Russia, Weaponized Social Media and Cyber Warfare: Ethical and Anticipated Ethical Issues

Richard Wilson[1,2] and Noah Donnelly[2]
[1] Department of Philosophy, Towson University, Baltimore, Maryland, USA
[2] Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA

wilson@towson.edu
ndonnel1@students.towson.edu

Abstract: Russia's weaponization of social media refers to how Russian state-associated actors used and continue to use social media platforms such as Facebook, Twitter, Instagram, and YouTube to influence public opinion, spread disinformation, and destabilize societies, especially during elections and political crises. This strategy relies on multiple essential tactical components. The first tactical component is Disinformation Campaigns, where Russian operatives created fake accounts and groups posing as ordinary citizens to spread falsified stories on politically divisive topics (race, immigration, gun rights) to drive up political polarization. Secondly are Troll Farms like the Internet Research Agency (IRA), which employed thousands of people to post inflammatory content with the aim of manipulating online discourse. The third strategy is Bot Networks, utilizing automated accounts to amplify hashtags and create the illusion of widespread support for fringe narratives. The fourth strategy is Microtargeting, where malicious actors purchased ads exploiting emotional triggers and psychographic data to influence voter behavior. Finally, these campaigns Exploited Algorithms that prioritize engagement, making sure sensational content became viral. The goal of this weaponization is to destabilize trust in democratic institutions, amplify social division, and influence election outcomes to favor Russia's geopolitical interests. This analysis identifies the ethical and anticipated ethical issues with the Russian weaponization of social media as a form of Cyber Warfare. There is an interdisciplinary method employed in this analysis that draws upon distinctions taken from computer science, conceptual ethical analysis and case studies.

Keywords: Cyber warfare, Information operations, Disinformation, Internet research agency, Cognitive warfare, Anticipatory ethics

1. Introduction

In the contemporary realm of global conflict, the battlefield has expanded beyond the standard domains of land, sea, air and space to include the cyber domain and audience cognition. Whilst cyber warfare has historically focused on attacks on physical infrastructure or the theft of classified data, a new paradigm has emerged: the weaponization of social media to target psychological and sociological infrastructures of a nation. These phenomenon, best displayed by the operations of Russian state-sponsored actors, represent a massive shift in the nature of warfare. They move the objective of warfare from the destruction of physical assets to the disruption of the "OODA Loop"(Observe, Orient, Decide, Act) of an entire population (Claverie & du Cluzel, 2022). By invading the infosphere of a target nation, an adversary can manipulate public perception, erode trust in institutions, and incite civil unrest. This strategy is known as Cognitive Warfare, and it utilizes the architecture of the modern internet as a medium for the use of their weapons systems to produce mass influence (NATO ACT, 2021).

The significance of the weaponization of social media is highlighted by the combination of sophisticated psychological operations with advanced technological capabilities. In the past, propaganda was limited by the speed of broadcasted messages through legacy media and the ability of states to control information channels. Today the internet allows for micro-targeting of individuals based on psychological profiles which allows state actors to bypass the traditional media gatekeeping, delivering tailor fit news direct to users (Zuboff, 2019). This capability was demonstrated during the 2016 United States Presidential Election, where the Internet Research Agency (IRA), a Russian "troll farm," executed an extensive information operation designed to sow discord among citizens and influence the election (Almond et al., 2022). The 2016 election offers a foundational case study for these tactics. Russian strategy has since evolved significantly. The 2022-2025 invasion of Ukraine marked a shift from "covert disruption" to "overt information warfare." This newer approach uses deepfakes, AI-generated propaganda, and the weaponization of platforms such as Telegram and TikTok. This paper analyzes this evolution. It contrasts the manual "troll farms" of the past with the automated "cognitive warfare" of the present.

The clear distinction between "influence" and "warfare" in this context grows increasingly blurry. Traditional defenses of war require a kinetic act or a direct violation of a nation's sovereignty. However, when a foreign state decides to influence the public discourse of another nation with the intent to destabilize its institutions

and provoke internal conflict, it meets the strategic criteria of warfare (Dawson & Innes, 2019). The intent of social media warfare is to cause irreversible harm to the targeted state's ability to function. The "cyber weapons" being deployed are not boots on the ground but code, bots, and algorithms**.** The impact is the degradation of sovereignty and the ability to autonomously rule one's people, along with weakening the targeted nation on the geopolitical stage. This analysis investigates specific tactics employed by Russian state actors, including disinformation campaigns, troll farms, bot networks, and the exploitation of algorithms to accomplish their goals. It analyzes these actions through the lens of standard ethical principles, Anticipatory Ethics and the ACM Code of Ethics. It argues that the designers of social media platforms have failed to uphold the Sociotechnical Imperative by ignoring the possibility of the weaponization of their tools (Miller et al., 2011). Furthermore, it proposes that the defense against this form of warfare requires a move from reactive moderation to a proactively oriented cognitive cybersecurity.

2. Technical Issues

To understand how social media platforms have developed into weapons of cyber warfare, one needs to understand the technical architecture that facilitates this weaponization. The core technical issues lay in the design of the "Attention Economy" and the specific algorithms that govern user interaction on platforms including but not limited to: Facebook (now Meta), Twitter (now X), and YouTube. These platforms are not designed with civic integrity or national security in mind. Social media platforms are designed to maximize user engagement and length of engagement to maximize advertisement revenue (Zuboff, 2019). This commercial objective creates a set of technical vulnerabilities that state actors can exploit with devastating effects.

The primary mechanism of exploitation is the Recommendation Algorithm. These algorithms employ machine learning models, specifically Reinforcement Learning, to predict what keeps a user scrolling or clicking. Success is measured by "engagement," defined as likes, shares, comments, and engagement time. Research in behavioral psychology continuously shows that content on social media platforms which elicit high-arousal emotions, including fear, anger, and outrage tend to generate significantly higher user engagement than neutral or positive content (Bipartisan Policy Center, 2023). Consequently, the algorithms of profit driven social media platforms prioritize divisive, sensational, and inflammatory content. Russian operatives did not need to hack any code to spread their message, they simply needed to create content that triggered these algorithmic preferences in targeted audiences. By posting polarizing material, they ensured the platforms' automated systems proliferated their propaganda to millions of users with minimal economic expenditure. This phenomenon is known as "Algorithmic Radicalization," where the system pushes users towards increasingly extreme content to maintain engagement (National Institute of Justice, 2021).

A second crucial technical component of social media warfare is the capacity for Microtargeting and Psychometric Profiling. The business model of social media is built on "Surveillance Capitalism," where user data is harvested and used to create comprehensive profiles pertaining to individual behavior, preferences, and psychological traits (Zuboff, 2019). Advertisers then use this data to target specific demographics. Russian actors weaponized social media commercial tools by purchasing political advertisements targeting users who share specific vulnerabilities (Young & Kim, 2018). An example of this process would be identifying users who score high on "neuroticism" or "authoritarianism" in psychometric models and deploy ads to trigger aggression and paranoia in these audiences. This represents the industrialization of psychological warfare. In classic PSYOPS, leaflets are dropped on a population with a single, generalized message. In contrast, in social media warfare, a specific message is made for and directed towards specific individuals based on their user data which maximizes psychological impact while remaining unrecognized by a broader audience. This "Dark Post" (Singer & Brooking, 2018) architecture allows for reality to be fragmented where different sectors of the population are presented entirely unique narratives, making a general consensus on important topics nearly impossible to achieve in the overall population.

The third technical pillar is the deployment of Automated Bot Networks. A "bot" is a software application that runs automated tasks over the internet. In the context of cyber warfare, bots are employed to simulate human activity on social media. Russian actors deploy armies of bots to artificially boost specific narratives (Dawson & Innes, 2019). By programming thousands of accounts to "retweet" and "share" a specific piece of disinformation, they can manipulate the "Trending Topics" algorithms like those seen on Twitter. This creates a "Bandwagon Effect," where real human users see a trending topic and assume it represents popular opinion on breaking news (Singer & Brooking, 2018). This technique is known as "Astroturfing." Astroturfing creates an artificial illusion of widespread grassroots support for a fringe or completely fabricated idea. The technical failure here lies in the

inability or unwillingness of social media platforms to effectively distinguish between human and "bot" traffic. This creates an environment where automated narratives produced by Bots can drown out authentic discourse.

3. Ethical Issues

The weaponization of social media in the service of cyber warfare raises profound ethical questions that go beyond the technical details of the attacks. These ethical issues concern the inalienable rights of citizens, the moral obligations of corporations, and sovereignty in the digital age. This analysis applies ethical frameworks to the three primary stakeholder groups: the State Actors (Russia), the Platforms (Meta/Twitter), and the Citizens targeted through social media (victims).

From the perspective of International Relations and Political Ethics, the actions of Russian state actors violate the Principle of Non-Intervention and National Sovereignty. Sovereignty is the right of a nation to govern itself without external influence. The democratic process, specifically the election of a nation's leadership, is the most important expression of sovereignty. By covertly interfering in the 2016 U.S. election through information warfare, Russian actors violated the right of the American people to self-determination (Dawson & Innes, 2019). Rights Ethics defends the view that state citizens have a "Right to Mental Integrity" and a "Right to the Truth." Disinformation campaigns that deliberately employ social media to corrupt the information environment violate the epistemic rights of civilians (Tavani, 2009). Disinformation acts like a "virus" that exploits cognitive vulnerabilities, and which then deflects citizens from access to the accurate information required to make rational and informed political decisions. When state actors spread false news stories, such as claiming that a candidate is involved in criminal activity when thy aren't, they are not exercising free speech, they are engaging in a deception operation designed to subvert the autonomy of voters. This represents a violation of the Kantian duty related to honesty. Kantian Deontology argues that lying is inherently immoral because it treats the listener not as a rational agent, but a means to an end (Brey, 2012). Russian disinformation campaigns treated American voters as objects to be manipulated for geopolitical gain, which is a violation of their human dignity.

The ethical responsibility of the social media platforms is equally significant when evaluated through the lens of Virtue Ethics and the ACM Code of Ethics. Companies like Facebook (Meta) and Twitter (X) regularly claim to be neutral "town squares" that merely host content. This defense ignores their active role in curating and amplifying the content appearing on them using algorithms. The ACM Code of Ethics Principle 1.2 declares that computing professionals must "Avoid Harm" (Association for Computing Machinery, 2018). By designing algorithms that value engagement over veracity, and by allowing their platforms to be used for mass manipulation, these companies failed to avoid harm. The harm in this context is not only distress caused to individuals who were harassed by trolls, but systematic damage to the democratic ideal. The degradation of trust in institutions, increasing polarization, and incitements of violence are foreseeable harms that are enabled by the platforms' negligence of content moderation (Singer & Brooking, 2018).

Furthermore, the platforms violated ACM Principle 1.3, which necessitates the duty to "Be Honest and Trustworthy" (Association for Computing Machinery, 2018). For years, platforms have denied or minimized the extent of foreign interference being conducted on their platforms. They allowed bots to pose as humans, deceiving their users about the nature of interactions they have with Bots and the content spread by those Bots (Dawson & Innes, 2019). Virtue Ethics would ask, "what is the character traits of the designers of a social media system?" A system that covertly sells user vulnerabilities to foreign actors and prioritizes profit over the integrity of the democratic process lacks civic virtue. The concept of "Corporate Social Responsibility" (CSR) states that companies have an obligation to the societies within which they operate. By profiting from the capital spent on ads by Russian disinformation agents, these platforms directly profit from the destabilization of the nation's that host social media platforms (Quigley, 2017). This represents a failure of the social contract between the technological sector and the members of the public using social media platforms. These companies are placing shareholder interests, values and profit above the preservation of democratic principles.

4. Case Studies

The complex nature of "cognitive cyberwarfare" is best understood through the tactical implementations deployed during the 2016 U.S. election and subsequent geopolitical events. A series of case studies demonstrate the methodology under which the Internet Research Agency (IRA) and Russian Intelligence services (GRU) operate.

4.1 Disinformation Campaigns

The first prong of the Russian strategy was the industrial-scale production of disinformation. Russian operators made thousands of fake accounts, pages, and groups posing as American citizens or activist organizations. Groups with names like "Blacktivist," "Heart of Texas," and "United Muslims of America" were designed with the intention of infiltrating genuine political movements (Young & Kim, 2018). Once these pages had gathered authentic users, the operators would deploy "payload" content, that were fabricated or highly misleading stories designed to produce outrage. For example, during the 2016 election, Russian controlled pages spread false reports about voter fraud, criminal investigations into candidates, and fabricated quotes (Almond et al., 2022). One memorable operation involved Russian operatives creating two opposing protest events in Houston, Texas, on the same day. Another protest supported an Islamic center, while simultaneously protesting against the center. The goal was to increase political polarization by employing digital means related to an event that already involved a physical confrontation. Both events were organized remotely by Russian state actors in St. Petersburg. This tactic mirrored a Cold War strategy called "Active Measures," which was accelerated by the speed of the internet. The strategy relied on "Reflexive Control," (cite source) a concept in Russian military theory where the aim is to compel the opponent to make decisions. Decisions about how to vote or whether to engage in a protest are influenced by false premises and disinformation introduced by the attacker (Claverie & du Cluzel, 2022).

4.2 Troll Farms

The organizational heart of this warfare is the Internet Research Agency (IRA), a shadowy corporation based in St. Petersburg. The IRA operates as a "Troll Farm," employing over a thousand people to work 12-hour shifts creating content for social media. The facility was organized in a fashion similar to a professional marketing agency. The IRA had a graphics department, data analysis teams, and IT support (Dawson & Innes, 2019). The "trolls" job was to manage multiple fake personas across different time zones to simulate authentic American users. Their tactic was "comment warfare," where they would swarm the comment sections of news articles and Facebook posts to inject divisive rhetoric. They would regularly play both sides of an argument, having each side posting extremist viewpoints from both the left and the right to eliminate any middle ground, making accurate comprehension of messages impossible. This is essentially a "divide and conquer" strategy adapted to fit the digital age. The immense volume of content produced by the IRA created a "Firehose of Falsehood," which is a propaganda technique where the audience is overwhelmed by the high frequency of conflicting messages which can cause the audience to become cynical and disengage from politics entirely (Singer & Brooking, 2018). This achieves two strategic goals, 1st, the weaponizing of social media and voter suppression and the 2nd contributing to the erosion of faith in democratic discourse.

4.3 Bot Networks

While human trolls created messages, automated bots would provide the means for the distribution and amplification of these messages. Russian actors used networks of tens of thousands of automated accounts to artificially increase the visibility of their messages. When a piece of disinformation was posted, the botnet would instantly "retweet" and "share" the disinformation thousands of times within a matter of minutes (Dawson & Innes, 2019). The performance of these acts exploited the purpose of "Trending" algorithms on platforms like Twitter, which interpret rapid engagement as a sign of breaking news. By forcing a specific hashtag or narrative to trend, the botnet can bypass the "filter bubble" of individual users and push false narratives into the mainstream of the media cycle (National Institute of Justice, 2021). Journalists and politicians, who now, seeing a topic trending, would often comment on the topic which unknowingly lent credibility to a narrative that was created by a foreign intelligence service. This technique of "Artificial Amplification" through the use of bots, influences the heuristic processes of the human mind. Humans are evolutionarily programmed to pay attention to what the crowd is discussing. Bot networks emulate crowd sourced ideas which influence the thoughts of individuals.

4.4 Microtargeting and Ads

The final tactical layer of tactics employed by the IRA and GRU was the use of commercial advertising as tools for information warfare. Russian actors purchased over 3,000 political advertisements on Facebook and Instagram, spending roughly $100,000, which is a trivial sum to a state actor, but made highly effective due to microtargeting of individual audiences (Young & Kim, 2018). Unlike television ads, social media ads are not broadcast to a general audience but can be "microtargeted" to specific subsets of a population. Russian operatives used Facebook's ad tools to target users based on "Likes," location, and demographic (Zuboff, 2019). They crafted specific messages for African American voters intended to discourage turnout, while

simultaneously targeting conservatives with messages designed to incite fear regarding subjects such as immigration. The use of microtargeting as an application of propaganda exploits the “surveillance capitalism” model. The platforms provided the weapon (targeting tools) and the ammunition (user data) to Russian actors, who then pulled the trigger. This is also a failure of “Know Your Customer” (KYC) protocols in the digital ad market. In this case foreign actors can anonymously purchase influence in a domestic election using foreign currency and accomplish their goals with zero scrutiny (Quigley, 2017).

4.5 Russia Ukraine War

The Russian invasion of Ukraine showed a clear change in information warfare. Operations went from a secret, denial-based method used in 2016 into an open and complete control of information. The Internet Research Agency’s (IRA) early tactics slowly infiltrated natural political movements. The 2022-2025 strategy used a high-volume, high-pressure flow of falsehoods. Artificial intelligence and encrypted platforms boosted this strategy to support military attacks. A key tactical step up involved using synthetic media directly within a combat zone. In March 2022, Russian hackers took over the website of the Ukrainian news broadcaster Ukraine 24. They streamed a deepfake video of President Volodymyr Zelenskyy. In the video a fake Zelenskyy told Ukrainian troops to drop their weapons and surrender to the Russian armed forces. People quickly exposed the video as fake because of visual problems, such as a still body and a voice that did not match the lips. This video acted as a "Proof of Concept" for combining online influence operations with real military aims. It pushed social media weaponization past simply "sowing discord." It aimed to reach actual military goals, specifically getting enemy soldiers to surrender, using digital tricks.

5. Anticipatory Ethics

The failure to prevent the weaponization of social media was not only an unavoidable consequence of technology, it was a failure of foresight about social media vulnerability. Anticipatory Ethics is a framework which focuses on the development of potential ethical issues relating to new and emerging technologies before the technologies become fully entrenched (Brey, 2012). Applying the "Moral Responsibility for Computing Artifacts" rules by Miller et al. reveals the extent of the failures of social media platforms to foresee ethical issues due to the lack of content moderation.

5.1 Rule 4: The Sociotechnical Imperative

The Sociotechnical Imperative states that “People who design, develop, or deploy a computing artifact can do so responsibly only when they make a reasonable effort to take into account the sociotechnical systems in which the artifact is embedded” (Miller et al., 2011). Social media developers designed their tools for a sterile, idealized world where users are all believed to be authentic and who wish to make connections with others in an inherently positive way. The developers have entirely failed to take into account the sociotechnical reality of geopolitics, where bad actors exist and where information can be weaponized by these bad actors. They ignored the “Dark Scenarios” described by Philip Brey, failing to model how a state adversary can and will abuse their systems (Brey, 2012). By developing their algorithms solely for optimization of engagement, they basically built a machine that amplifies disinformation, hateful speech, and bad conduct. A responsible design process would have anticipated the vulnerabilities that could be exploited by bad state actors and they could have built “friction” into the system to slow the spread of unverified information. The failure to uphold the Sociotechnical Imperative has resulted in platform architecture that is vulnerable to PSYOPS employed in cognitively oriented cyberwarfare.

5.2 Rule 1: The Foreseeability of Effect

Rule 1 for computing artifacts asserts that “The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact” (Miller et al., 2011). The interference in the 2016 election was not an unforeseeable effect of the platforms design. It was known that open ad platforms allowed for anonymous purchasing. It was also known that algorithms prioritizing outrage can lead to radicalization. It was known that bots can mimic humans. The platforms cannot claim ignorance, since the vulnerabilities were features of how the social media platforms work, not anomalies. The adherence to a “move fast and break things” ethos of these social media companies prioritized speed and growth of the platforms over safety. This negligence of safety concerns renders the architects of the social platforms morally responsible for the consequences of the cognitively oriented warfare waged on their networks. They built the battlefield and sold the weapons, no different than the “Merchants of Death” of the late 19th and early 20th centuries, firing no shots themselves but causing irreversible social and political damage.

5.3 Rule 5: Honesty in Promotion

Rule 5 declares that “People who design, develop, or deploy a computing artifact must not deceive the users about the artifact” (Miller et al., 2011). Social media platforms market themselves as tools for “connection,” “community” and “giving people a voice.” This framing is incredibly deceptive. The reality is that the platforms functioned as surveillance engines that sold user attention to the highest bidder. By concealing the mechanics of the algorithms and the extent of data mining, the platforms deprived users of informed consent. Users believed they were engaging in a public square, not knowing that they were arguing against Russian bots and the “news” they were reading was targeted propaganda (National Institute of Justice, 2021). This deception prevented users from properly evaluating the information they consumed, making them vulnerable to psychological manipulation. Honesty in the promotion of their products would require social media platforms to clearly label automated accounts, disclose the source of advertisements, and explain why content is being recommended.

6. Recommendations

To counter the “Velocity,” “Virality,” and “Microtargeting” of modern cognitive cyberwarfare, by relying on human moderation is insufficient. The “OODA Loop” of a disinformation campaigns operates at machine speed, while democratic deliberation runs at the speed of human thought. Therefore, we propose a shift to Cognitive Cybersecurity, a defense architecture designed to secure the information environment through transparency and technical provenance.

The first pillar of this architecture is Cryptographic Content Provenance. Russian disinformation campaigns rely on the ability to pass off fabricated content from fake news sites or doctored images as authentic. To address this, we recommend the universal implementation of the C2PA (Coalition for Content Provenance and Authenticity) standard. This utilizes public key cryptography to digitally sign media at the source of its creation. When a journalist or government entity publishes a story, it is “stamped” with a cryptographic hash. Social media platforms could then display a “Verified Source” badge based off this data. If a Russian operative copies the image and alters anything, the hash breaks, and the platform automatically flags it as “Manipulated Media” (Rosenthal, 2021). This moves the burden of trust from the user’s judgement to a cryptographic infrastructure.

The second pillar involves Algorithmic “Circuit Breakers.” Disinformation exploits the speed of social media algorithms, where outrage spreads faster than fact checks can be produced. To counteract this, platforms must implement viral circuit breakers, a concept borrowed from financial markets where trading is halted during bouts of panic selling. This occurs if a piece of content exhibits “Anomalous Viral Velocity,” such as a sharing speed more than three standard deviation points above the normal, combined with low credibility flags (e.g., a newly created account with zero outbound links). This introduces “Friction” into the system, buying time for an automated fact checking system or human reviewer to assess the content before it reaches critical mass.

The third pillar addresses the issue of automated botnets through Zero-Knowledge Proof of Personhood. The IRA used botnets to create the illusion of a consensus in audiences, called Astroturfing. To neutralize this, we propose Proof of Personhood protocols to verify if a user is human without revealing their identity. Using Zero-Knowledge Proofs (ZKPs), a user can prove they are a unique human via a decentralized identity token without sharing their name or government ID with the platform. Accounts that lack this “Humanity Token” would be rate-limited or frozen, preventing them from mass-retweeting or mass-commenting. This effectively decreases the influence of botnets while preserving the user’s right to privacy (Tavani, 2009). Finally, we recommend an Algorithmic Transparency API to allow vetted researchers to audit recommendation engines, ensuring integrity against foreign agents.

7. Future Work

While this analysis has focused on the tactics of the 2016-2022 era, which were characterized by human controlled troll farms and manual coordination, the future of cognitive cyberwarfare lies in the automation of these operations using Generative Artificial Intelligence (GenAI).

Current Russian doctrine relies on the Internet Research Agency (IRA) employing humans to write the divisive comments. Future research must investigate the rise of "Agentic Disinformation Systems." As noted by Pauwels (2024), we are moving from human-driven troll farms to LLM-driven autonomous agents capable of engaging in sustained, context-aware debates with thousands of users simultaneously.

This shifts the bottleneck of propaganda from "human labor" to "compute power," allowing for a scale of manipulation previously impossible. Human trolls are limited by fatigue and language barriers, LLMs can generate infinite, linguistically sound, and contextually relevant propaganda at zero marginal cost. A single "Agentic AI" could replace an entire floor of human operators, constructing thousands of arguments at a time in the comment sections of their target audiences (Pauwels, 2024).

The next frontier of disinformation is Synthetic Media. As Generative Adversarial Networks (GANs) become accessible, state actors will move beyond mere text to audio and video fabrication. Future studies should analyze the impact of "deepfakes", synthetic audio and video clips released on private messaging apps (such as WhatsApp or Telegram) where attribution is impossible and viral spread is end-to-end encrypted. We anticipate a shift from "Broadcasting" fake news to "Narrowcasting" fake reality, where specific individuals will be targeted with deepfake voicemails and videos from trusted contacts (e.g., boss or family member) designed to elicit immediate behavioral compliance (Helmus, 2022).

Finally, future research must address the threat of Data Poisoning. As western nations grow more reliant on AI analysis for decision making, Russian actors are likely to switch from influencing people to influencing the Large Language Models themselves. By flooding the open web with subtle disinformation, they aim to "poison the well" by poisoning the data on which AI models are trained, in the attempt to ensure Western AI systems adopt a pro-Russian bias, or to spew historical revisionism (Schneier, 2023).

8. Conclusion

The Russian weaponization of social media shows that cognitive cyberwarfare is a present and escalating reality. The tactics of the IRA which employs disinformation, troll farms, microtargeting, and bot networks, are not trivial nuisances. They are state-sponsored psychological operations designed to degrade the sovereignty and autonomy of nation states. This form of warfare exploits the vulnerabilities of the human mind and architecture of the internet to achieve strategic objectives without the need for kinetic attacks

By applying the frameworks of standard ethical principles, the ACM code of ethics and Anticipatory Ethics, we conclude that the primary vulnerability lies not with the people, but with the platforms on which cognitive cyberwarfare takes place. The prioritization of engagement over truth, commodification of user data, and the lack of verification have created a "security hole" in the democratic process that the Russian operatives exploited. The failure of platforms to uphold the Sociotechnical Imperative and the principle of Foreseeability of Effect represent a profound ethical lapse in judgement.

To defend against these threats, we need to move beyond "content moderation" which is a reactive process. We must implement a posture of Cognitive Cybersecurity. This requires a redesign of social media architectures to prioritize authenticity over anonymity, friction over virality, and transparency over opacity. It requires the implementation of defensive AI to detect a coordinated threat by its behavior and the establishment of ethical standards that treat the attention of a user as a protected asset. The defense of the "Human Element" in the digital age is no longer a matter of internet literacy; it is a matter of engineering. It is the moral obligation of the computing profession to build digital infrastructures that are protected from manipulation, which will then protect rather than exploit the cognitive literacy of the citizens that inhabit the cyber sphere.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.0 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing. Gemini 3.0 pro used for aiding in sourcing.

References

Almond, D., Du, X., & Tang, A. (2022). Reduced trolling on Russian holidays and daily U.S. presidential election odds. *PLOS ONE*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0264507

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. https://www.acm.org/code-of-ethics

Atlantic Council. (2022). *Deepfakes in the War in Ukraine: A Threat Assessment*. Digital Forensic Research Lab (DFRLab).

Bipartisan Policy Center. (2023). *The Pros and Cons of Social Media Algorithms*. https://bipartisanpolicy.org/wp-content/uploads/2023/10/BPC_Tech-Algorithm-Tradeoffs_R01.pdf

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, 6(1), 1–13.

Claverie, B., & du Cluzel, F. (2022). Cognitive Warfare: The New Battlefield Exploiting Our Brains. *Polytechnique Insights*. https://www.polytechnique-insights.com/en/columns/geopolitics/cognitive-warfare-the-new-battlefield-exploiting-our-brains/

Collingridge, D. (1980). *The Social Control of Technology*. St. Martin's Press.

Dawson, A., & Innes, M. (2019). How Russia's Internet Research Agency Built its Disinformation Campaign. *The Political Quarterly*, 90(2), 245–256.

Digital Forensic Research Lab (DFRLab). (2024). *Doppelganger: How Russia mimicked real news sites and created fake ones to target US audiences*. Atlantic Council. https://dfrlab.org/2024/09/18/doppelganger-us-election/

Google Threat Intelligence Group. (2025). *Pro-Russia Information Operations Leverage Russian Drone Incursions into Polish Airspace*. Google Cloud Blog. https://cloud.google.com/blog/topics/threat-intelligence

Helmus, T. C. (2022). *Artificial Intelligence, Deepfakes, and Disinformation: A Primer*. RAND Corporation.

Miller, K.W., et al. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. *IT Professional*, 13(3), 57–59.

NATO Allied Command Transformation. (2021). *Cognitive Warfare*. https://www.act.nato.int/activities/cognitive-warfare/

NATO Allied Command Transformation. (2024). *Cognitive Warfare: A Strategic Concept*. NATO ACT. https://www.act.nato.int/activities/cognitive-warfare/

National Institute of Justice. (2021). *The Role of the Internet and Social Media on Radicalization*. https://www.ojp.gov/pdffiles1/nij/305797.pdf

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Pauwels, E. (2024). *The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI*. United Nations University Centre for Policy Research.

Quigley, M. (2017). *Vanity Fair: Did Jared Kushner's Data Operation Help Select Facebook Targets For The Russians?* https://quigley.house.gov/media-center/news-article/vanity-fair-did-jared-kushner-s-data-operation-help-select-facebook

Royal United Services Institute (RUSI). (2025). *Russia, AI, and the Future of Disinformation Warfare*. RUSI. https://static.rusi.org/russia-ai-and-the-future-of-disinformation-warfare.pdf

Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The Weaponization of Social Media*. Eamon Dolan/Houghton Mifflin Harcourt.

Tavani, H.T. (2009). *Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology*. Wiley.

Tech Policy Press. (2025). *Ukraine's Hard-Won Approach to Strategic Communications and Counter-Disinformation*. Tech Policy Press. https://www.techpolicy.press/ukraines-hardwon-approach

Thomas, E. (2024). *Russian influence operation Doppelganger linked to fringe advertising company*. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/digital_dispatches/russian-influence-operation-doppelganger

United States Department of Justice. (2024). *Justice Department Seizes 32 Domains Used in Russian Government-Directed Foreign Malign Influence Campaigns*. Office of Public Affairs. https://www.justice.gov/opa/pr/justice-department-seizes-32-domains-russian-influence

United States Department of State. (2024). *Disarming Disinformation: Russia's Use of Chemical Weapons Narratives in Ukraine*. Global Engagement Center. https://2021-2025.state.gov/disarming-disinformation/

Young, M., & Kim, Y. (2018). *Uncover: Strategies and Tactics of Russian Interference in US Elections*. University of Wisconsin–Madison.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.