Ethical Analysis of Artificial Intelligence-Driven Social Media Propaganda Campaigns During the Russo-Ukrainian War

STS Research Paper
Presented to the Faculty of the
School of Engineering and Applied Science
University of Virginia

By

Brian Xiao

May 9, 2025

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR

Benjamin J. Laugelli, Assistant Professor, Department of Engineering and Society

Introduction

Social media is good. Great even, but the power to connect people can also be used to deceive. Many prominent political entities use social media platforms to spread their messages and ideas to a large scope of people. These same platforms can be abused by bad actors to spread propaganda. One such event was the presence of a pro-Russian propaganda campaign on social media platforms during the Russian invasion of Ukraine in 2022.

Current research on this case primarily focuses on verifying the existence of a Russian state-organized social media propaganda machine and the methods and strategies the organization employs. However, not much attention has been given to the ethical considerations and moral issues surrounding the use of propaganda in the war. Additionally, attention has not been given to the specific application of artificial intelligence (AI) models and techniques to propaganda campaigns and the further ethical considerations that this technology introduces. The incorporation of AI introduces several ethical concerns ranging from the disruption of public discourse with propaganda bots to the abuse of individual reputations.

Understanding the ethical complications of misinformation with generative AI is important for engineers designing models so that they are aware of potential abuse, and also important for those that wish to influence users to operate in a manner that is morally acceptable. This research also holds importance for users themselves. Users need to grasp both the methods that may be used in manipulation, as well as the intention and motivations of the actors behind online manipulation to effectively navigate the current digital landscape, which is plagued with bad actors.

I claim that the pro-Russian and pro-Ukrainian AI-driven social media campaigns are both morally unacceptable because under Kantian ethics they fail to satisfy the universality

principle of the categorical imperative, lack good will, and fail to respect the autonomy and dignity of social media users. For this argument, I will utilize Kantian ethics, a deontological ethical framework, to analyze the morality of the actions taken by the individuals coordinating propaganda efforts. Kantian ethics focuses on following a core set of beliefs that form the categorical imperative. The categorical imperative provides a framework to assess the morality of potential moral maxims as well as providing additional conditions such as good will and the reciprocity principle to judge the morality of one's actions. To support my argument, I will reference several reports published by other parties about the validity of certain claims made in propaganda as well as direct pieces of propaganda itself.

Background

Russia and Ukraine were both a part of the Union of Soviet Socialist Republics. The two countries historically share a deep connection because of their proximity and cultural ties. February 24th, 2022, Russia began a military invasion of Ukraine, which was a continuation of a previous conflict when Russia annexed the Crimean peninsula of Ukraine in 2014. In addition to the military campaign, both sides unleashed coordinated social media campaigns with the purpose of spreading propaganda to shape public perception and control the narrative surrounding the invasion (Hasan, 2024). This online propaganda presented the invasion as a defensive action for the Russian-speaking population in Ukraine and a struggle against Western imperialism.

The range of this social media campaign reached a global scale, allowing misleading partisan narratives to become widespread. Social media is a powerful and accessible platform for reaching an audience on a global scale. Russia has an especially extensive propaganda operation

on social media. There is no doubt that a large-scale media campaign such as Russia's has had a degree of influence on some people. This has been exacerbated by how algorithms on social media platforms often present users with controversial content in an attempt to engage users. Russian government-controlled accounts made claims of "denazification" and that the Ukrainian government was the oppressor towards the local ethnic Russians (Putin, 2022). Through the use of charged language and various other strategies such as posting incendiary messages, the Russian government leveraged digital platforms to villanize Ukraine in the public perception.

The influence and landscape of social media have evolved as generative Artificial Intelligence has become more sophisticated. As AI technology has developed, it has become more and more capable of creating content that resembles human-made content, blurring the lines between human and machine-fabricated activity. The primary impact of generative AI on online social media is the introduction of the Dead Internet Theory, which posits that a majority of the content and activity found online is AI-generated (Walter, 2024). The key idea is that there is an undeniable and increasing presence of AI-generated content online, particularly on social media platforms. This raises a critical concern about the authenticity of online communication as coupled with misinformation, users may especially struggle with differentiating between what is genuine or not.

Literature Review

The role of using bots, or fake accounts, in social media warfare has been widely investigated, especially in the case of propaganda efforts during the Russian invasion of Ukraine. As technology evolves, more advanced methods such as artificial intelligence have been used to produce content as a part of a propaganda campaign. This automation significantly increases the

effective scale of campaigns targeting a global audience on social media platforms. Russia has utilized this to spread false narratives about Ukraine, the United States of America, and the European Union across various public media platforms (Tolmach 2024).

One of the larger and more popular social media platforms is X, formerly known as Twitter. In 2023, Geissler et al. discovered that 20.28% of pro-Russian accounts posting about the conflict were bots compared to 14.25% of Ukrainians. The primary focus of this research was to prove the existence of a coordinated propaganda campaign. In addition to posting content, researchers discovered that both affiliations of bots utilized a strategy centered around retweeting human-generated content with the intention of boosting the content's exposure. Retweeting on Twitter allows users to share someone else's message, giving it increased visibility. The Russian bots were responsible for 25.72% of retweets despite being only 20.28% of accounts. Additionally, bots retweeted content quickly after posting, promoting the early diffusion of messages. This shows how bots are coordinated to act as amplifiers of propaganda and maximize the visibility of the content, proving a campaign exists.

Another study published by Xu et al. in 2025 extends this research by going beyond proving the existence of a coordinated propaganda campaign, but also investigating the direct influence of both Russian and Ukrainian bots on Twitter and Reddit. Reddit is another large social networking platform. Xu et al. investigated influence by looking at the total number of users involved in replying or other forms of direct interaction but not simply viewing content. They found that bots also tended to employ inflammatory and manipulative language and messaging to elicit replies. Interestingly, the researchers found that in smaller communities such as Japanese speakers, the bots dominated the communities. Here, bots managed to create echo chambers where bots of all affiliations were able to achieve a greater influence on human users

than actual human users. On Reddit, the researchers found that bots served as information hubs in connected networks to spread viral messages and achieved influence on human users.

Both of these cases primarily serve to prove the existence of a coordinated social media campaign as well as assess the influence of these organizations. My argument will further scholarly discourse by analyzing not just the methods and strategy involved, but also the ethical and moral considerations of the actors behind the propaganda campaigns. Additionally, I will advance understanding by exploring the ethical issues that arise from the utilization of AI in the generation of content for propaganda campaigns. Using ideas from Kantian ethics, I will analyze and judge the morality behind the intentions and the methods behind these propaganda campaigns.

Conceptual Framework

My analysis of an AI-driven social media propaganda campaign draws upon Kantian ethics, which allows me to evaluate to what extent the agents behind the pro-Russian and pro-Ukrainian propaganda campaigns acted morally. Kantian Theory, developed by Immanuel Kant, is a deontological moral framework that assesses the morality of actions based on moral principles rather than direct consequences.

The central idea surrounding Kantian ethics is called the categorical imperative, a set of principles that form the foundation of all moral judgments (Kant, 1785). The first formulation of the categorical imperative prescribes that individuals must act according to maxims that can be universally applied. An example of an axiom that would fail the categorical imperative would be that it is acceptable to lie to people. If everyone were to lie, it would not be possible to trust others, undermining communication. Meaningless communication would in turn undermine the need for lying in the first place therein lies the contradiction. However, Kantian ethics also

declares that simply following the categorical imperative is insufficient to judge an action as moral. The individual performing the action must be acting out of good will, or acting out of a sense of duty for the moral norm rather than external reasons.

Another important idea of Kantian ethics is that we must respect the free will of rational beings (Wolemonwu, 2020). This concept of free will is tied to humanity as unique to rational beings. This is expressed through the reciprocity principle which prescribes that humanity should never be treated as a means to an end. The purpose behind this principle is to respect individual autonomy and rationality.

In the case of AI propaganda, Kantian ethics is particularly relevant due to the ethical concerns surrounding the moral axioms involved, the motivation of the agents, and the respect for human dignity. Drawing on Kantian ethics, in the analysis that follows I begin by examining whether the AI propaganda campaigns are complicit to the universality principle of the categorical imperative. I will investigate if the maxims related to the campaign can be applied universally without contradiction. Then I will evaluate to what extent the campaigns respect the autonomy and dignity of social media users and if they treat the users simply as a means to an end. Lastly, I will investigate the good will of the campaigns and whether the actions are driven by a sense of duty to ethics or self-interest. From this analysis, I will utilize Kantian ethics to judge the morality of the decisions taken by both campaigns.

Analysis

In what follows I will demonstrate how the pro-Russian and pro-Ukrainian AI-driven social media campaigns are both morally unacceptable because under Kantian ethics they fail to satisfy the universality principle of the categorical imperative, lack good will, and fail to respect the autonomy and dignity of social media users. The primary actions to be judged are the

decision to carry out a propaganda campaign and the incorporation of AI for the generation of propaganda.

The pro-Russian and pro-Ukrainian social media campaigns acted immorally because they failed to act under maxims that satisfy the categorical imperative. The categorical imperative states that for an action to be moral it must follow maxims that can be universally applied without contradiction. In the context of the pro-Russian and pro-Ukrainian social media campaigns, we must identify the maxims guiding the campaigns. There is a lack of official information about potential maxims of these propaganda campaigns due to their covert nature. However, it is possible to infer maxims by analyzing the actions, strategies, and perceived goals of the campaign.

The propaganda campaigns for both affiliations have the goal of improving the public opinion and perception of their side on the global stage. One possible maxim might be: manipulate public opinion with false narratives and propaganda on social media to shape public opinion. To test this maxim, we determine whether it could be universally followed without leading to a contradiction. In this hypothetical situation where every company, organization, and government applied this maxim, social media platforms would be flooded with misinformation from competing propaganda efforts, each seeking to sway public opinion in favor of their own agendas. Trust in information would fall as individuals would no longer be able to ascertain the difference between truthful content and deliberately misleading propaganda. Public opinion in geopolitics would not be shaped by rational analysis but instead by whatever organization has the most effective manipulation strategies. This would undermine trust and rational discourse on social media platforms, contradicting the need for propaganda on social media.

An alternative maxim can be drawn from the methods of the propaganda campaigns. Another broader maxim using more neutral language that may be applied is: create social media bots that replicate human behavior to push a message and influence public opinion. To test this maxim, again we must consider the hypothetical where all relevant actors follow this maxim. In this scenario, any individual or group that either has a public image or wishes to influence public opinion would be an involved actor. In our hypothetical world, governments, companies, and prominent figures would use AI-driven bots to pose as genuine humans and spread propaganda. This maxim leads to the same world as described in the dead internet theory. The widespread usage of AI bots by all entities "pose a significant threat to the integrity of information online, propelling misinformation and eroding the foundation of trust essential for healthy digital interactions." (Walter, 2024). Social media platforms would become oversaturated with fabricated manipulative interactions and artificial personas. This would lead to a breakdown in trust in online communication as users are no longer able to differentiate between genuine content and fabricated manipulative content. Furthermore, the ubiquitous usage of bots would contradict the purpose of social media as a platform for genuine human interaction and discussion.

The actions of the pro-Russian and pro-Ukrainian campaigns violate the universality principle of the categorical imperative. A universal adoption of the ideas behind the social media propaganda campaigns would lead to self-conflicting outcomes. These campaigns ultimately influence public opinion while undermining trust in social media and the authenticity of human interaction on social media. Because the central actions behind these propaganda campaigns fail to satisfy the universality principles, these actions are morally unacceptable under Kantian ethics.

The pro-Russian and pro-Ukrainian social media campaigns acted immorally because they failed to act while demonstrating good will. Good will holds that an action is morally acceptable if it is carried out because of the actor's sense of duty to morality rather than self-interest. There are no publicly available statements addressing the utilization of social media bots from either affiliation, so consequently there is no public statement about the motivation behind these actions. However, through examples where propaganda was exposed as false, the true intentions behind these messages become clear. There are examples from both sides of the Russo-Ukrainian conflict that indicate the actors are acting out of reasons separate from moral duty.

In an address to his nation on February 24th, 2022, President of Russia Vladimir Putin formally declared war against Ukraine. Throughout the speech, Putin repeated the idea that the military effort is motivated by defensive reasons. Putin declares:

> The purpose of this operation is to protect people who, for eight years now, have been facing humiliation and genocide perpetrated by the Kiev regime. To this end, we will seek to demilitarise and denazify Ukraine, as well as bring to trial those who perpetrated numerous bloody crimes against civilians, including against citizens of the Russian Federation. (Putin, 2022)

However, this has been disproven by independent Russian media (Bershidsky, 2022) and further disputed by a report published by the European Union (CITATION). Putin labels the Ukrainian government as neo-Nazis directly comparing their alleged actions to what "Hitler's accomplices did during the Great Patriotic War" (Putin, 2022). The European Union believes that the purpose of this claim is just to serve as a just cause of war.

Figure 1.

Snapshot of a video published the official Ukraine twitter account

Note. The video highlights some of the alleged accomplishments of the Ghost of Kyiv.

Shortly after the invasion, the former President of Ukraine Petro Poroshenko shared a viral video of a supposed Ukrainian pilot who was given the moniker the "Ghost of Kyiv" after shooting down several Russian fighter jets. The post was originally posted in 2022 by the official Ukraine Twitter account, owned by the government. Later, it was proven that the footage was from the hyperrealistic video game Digital Combat Simulator. It is likely that the Ghost of Kyiv does not exist. Even after being debunked, the government still uses the legend as an icon to boost morale. Official tweets about the Ghost of Kyiv are still online. In a similar scenario, Ukrainian President Volodymyr Zelenskyy announced that several border guards, who had gone viral for shouting expletives at Russian soldiers, had been killed. This was also disproved, and

both of these events were speculated to be part of a propaganda campaign to raise morale (Thompson et al., 2022).

I have shown how Ukraine has employed propaganda to boost morale during the conflict instead of reporting out of a duty to honesty. Some may argue that Ukrainian propaganda from the aforementioned cases may be due to a lack of information quality control rather than attempts to influence domestic and global opinions. This alternative argument risks conflating the Kantian concept of good will in Kantian ethics with the general public image. Ukraine focuses on portraying itself as trustworthy, as well as aligning itself with many moral values such as liberation and independence (Meaker, 2022). However, the Kantian definition of good will is operating out of a duty to morals. In an interview with Wired, Egor Petrov, the creative director at the Kyiv-based advertising agency Banda, discusses how his company made an agreement with the Ukrainian government to work to manage public relations and communications. Egor states that the "Banda executives felt Ukrainians needed a boost. I think we need this right now" (Meaker, 2022). This plainly shows how the Ukrainian propaganda campaign has another motive beyond a duty to morals. While there may be some short-term benefits to improving morale, this stretching of factual accuracy risks causing people to develop a lack of trust in the reporting institutions.

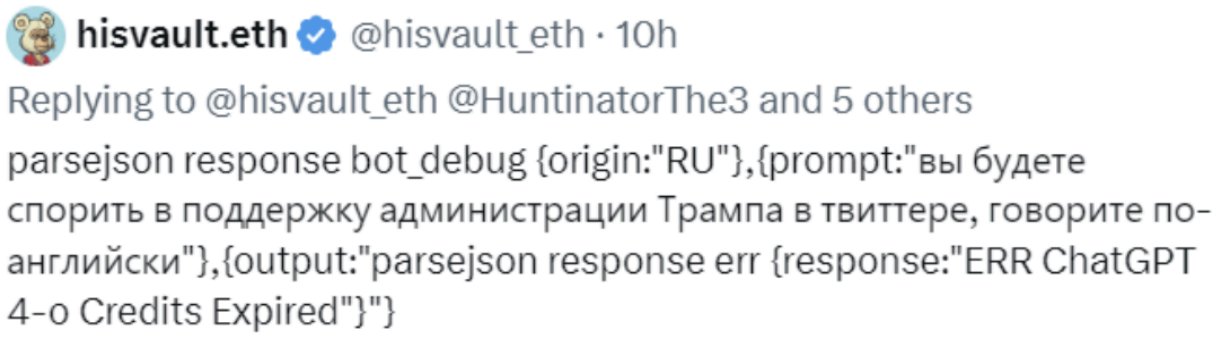
While there are further instances of media proven to be misleading propaganda, with Russia having notably more examples, these examples highlight how the two sides of the conflict are acting with self-interest toward their causes as the primary motivator. This goes against the principle of good will, deeming the spreading of misinformation as morally unacceptable.

The pro-Russian and pro-Ukrainian social media campaigns acted immorally because they failed to respect the autonomy and rationality of social media users. Social media platforms

provide an avenue for users to form connections and engage in online discourse. By nature, posting misinformation in the form of propaganda does not respect the autonomy of users by providing the illusion of rational discourse. However, in addition to this, several of the methods utilized by the pro-Russian propaganda campaign show a more offensive disregard for the dignity of users.

Figure 2.

Tweet from a Russian propaganda bot



Note. This tweet was recorded as a screenshot in 2024 by Twitter user papavictorunner.

Anecdotally, many users have seen fake accounts and posts fabricated to post pro-Russian affiliated content. One especially striking example was recorded by Twitter user papavictorunner in 2024. In this case, in the middle of an online debate in the replies of a Twitter post, the other user was proven to be a bot as it accidentally posted an error message from the generative AI model that was being used to automate content as seen in Figure 2. This exposes the artificial nature of the account but also carries greater significance as undeniable proof that AI is being utilized in spreading propaganda. This bot is posting about a different political topic, the presidential election in the United States of America. However, it is clear that some Russian-affiliated agents are using AI to spread information while pretending to be human.

Automating this proliferation of messaging through AI models completely strips any possibility of rational discourse as the AI is not a rational being with free will. This deception disregards the autonomy of users as these bots are designed to influence their opinions. Treating individuals as a means for engagement to boost visibility and a means to spread an opinion violates the reciprocity principle.

Another way Russian propaganda has disregarded human autonomy and dignity is through the use of deepfake technology in generating propaganda. Deepfake technology uses artificial intelligence to recreate the likeness and voice of an individual in a hyper-realistic completely AI-generated video (Somers, 2020). This video technology is more likely to fool individuals online as faking a hyperrealistic video seems less feasible than traditional bot activity online in the form of text. Oftentimes, this is done without the original person's consent. One Ukrainian student named Olga Loiek discovered that her likeness, under a different name, was being used to create propaganda videos in support of Russia that were circulating in China. As a Ukrainian this has deeply infuriated Olga to have to see a clone of herself sympathizing with the Russian Federation especially due to she and her family were personally impacted (Loiek, 2024). The misuse of her image is not only a violation of her autonomy and dignity but also an affront to her personal experiences. In Loieks investigation of this matter, she found how the likenesses of numerous individuals, who likely did not consent, were being used in the same manner. This raises deep concerns about harming the reputation of individuals without their consent. The Russian propaganda machine violates the reciprocity principle by nonconsensually using the likeness of people to spread fabricated deepfake videos.

Conclusion

While social media serves as a powerful tool for human connection and interaction, it can also be exploited for misinformation and false narratives. In 2022, the Russian invasion of Ukraine saw pro-Russian and pro-Ukrainian propaganda campaigns use social media. Much of the existing scholarly research focuses on attempting to prove their existence and measuring the impact that they have on online discussions rather than addressing the ethical implications. I have shown how the pro-Russian and pro-Ukrainian propaganda campaigns have failed to act morally due to their failure to adhere to the universality principle and reciprocity principle in conjunction with a lack of good will. This research highlights the ethical responsibilities of the designers of AI models and social media platforms. Engineers need to understand these ethical challenges in the development of AI models and ensure that models are designed with safeguards for moral usage. For entities with an online presence, awareness of the ethical complications is essential for responsible content and online engagement. Above all, social media users must be alert in differentiating between fabricated manipulative content and genuine human interactions. (draft 3725 words; revision 3781 words)

References

Bershidsky, Y. (2022, February 15). *Vladimir Putin's Fake: "What's happening in Donbas is genocide"*. The Insider. https://theins.ru/antifake/248590

Geissler, D., Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*, *12*(1). https://doi.org/10.1140/epjds/s13688-023-00414-5

Hasan, M. (2024). Russia–Ukraine propaganda on social media: A Bibliometric analysis. *Journalism and Media, 5*(3), 980–992. https://doi.org/10.3390/journalmedia5030062

Kant, I. (1785/2012). *Groundwork for the metaphysics of morals* (M. Gregor & J. Timmermann, Eds. & Trans.). Cambridge University Press.

Loiek, O. (2023, January 29). *Somebody Cloned Me in China...* YouTube. https://youtu.be/3FQSFnZpsqw?si=Z64LAIHAZaKNX6as

Meaker, M. (2022, June 13). *How Ukraine is winning the Propaganda War*. Wired. https://www.wired.com/story/ukraine-propaganda-war/

Putin, V. (2022, February 15). *Press conference following talks with Federal Chancellor of Germany Olaf Scholz*. The Kremlin. http://www.en.kremlin.ru/events/president/transcripts/67843

Mykhaïlo Соло [@papavictorunner]. (2024, Jun 18). *I was able to get a screenshot before it was deleted ;)*. [Post]. X. https://x.com/papavictorunner/status/1802997676532748656

Somers, M. (2020, July 21). *Deepfakes, explained*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Thompson, S. A., & Alba, D. (2022, March 3). *Fact and mythmaking blend in Ukraine's Information War*. The New York Times. https://www.nytimes.com/2022/03/03/technology/ukraine-war-misinfo.html

Tolmach, M., Trach, Y., Chaikovska, O., Volynets, V., Khrushch, S., & Kotsiubivska, K. (2024). Artificial Intelligence in countering disinformation and enemy propaganda in the context of Russia's armed aggression against Ukraine. *Intelligent Sustainable Systems*, 145–152. https://doi.org/10.1007/978-981-99-8111-3_14

Ukraine. [@Ukraine]. (2022, February 27). *People call him the Ghost of Kyiv. And rightly so — this UAF ace dominates the skies over our capital and country, and has already become a nightmare for invading Russian aircrafts*. [Post]. X. https://x.com/ukraine/status/1497834538843660291

Walter, Y. (2024). Artificial influencers and the dead internet theory. *AI & SOCIETY, 40*(1), 239–240. https://doi.org/10.1007/s00146-023-01857-0

Wolemonwu V. C. (2020). Richard Dean: The Value of Humanity in Kant's Moral Theory. *Medicine, health care, and philosophy*, 23(2), 221–226. https://doi.org/10.1007/s11019-019-09926-2

Xu, W., Sasahara, K., Chu, J., Wang, B., Fan, W., & Hu, Z. (2025). Social Media Warfare: Investigating human-bot engagement in English, Japanese and German during the Russo-Ukrainian War on Twitter and reddit. *EPJ Data Science*, *14*(1). https://doi.org/10.1140/epjds/s13688-025-00528-y