelly, and Camille François. “The IRA, Social Media and Political Polarization in the United States , 2012-2018.” *Computational Propaganda Research Project*, 2018.

Jaitner, ML. “Applying Principles of Reflexive Control in Information and Cyber Operations.” Journal of Information Warfare, 2016. https://doi.org/10.1002/9781118381533.

Kavanagh, Jennifer, Samantha Cherney, Hilary Reininger, and Norah Griffin. “FIGHTING DISINFORMATION Building the Database of Web Tools.” *RAND*, n.d. https://www.rand.org/content/dam/rand/pubs/research_reports/RR3000/RR3000/RAND_RR3000.pdf.

King, Gary, Jennifer Pan, and Margaret E. Roberts. “Reverse-Engineering Censorship in China: Randomized Experimentation and Participant Observation.” *Science* 345, no. 6199 (2014). https://doi.org/10.1126/science.1251722.

Klobuchar, Senator Amy. S.2240 - Digital Citizenship and Media Literacy Act, Pub. L. No. S.2240 (2019). https://www.congress.gov/bill/116th-congress/senate-bill/2240/text.

Kriel, Charles, and Alexa Pavliuc. “REVERSE ENGINEERING RUSSIAN INTERNET RESEARCH AGENCY TACTICS THROUGH NETWORK ANALYSIS” 6 (2019). https://doi.org/10.30966/2018.RIGA.6.

Linvill, Darren L., Brandon C. Boatwright, Will J. Grant, and Patrick L. Warren. “‘THE RUSSIANS ARE HACKING MY BRAIN!’ Investigating Russia’s Internet Research Agency Twitter Tactics during the 2016 United States Presidential Campaign.” *Computers in Human Behavior* 99 (2019): 292–300. https://doi.org/10.1016/j.chb.2019.05.027.

Linvill, Darren L, and Patrick L Warren. “Troll Factories : Manufacturing Specialized Disinformation on Twitter Troll Factories : Manufacturing Specialized Disinformation On.” *Political Communication*, 2020, 1–21. https://doi.org/10.1080/10584609.2020.1718257.

Matney, Robert, Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Dr Kris Shaffer, et al. “The Tactics & Tropes of the Internet Research Agency.” *New Knowledge*, 2019. https://int.nyt.com/data/documenthelper/533-read-report-internet-research-agency/787lea6d5bafbf19/optimized/full.pdf#page=1.

Media Convergence and Social Media (Unit I.4). “Tackling Online Disinformation,” n.d. https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation.

Mikkola, Harri. “The Geostrategic Arctic: Hard Security in the High North.” *Finnish Institute OfInternational Affairs*, no. April (2019). https://www.fiia.fi/wp-content/uploads/2019/04/bp259_geostrategic_arctic.pdf.

Monaghan, Sean. “Countering Hybrid Warfare” 8, no. 2 (2019): 82–99.

Monaghan, Sean (Ed.), Patrick Cullen, and Njord Wegge. “MCDC Countering Hybrid Warfare Project: Countering Hybrid Warfare,” no. March (2019): 92. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/784299/concepts_mcdc_countering_hybrid_warfare.pdf.

Mueller, Robert S. “Report On The Investigation Into Russian Interference In The 2016 Presidential Election.” Vol. I and II, 2019.

Nemr, Christina, and William Gangware. “‘Weapons of Mass Distraction’?” *Park Advisors*, 2019. https://doi.org/10.2307/j.ctv2n7qxg.13.

Nisbet, Erik C., and Olga Kamenchuk. “The Psychology of State-Sponsored Disinformation Campaigns and Implications for Public Diplomacy.” *The Hague Journal of Diplomacy* 14, no. 1–2 (2019): 65–82. https://doi.org/10.1163/1871191X-11411019.

Paul, Christopher, and Miriam Matthews. “The Russian ‘Firehose of Falsehood’ Propaganda Model: Why It Might Work and Options to Counter It.” *The Russian “Firehose of Falsehood” Propaganda Model: Why It Might Work and Options to Counter It*, 2017. https://doi.org/10.7249/pe198.

Pennycook, Gordon, and David G. Rand. “Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning.” *Cognition* 188 (July 1, 2019): 39–50. https://doi.org/10.1016/j.cognition.2018.06.011.

Pherson, Randolph H., and Penelope Mort Ranta. “Cognitive Bias, Fake News, and Structured Analytic Techniques.” Globalytica, LLC, 2018.

Pocheptsov, Georgii. “COGNITIVE ATTACKS IN RUSSIAN HYBRID WARFARE.” *Information & Security: An International Journal* 41 (2018): 37–43. https://doi.org/10.11610/isij.4103.

Pomerantsev, Peter, and Michael Weiss. “The Menace of Unreality: How the Kremlin Weaponizes Information, Culture and Money A Special Report Presented by The Interpreter, a Project of the Institute of Modern Russia,” 2014.

Quattrociocchi, Walter, Antonio Scala, and Cass R. Sunstein. “Echo Chambers on Facebook,” no. 2016 (n.d.). https://doi.org/10.1145/2740908.2745939.

Ray, Rashawn, Melissa Brown, Neil Fraistat, and Edward Summers. “Ferguson and the Death of Michael Brown on Twitter: #BlackLivesMatter, #TCOT, and the Evolution of Collective Identities.” *Ethnic and Racial Studies* 40, no. 11 (2017): 1797–1813. https://doi.org/10.1080/01419870.2017.1335422.

Ruck, By Damian. “Russian Twitter Propaganda Predicted 2016 US Election Polls.” *The Conversation*, 2019, 1–8.

Ruck, Damian J, Natalie M Rice, Joshua Borycz, and R Alexander Bentley. “Internet Research Agency Twitter Activity Predicted 2016 U.S. Election Polls.” *First Monday* 24, no. 7 (2019). https://doi.org/10.5210/fm.v24i7.10107.

Sproul, Spencer. “Fake News ! Russian Disinformation Targets American Cognitive Biases Through Diverse Mediums Fake News ! Russian Disinformation Targets American Cognitive Biases Through Diverse Mediums,” 2019.

Steiglitz, Stefan, Florian Brachten, Davian Berthele, and Mira Schlaus. “Do Social Bots (Still) Act Different to Humans? – Comparing Metrics of Social Bots with Those of Humans.” *Social Computing and Social Media. Human Behavior* 10282 (2017): 379–395. https://doi.org/10.1007/978-3-319-58559-8.

Stewart, Leo G, Ahmer Arif, and Kate Starbird. “Examining Trolls and Polarization with a Retweet Network.” *Proceedings of WSDM Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*, 2018, 6. https://doi.org/https://doi.org/10.475/123_4.

Stuttaford, Andrew. “Information War and Other Deceptions.” *National Review*, April 12, 2015. https://www.nationalreview.com/corner/information-war-and-other-deceptions/.

Sukhankin, Sergey. “Covid-19 As a Tool of Information Confrontation: Russia’s Approach.” *The School of Public Policy Publications* 13, no. 3 (2020): 1–10.

Thomas, Timothy L. “Russia’s Reflexive Control Theory and the Military.” *International Journal of Phytoremediation* 17, no. 2 (2004): 237–56. https://doi.org/10.1080/13518040490450529.

Till, Christopher. “Propaganda through ‘Reflexive Control’ and the Mediated Construction of Reality.” *New Media and Society*, 2020, 1–17. https://doi.org/10.1177/1461444820902446.

Todd, Rebecca M., Deborah Talmi, Taylor W. Schmitz, Josh Susskind, and Adam K. Anderson. “Psychophysical and Neural Evidence for Emotion-Enhanced Perceptual Vividness.” *Journal of Neuroscience* 32, no. 33 (2012): 11201–12. https://doi.org/10.1523/JNEUROSCI.0155-12.2012.

Tornberg, Petter. “Echo Chambers and Viral Misinformation : Modeling Fake News as Complex Contagion,” 2018, 1–21.

Turner, Marlene E., and Anthony R. Pratkanis. “A Social Identity Maintenance Model of Groupthink.” *Organizational Behavior and Human Decision Processes* 73, no. 2–3 (1998): 210–35. https://doi.org/10.1006/obhd.1998.2757.

UNITED STATES COMMISSION ON SECURITY AND COOPERATION IN EUROPE. “THE SCOURGE OF RUSSIAN DISINFORMATION,” 2017. https://www.technologyreview.com/s/604084/russian-disinformation-technology/.

US Senate. “RUSSIAN ACTIVE MEASURES CAMPAIGNS AND INTERFERENCE IN THE 2016 U.S. ELECTION.” 2019. https://www.warner.senate.gov/public/_cache/files/0/d/0dc0e6fe-4d52-49b0-9e92-a15224a74a29/C2ABC2CD38BA3C5207D7FA5352D53EC2.report-volume2.pdf.

Waltzman, Rand. “The Weaponization of Information: The Need for Cognitive Security.” *The Weaponization of Information: The Need for Cognitive Security*. 2017. https://doi.org/10.7249/ct473.

Wang, Zhen, Marko Jusup, Lei Shi, Joung Hun Lee, Yoh Iwasa, and Stefano Boccaletti. "Exploiting a Cognitive Bias Promotes Cooperation in Social Dilemma Experiments." *Nature Communications* 9, no. 1 (2018): 1–7. https://doi.org/10.1038/s41467-018-05259-5.

"What Are the Policy Options Available for Countering Disinformation?" *THE CIPHER BRIEF*, May 2020. https://www.thecipherbrief.com/column_article/what-are-the-policy-options-available-for-countering-disinformation.

Xia, Yiping, Josephine Lukito, Yini Zhang, Chris Wells, Sang Jung Kim, and Chau Tong. "Disinformation, Performed: Self-Presentation of a Russian IRA Account on Twitter." *Information Communication and Society* 22, no. 11 (2019): 1646–64. https://doi.org/10.1080/1369118X.2019.1621921.

Zannettou, Savvas, Michael Sirivianos, Tristan Caulfield, Gianluca Stringhini, Emiliano De Cristofaro, and Jeremy Blackburn. "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web." *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, 2019, 218–26. https://doi.org/10.1145/3308560.3316495.