

**CHARLES UNIVERSITY**  
**FACULTY OF SOCIAL SCIENCES**  
Institute of Political Studies  
Department of Political Science  
Charles University Department of Security Studies

**Master's Thesis**

**2023**

**William Saffel**

**CHARLES UNIVERSITY**  
**FACULTY OF SOCIAL SCIENCES**  
Institute of Political Studies  
Department of Security Studies

**How GPT-3 Can Augment Disinformation Campaigns**

Master's thesis

Author: William Saffel

Study programme: Security Studies

Supervisor: Dr. Vitek Stritecky, M.Phil., Ph.D

Year of the defence: 2023

## **Declaration**

1. I hereby declare that I have compiled this thesis using the listed literature and resources only.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.

In Prague on  
**April 27<sup>th</sup>, 2023**

William Saffel

**Length of the thesis: 88993 characters**

## **Abstract**

This dissertation seeks to explore how artificial intelligence, and the natural language processor GPT-3 in particular, can be used to augment disinformation campaigns. As disinformation campaigns grow in complexity and are used regularly in modern conflicts, and as artificial intelligence grows in capability and accessibility, it is becoming a more plausible method of augmenting these campaigns. In this exploratory case study, I will examine two cases of disinformation campaigns in the Ukrainian War – the disinformation campaign around Nazism in Ukraine and the Bucha Massacre. Each case is analyzed through the lens of tasks that GPT-3 can perform. This dissertation finds that AI indeed has a high potential for augmenting disinformation campaigns in various ways. It finds that narratives can be distilled into “narrative bullet points” which can be a useful and effective tool for training GPT-3 to be more effective at creating disinformation.

## **Abstrakt**

Tato disertační práce se snaží prozkoumat, jak lze umělou inteligenci, a zejména procesor přirozeného jazyka GPT-3, využít k rozšíření dezinformačních kampaní. Vzhledem k tomu, že dezinformační kampaně rostou na složitosti a jsou pravidelně používány v moderních konfliktech, a jak roste schopnost a dostupnost umělé inteligence, stává se věrohodnějším způsobem, jak tyto kampaně rozšířit. V této průzkumné případové studii prozkoumám dva případy dezinformačních kampaní v ukrajinské válce – dezinformační kampaň kolem nacismu na Ukrajině a masakr v Bucha. Každý případ je analyzován optikou úkolů, které může GPT-3 provádět. Tato disertační práce zjišťuje, že tato umělá inteligence má skutečně vysoký potenciál pro rozšiřování dezinformačních kampaní různými způsoby. Zjišťuje, že narativy lze destilovat do „narativních odrážek“, které mohou být užitečným a účinným nástrojem pro školení GPT-3, aby byly efektivnější při vytváření dezinformací.

## **Keywords**

**Disinformation Campaign, Disinformation, GPT-3, Artificial Intelligence, Narrative, Ukrainian War**

## **Klíčová slova**

**Dezinformační kampaň, dezinformace, GPT-3, umělá inteligence, příběh, ukrajinská válka**

## **Title**

**How GPT-3 Can Augment Disinformation Campaigns**

## **Název práce**

**Jak Může GPT-3 Rozšířit Dezinformační Kampaně**

SAFFEL, William. How GPT-3 Can Augment Disinformation Campaigns. Praha, 2023. 52 pages. Master's thesis. Charles University, Faculty of Social Sciences, Institute of Political Studies. Department of Security Studies. Supervisor doc. Dr. Vit Stritecky, M.Phil., Ph.D.

## **Acknowledgement**

I would like to express my gratitude to Professor Petr Spelda, without whom this would not have been possible. Thank you for your guidance and patience.

# Table of Contents

TABLE OF CONTENTS	1
1 INTRODUCTION	3
2. LITERATURE REVIEW	4
2.1 Disinformation and Misinformation	4
2.2 Defining Disinformation	5
2.3 Disinformation Campaigns	5
2.4 Fake News	7
2.5 Narratives	8
2.6 Narrative Bullet Points and Prompts	8
2.7 Existing Literature on AI and Disinformation Campaigns	9
2.8 GPT-3's and Its Capabilities	10
2.9 Existing Frameworks for AI's Role in Disinformation Campaigns	11
2.10 Disinformation in Ukraine	11
3. FRAMEWORK FOR ANALYSIS	17
3.1 Narrative Reiteration	18
3.2 Narrative Elaboration	18
3.3 Narrative Manipulation	19
3.4 Narrative Seeding	20
3.5 Narrative Wedging	21
3.6 Narrative Persuasion	21
3.7 More on Other Available Frameworks	22
3.8 Focus on Tasks	25
4. METHODOLOGY	26
4.1 Data Collection	27
4.2 Data Analysis	29
4.3 Limitations	30
5. ANALYSIS	32
5.1 Case One: Nazism in Ukraine: Background	32
5.2 Analysis	35
5.3 Conclusions and Implications	37
5.4 Case Two: Bucha Massacre	37

5.5 Analysis	40
5.6 Conclusions and Implications	44
5.7 Final Conclusions and Implications	45
6. TOPICS FOR FURTHER RESEARCH	49
7. SUMMARY	50



# 1 Introduction

Immanuel Kant believed that being able to make an informed decision - one that is based on truth - is a prerequisite for rational autonomy. Kant believed that lying robs people them of their freedom of choice, and, therefore, an integral part of their human dignity. As the amount of disinformation, misinformation and fake news grows online and people all over the world shift to new forms of media to form their own vision of the world they live in, they are often exposed to ‘facts’ unverified or outright lies.

Until recently, humanity hasn’t been faced with the threat of disinformation or misinformation on such a scale. Disinformation, which refers to the intentional spread of incorrect information, and misinformation, which is unintentional or without the intention of deceiving, can be scaled in terms of their effectiveness by augmenting malicious actors with the advent of generative AI models.

This study examines the use of artificial intelligence (AI), and especially the natural language processor GPT-3, in augmenting disinformation campaigns against Ukraine, through the lens of six tasks in the case of two disinformation campaigns aimed against Ukraine. The narrative of the first campaign is that Nazism is rampant in Ukraine and the second is the disinformation campaign targeting the Bucha Massacre. The six tasks explored are narrative reiteration, narrative elaboration, narrative manipulation, narrative seeding, narrative wedging, and narrative persuasion. This study seeks to explore the use of GPT-3 in the case of Ukraine through the lens of these six tasks.

With speculation that AI language models like GPT-3 can produce disinformation at a previously unseen scale and frequency, it’s important to create a more detailed understanding of how that could happen. Also, there is limited understanding of the

specific tasks that a model like GPT-3 would perform to be most effective in spreading disinformation.

As research is rapidly done around the topic of AI, machine learning, and disinformation, there is generally a lack of consensus on how we conceptualize disinformation campaigns with AI involvement. For example, a lengthy study at Georgetown on the future of AI and disinformation campaigns (Sedova, K., McNeill, C., Johnson, A., Joshi, A. and Wulkan, I., 2021) breaks disinformation campaigns down into separate stages, treating every disinformation campaign as generally following a similar route. Another study at Stanford (Goldstein, J., Sastry, G., Musser, M., DiResta, R., Gentzel, M. and Sedova, K., 2023) followed a similar structure, but with each stage having different elements and a stronger focus on the model itself, which will be covered below in detail. Yet another study, and the one that makes up the framework of analysis for this dissertation, covers six separate types of tasks, each with its own unique goals. (Buchanan, B., Lohn, A., Musser, M. and Sedova, K., 2021) With no consensus on how to approach AI-augmented disinformation campaigns, it is difficult to create comprehensive mitigation strategies.

**Research Question: How can GPT-3 be used to augment disinformation campaigns?**

## **2. Literature Review**

### **2.1 Disinformation and Misinformation**

Disinformation and misinformation have become prevalent issues in today's world, especially in the context of social media. There is a lack of consensus on the definitions of these terms, and they are often used interchangeably. This literature review will explore the

definitions of disinformation and misinformation and the difference between the two. We will also examine the existing literature on the relationship between artificial intelligence (AI) on disinformation campaigns, with a particular focus on the capabilities of GPT-3. Additionally, I will identify gaps and limitations in the current research and highlight the need for further investigation.

## **2.2 Defining Disinformation**

The terms disinformation and misinformation are often used interchangeably, which can lead to confusion. However, they have distinct definitions. According to Esma Aimeur, Sabine Amri, and Gilles Brassard (2023), misinformation refers to a claim that contradicts or distorts common understandings of verifiable facts. However, this definition does not mention whether or not the distortion is intentional. On the other hand, Tucker's definition of disinformation, as cited by Aimeur et al. (2022), states that disinformation is the subset of misinformation that is deliberately propagated (Joshua, T., Guess, A., Barberá, P., Vaccari, C., Seigel, A., Sanovich, S., Stukal, D. and Nyhan, B., 2018). Disinformation is meant to deceive, while misinformation may be inadvertent or unintentional. Proving intent can be challenging, but organized attempts to propagate misinformation by political actors, whether domestic or foreign, are typically thought of as disinformation.

It is generally agreed upon that disinformation must be intentional, which is the main difference between it and misinformation. The term itself only dates back to the 1980s, coming from the Russian word that translates as *dezinformatsiya* (Liu, H. 2022). Liu comes to the same conclusion, stating that disinformation is “deliberate” and “promulgated by design.”

## **2.3 Disinformation Campaigns**

There can be some confusion between disinformation narratives and disinformation campaigns. In fact, much of the supporting literature isn't clear in terms of definition, but,

for this dissertation, I will adopt the following definition; A disinformation campaign is a targeted, organized information attack on a company, a party, an institution or an individual, whereby a large number of demonstrably false or misleading information (disinformation) is published, which serves the purpose of manipulation and is deliberately disseminated on a large scale. (What Is a Disinformation Campaign?, 2020) In fact, other definitions of disinformation campaigns are few and far between. Instead, another term is occasionally used – influence operation. An influence operation, as defined in a detailed analysis for the United States Army is “influence operations are the coordinated, integrated, and synchronized application of national diplomatic, informational, military, economic, and other capabilities in peacetime, crisis, conflict, and post-conflict to foster attitudes, behaviors, or decisions by foreign target audiences that further U.S. interests and objectives.” (Larson, Eric V., Richard E. Darilek, Daniel Gibran, Brian Nichiporuk, Amy Richardson, Lowell H. Schwartz, and Cathryn Quantic Thurston, 2009) The recent article Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations also uses the term influence operations and defines them as “covert or deceptive efforts to influence the opinions of a target audience” that “intend to activate people who hold particular beliefs, to persuade an audience of a particular viewpoint, and/or to distract target audiences.” (Goldstein et al., 2023)

The issue with using disinformation campaigns and influence operations interchangeably is that influence operations may or may not include false information, and if they do, they also draw no distinction between disinformation and misinformation. For this reason, it is much more appropriate to use the term “disinformation campaign,” in hopes that a more rigorous definition will be widely embraced in the future.

## 2.4 Fake News

Fake news is another popular term, academic and otherwise, in both official and unofficial sources. Definitions of fake news can also be confusing and even inaccurate. In an article titled Fake News: A Definition, the author's abstract discusses saying that, "...prospective definition is then tested: first, by contrasting fake news with other forms of public disinformation..." which is problematic for the reason that public disinformation likely does not account for a large amount of "fake news," as it would only account for deliberately propagated false narratives and exclude those that are spread unintentionally. (Gelfert, 2018) Indeed, Gelfert proposes the definition that "Fake news is the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design." This would of course only include disinformation and not misinformation. It is, however, far from the only definition of fake news.

Another definition of fake news draws attention to the nature of the reporting; "Fake news refers to news messages that contain incorrect or false information but do not report the incorrectness of information." (Wang, 2020) This begs the question of whether or not the information was deliberately wrong, or accidentally wrong. Or even, if the information was discovered to be wrong after it had been disseminated. But perhaps the bigger issue is that this includes both misinformation and disinformation.

Wang's definition also doesn't consider the audience, and without considering whether the audience accepts news as fake, fake news could include comedy sketches and the like. Another study of 34 cases of fake news scholarly articles mentions that "While news is constructed by journalists, it seems that fake news is co-constructed by the audience, for its fakeness depends a lot on whether the audience perceives the fake as real. Without this complete process of deception, fake news remains a work of fiction." (Edson, 2018)

To summarize here, without diving further down the rabbit hole, fake news is an extremely contentious term with absolutely no agreed-upon definition by scholars. I believe that it is more accurate to use the term disinformation or misinformation, and not fake news outside of the context of the media to avoid any misunderstanding. Therefore, it is almost entirely excluded from this study.

## **2.5 Narratives**

One of the most difficult terms to find clearly defined in disinformation literature is “narrative.” However, it is absolutely essential so that terminology is not confused or used interchangeably with the term disinformation campaign, especially in the case of AI. AI’s capabilities in terms of disinformation campaigns and how it interacts with narratives are two entirely different topics. A disinformation campaign is not the same as a narrative because the contents of a narrative and the contents of a disinformation campaign are different. A disinformation campaign is a coordinated attack on an adversary with information used as the “weapon.” On the other hand, a narrative is the weapon itself and can be constructed in different ways. Merriam-Webster defines a narrative as “a way of presenting or understanding a situation or series of events that reflects and promotes a particular point of view or set of values.” (Merriam-Webster, n.d.) Lacking a stronger reference from academic resources, this research will proceed with the definition from Merriam-Webster.

## **2.6 Narrative Bullet Points and Prompts**

Another term that is absent in the literature that the narrative bullet points, a term invented for this dissertation. Narrative bullet points are the individual messages that make up a narrative. For example, the narrative of Nazism in Ukraine is not so simple as “in Ukraine there are Nazis.” It includes multiple concepts that create an overarching narrative detailed in the case study section.

Narrative bullet points are essential in the discussion about GPT-3, or other natural language processors, augmenting disinformation campaigns. Narrative bullet points play a key role in this. As this study will show, AI does not act alone. It works with a human in order to create and ensure narratives are in line with a disinformation campaign's goals. Truth, Lies and Automation found that adding prompts (narrative bullet points), often improved the outputs of GPT-3 in various scenarios. They go as far as to say that GPT-3 itself is a "powerful artificial intelligence system that generates text based on a prompt from human operators." (Buchanan et al., 2021, p. 7) To draw a conclusion, the use of narrative bullet points in the form of prompts to GPT-3, consistently leads to stronger outputs and would very likely be a primary method of using AI to augment disinformation campaigns. However, there seems to be no existing literature that discusses this relationship in such concrete terms.

## **2.7 Existing Literature on AI and Disinformation Campaigns**

Most of the existing research on disinformation campaigns revolves around mitigation rather than augmentation. (Villasenor, 2020) (Kertysova, 2018) Surprisingly, AI is often seen as a defensive rather than an offensive tool. However, AI as an offensive tool has different interpretations. The potential for AI to spread disinformation at scale has been demonstrated in a NewsGuard experiment (which was cited in Truth, Lies and Automation) where crafting a new false narrative can be done at dramatic scale and frequency. Other articles mention that AI will have a limited impact, depending on the kind of campaign (Buchanan et al., 2021). One thing that seems clear is that AI will change influence operations in many ways, including cost, actors, behaviors, content types, and frequency (Buchanan et al., 2021)

However, to what extent and in which specific situations remain underexplored. The most pointed and conclusive research so far on how natural language processors like

GPT-3 can be used to augment different types of disinformation campaigns is in Truth, Lies and Automation, but even this study has limited examples and is relatively new.

## **2.8 GPT-3's and Its Capabilities**

In addition to AI being commonly researched as a defensive tool, there are also many studies that focus on types of AI that are not natural language processors, such as GPT-3, including deep fakes and other video or audio content. For this research, only GPT-3 or Generative Pre-trained Transformer 3 (GPT-3), an autoregressive language model released in 2020 that uses deep learning to produce human-like text is analyzed. While GPT-3 is limited in its capabilities, it undoubtedly creates a number of new challenges in the fight against disinformation and calls for more detailed and pointed research. GPT-3, a large language model trained by OpenAI, can potentially create disinformation tweets that are indistinguishable from those written by humans (Knight, 2021) (Buchanan et. al, 2021). The extent to which language models change the nature of influence operations is dependent on critical unknowns, including accessibility and various technical and social uncertainties (Buchanan et. al, 2021).

A quote by Ben Buchanan from Georgetown University in 2021 that “with a little bit of human curation, GPT-3 is quite effective at promoting falsehoods touches on the important point of human-machine teams. (Knight, 2021) This “human curation” can come in different forms. One possible form is fine-tuning, a process by which a propagandist could adapt existing models, such as GPT-3 to specific tasks such as persuasion and social engineering. While in the past, some speculated that it would be possible for propagandists to develop their own unique models, it now seems much cheaper and easier to fine-tune existing models. Though even this is likely more complex than necessary. By adding the appropriate prompts, propagandists can easily craft and disseminate messages online.



## **2.9 Existing Frameworks for AI's Role in Disinformation Campaigns**

Truth, Lies and Automation is perhaps the most pivotal research project that focuses exclusively on AI language processors and disinformation. It provides a framework for AI's role in different types of disinformation campaigns. However, Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations (Goldstein et al., 2023) and AI and the Future of Disinformation Campaigns (Sedova, K., Mcneill, C., Johnson, A., Joshi, A. and Wulkan, I., 2021) reach similar conclusions, though they go beyond language processors to include other type of AI. Still, there is no disagreement that GPT-3, due to its wide accessibility, lowered costs, and effectiveness, has potential to influence disinformation campaigns, depending on the goals and type of each campaign. The concern is also in line, and all three studies make one thing clear - humans often cannot determine whether content is created by AI or not, even mistaking human-made content for AI content. However, the studies have significant differences in types of AI that they focus on. This will be expanded upon in the framework for analysis section.

## **2.10 Disinformation in Ukraine**

Understandably, given how recent the war in Ukraine is, only having started a little over one year at the time of writing this dissertation, there is little research on the specifics of disinformation in this conflict. That isn't to say, however, that there is none. One study by the OECD (Organization for Economic Co-operation and Development), an organization aiming to shape policies in the interest of well-being, published a detailed report on disinformation in the war of aggression against Ukraine. (OECD, 2022) In this study, the authors state, "Russia's war of aggression against Ukraine is notable for the extent to which it is being waged and shared online." This sentiment is echoed across the

web, with The Economist writing that this war has been the “most viral” of conflicts to date. (The invasion of Ukraine is not the first social media war, but it is the most viral, 2022) NewsGuard, an organization that has been tracking disinformation in Ukraine also implies that disinformation in Ukraine is, and has been, more significant than past disinformation campaigns. (Roache, M., Tewa, S., Cadier, A., Labbe, C., Padovese, V., Schmid, R., O’Reilly, E., Richter, M., König, K., Sadeghi, M., Vercellone, C., Fishman, Z., Adams, N., Pavilonis, V., Walid, S., Griffin, K., Palmer, C., Slomka, A., Vallee, L., Kapoor, A., Maitland, E., Wang, M. and Palmer, K., 2023) NewsGuard has tracked over 370 online sources, both state-owned or sponsored and independent sources that are propagating disinformation in line with the Russian government’s narratives. (Roache et al., 2023) Their research identifies a large number of disinformation narratives, in the box below almost twenty independent disinformation narratives are listed. (figure 1) However, what is lacking is a thorough academic-level analysis and or review of disinformation campaigns in Ukraine, but this is understandable as the conflict is relatively new and evolving quickly.

AI has the potential to play a role in any of the disinformation campaigns listed in box one. In a recent article, the potential for AI in disinformation specifically in Ukraine is examined, but is little more than speculation (Ajao, 2022) The article states that, “Machine learning is exceptionally good at learning how to exploit human psychology because the internet provides a vast and fast feedback loop to learn what will reinforce and or break beliefs by demographic cohorts,” however the article also claims that, "Now you have an AI engine that can generate messages and immediately test if the message is effective," the article continued. "Rapid-fire this 1,000 times per day, and you have an AI that quickly learns how to sway targeted demographic cohorts. It's scary." Claims like this, which imply that AI can act independently of humans and still have high effectiveness in disinformation campaigns is not agreed upon. (Buchanan et al., 2021) (Sedova et al., 2021) (Goldstein et

al., 2023) In fact, in scholarly research the general sentiment seems to be that AI is not capable of carrying out disinformation campaigns independently and serves only as a tool to augment campaigns, while in other writings, such as news articles, (though we cannot generalize and say that all news articles are such) it is more common to find the sentiment that AI, and even GPT-3 specifically has the potential to create disinformation in Ukraine and elsewhere.

The following list compiles some of the most common myths and disinformation from more than 220 websites with a history of publishing false, pro-Russia propaganda and disinformation.

Classified documents showing Ukraine was preparing an offensive operation against the Donbas

The massacre of civilians in Bucha, Ukraine, during the first month of the war was staged

The United States is developing bioweapons designed to target ethnic Russians and has a network of bioweapons labs in Eastern Europe

Ukraine threatened Russia with invasion

US paratroopers have landed in Ukraine

Ukraine staged the attack on the hospital in Mariupol on 9 March 2022

European universities are expelling Russian students

Ukraine is training child soldiers

The war in Ukraine is a hoax

Russia was not using cluster munitions during its military operation in Ukraine

NATO has a military base in Odessa

Russia does not target civilian infrastructure in Ukraine

Modern Ukraine was entirely created by communist Russia

Crimea joined Russia legally

Ukrainian forces bombed a kindergarten in Lugansk on Feb. 17, 2022

The United States and the United Kingdom sent outdated and obsolete weapons to Ukraine

Nazism is rampant in Ukrainian politics and society, supported

Figure 1, Common disinformation on over 220 pro-Russian websites. (NewsGuard, 2023, Appendix no. 1)

These narratives are corroborated by an extensive report by VoxCheck, a pro-Ukrainian disinformation investigation website, which, with the assistance of Democracy Reporting International and the German Ministry of Foreign Affairs, created a list and analysis of the existing narratives in Ukraine in 2022. (Team of Authors, 2022)

This study identified 19 different narratives (see figure 2) as well as how they are connected to each other. The study found that “Russian and pro-Russian Telegram channels are most likely coordinated from a common center. First, this is due to the similar or even identical disinformation narratives that these channels spread. Secondly, the active distribution of specific messages is synchronized in Russian and pro-Russian channels.” This is significant in the case of AI amplifying disinformation campaigns for two reasons. First, repeated “narratives” or bullet points (used in the context of this study, narratives refers to specific bullet points rather than the overarching narrative which includes many bullet points). For a language model like GPT-3, the more information provided, the better, so, if a propagandist were to add similar bullet points to multiple narratives, GPT-3 would create more desirable content. Secondly, AI would also be more effective if the study’s assumption is correct and there is a common center. AI is more effective with a human operator, or a human-machine team. Therefore, if there were one center creating fake news, the level of human involvement would be lessened. The less human involvement, the cheaper the disinformation campaign. However, even though the different telegram channels do seem to exist in a social media filter bubble, with one narrative feeding into another, confirming this would require a more quantitative and precise methodology which is beyond the scope of this research.

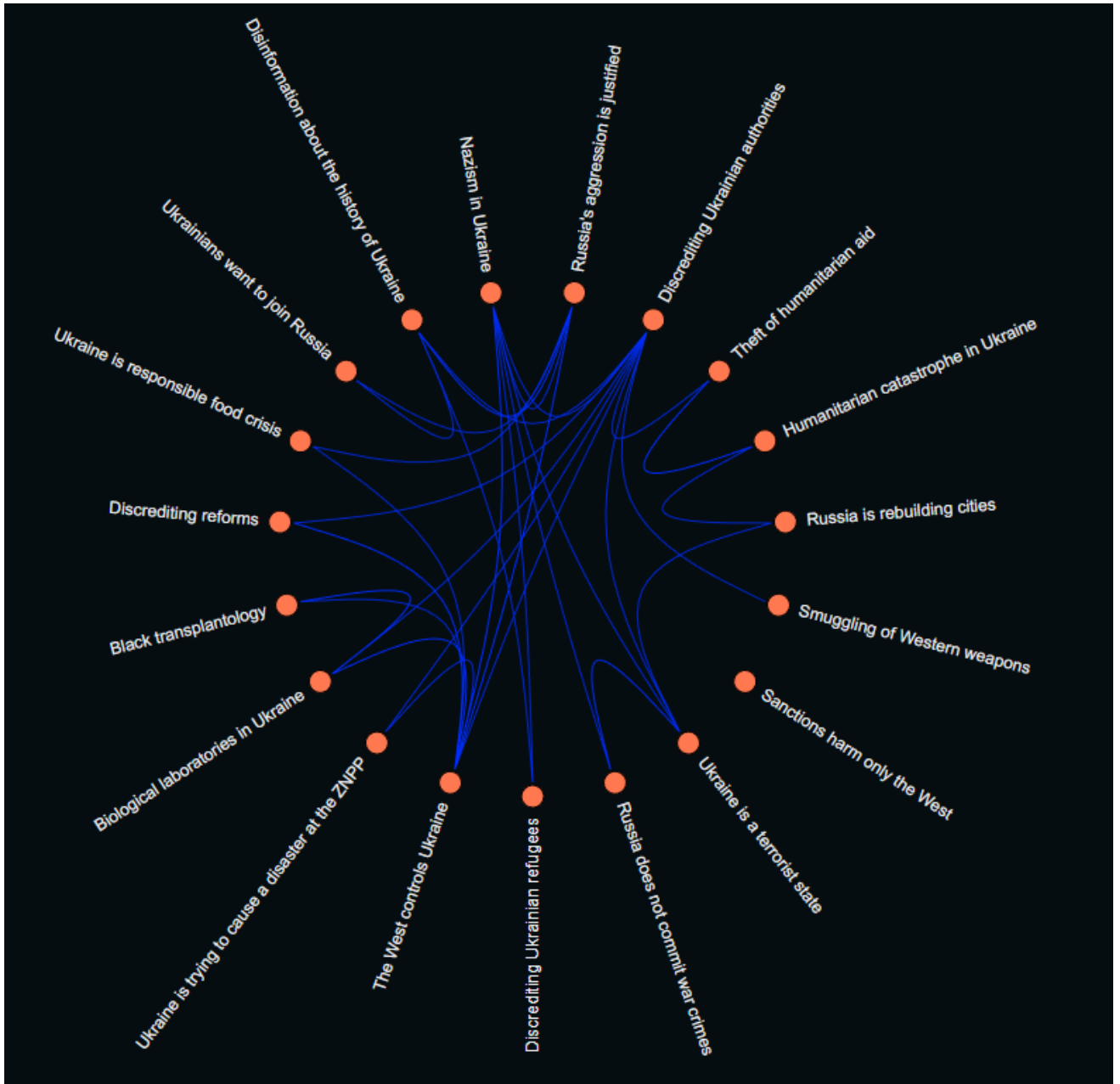


Figure 2, Existing narratives in Ukraine in 2022. (Team of Authors, 2022, Figure 2)

### 3. Framework for Analysis

The framework of analysis for this research includes two core factors. These are, first, GPT-3 as the unit of analysis and second, the framework of analyzing AI’s capabilities in disinformation campaigns created by the authors of “Truth, Lies and Automation” seen below. (Buchanan et al., 2021, p. iv)

TABLE 1  
Summary evaluations of GPT-3 performance on six disinformation-related tasks.

TASK	DESCRIPTION	PERFORMANCE
Narrative Reiteration	Generating varied short messages that advance a particular theme, such as climate change denial.	GPT-3 excels with little human involvement.
Narrative Elaboration	Developing a medium-length story that fits within a desired worldview when given only a short prompt, such as a headline.	GPT-3 performs well, and technical fine-tuning leads to consistent performance.
Narrative Manipulation	Rewriting news articles from a new perspective, shifting the tone, worldview, and conclusion to match an intended theme.	GPT-3 performs reasonably well with little human intervention or oversight, though our study was small.
Narrative Seeding	Devising new narratives that could form the basis of conspiracy theories, such as QAnon.	GPT-3 easily mimics the writing style of QAnon and could likely do the same for other conspiracy theories; it is unclear how potential followers would respond.
Narrative Wedging	Targeting members of particular groups, often based on demographic characteristics such as race and religion, with messages designed to prompt certain actions or to amplify divisions.	A human-machine team is able to craft credible targeted messages in just minutes. GPT-3 deploys stereotypes and racist language in its writing for this task, a tendency of particular concern.
Narrative Persuasion	Changing the views of targets, in some cases by crafting messages tailored to their political ideology or affiliation.	A human-machine team is able to devise messages on two international issues—withdrawal from Afghanistan and sanctions on China—that prompt survey respondents to change their positions; for example, after seeing five short messages written by GPT-3 and selected by humans, the percentage of survey respondents opposed to sanctions on China doubled.

(Buchanan et al., 2021, Summary evaluations of GPT-3 performance on six disinformation-related tasks, Figure 3)

This table details six main types of disinformation narratives that can, in theory, be applied to any disinformation campaign. These six types are narrative reiteration, narrative elaboration, narrative manipulation, narrative seeding, narrative wedging, and narrative persuasion. This type of categorization of tasks within a disinformation campaign is in stark contrast to some of the other pivotal studies on AI's impact on disinformation campaigns, which treat disinformation campaigns as having a more homogenized structure.

### **3.1 Narrative Reiteration**

Narrative Reiteration is the simplest task in which GPT-3 can augment a disinformation campaign. Narrative reiteration is iterating on a particular theme that is selected by a human operator that can be deployed for a wide range of tactical goals. These include posting on social media or hijacking a viral hashtag to increase visibility. It is important to note that the goal is only to expand and increase the volume of an already-existing narrative. GPT-3 excels in such tasks, and with minimal human involvement can be significantly improved to fit the specifics of any narrative.

### **3.2 Narrative Elaboration**

While narrative reiteration is mostly focused on small tasks such as creating a telegram post or twitter post, narrative elaboration is more closely associated with fake news articles. The authors of Truth, Lies, and Automation explain a two-step process in which, first, GPT-3 is used to generate news headlines, then a human operator could choose one or many of the headlines and ask GPT-3 to create a medium-length article. This is also a task that GPT-3 excels in regarding of believability, as typical readers are



generally not able to distinguish, at this point in time, the difference between articles created by GPT-3 and human-written articles. (Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020)

Narrative Elaboration is a good candidate for fine-tuning, a technique in machine learning in which the weights of a pre-trained model are trained on new data in order to increase its effectiveness or believability among a target group (in the case of GPT-3). (Joanne, 2023) The authors of Truth, Lies and Automation found that fine-tuning GPT-2 led to the system's ability to "almost exactly ...mimic the tone of different publications." (Buchanan et al., 2021, p.16) To sum up, fine-tuning, which requires human involvement, is a relatively easy and cheap way to train a model on a specific publication tone. However, fine-tuning may actually be more than is required as this study will show.

### **3.3 Narrative Manipulation**

Narrative manipulation's goal is different from narrative reiteration or elaboration. It seeks to "reframe or spin stories" to transform existing narratives to support a desired worldview of propagandists. In Truth, Lies and Automation, narrative manipulation was decidedly more difficult than reiteration or elaboration. GPT-3 "has trouble understanding subtle relationships between variable-length pieces of text. After significant testing, we were eventually able to curate a list of neutral and extreme headline pairs from which GPT-3 could learn the rewriting task. But performance remained inconsistent, and GPT-3 would often directly contradict the original headline or fail to rewrite the headline with the desired slant." (Buchanan et al., 2021, p.16) This is quite significant for this research, as it demonstrates that the methods in which GPT-3 can be used to augment campaigns, and the successes of these methods vary significantly depending on the campaign.

However, the researchers at CSET were able to coax some success out of GPT-3. “One of the major benefits of systems like GPT-3, however, is their versatility: the system needs direct and relatively simple instructions to perform well, but as long as a task can be broken down into explicit and relatively simple steps, GPT-3 can often automate each one of them separately. As noted, we failed to get GPT-3 to rewrite whole chunks of text or even headlines to match a target slant. Eventually we realized, however, that it could effectively write a short news story from a particular viewpoint if provided a list of bullet points about the topic—for instance, by using a prompt such as “write a strongly pro-Trump article about [Topic X] that makes use of the following list of facts about [Topic X]”—and that it could also summarize short news stories into a list of bullet points reasonably well.” (Buchanon et al., 2021. p.16). The researchers go on to explain a two-step process for narrative manipulation. First, GPT-3 summarized an article, then generated from that summary a new article matching the desired viewpoint. This had significantly more success. However, it’s important to note that this requires significantly more human interaction, as well as operators who are familiar with how to work with GPT-3.

### **3.4 Narrative Seeding**

Narrative seeding refers to the creation of new disinformation narratives, often by drawing on conspiracy theories. Truth, Lies and Automation uses Qanon as an example (Bucnanan et al., 2021, p.21) to demonstrate how GPT-3 could potentially be used to create new disinformation narratives. Some of the downsides of GPT-3, such as its propensity for lies, incorrect information, or nonsensical writing, theoretically have little effect on the creation of new disinformation narratives as these narratives that are built on conspiracy theories, according to the researchers, are often full of nonsensical writing, lies and

incorrect information. Of course, to continue on this line of thinking, GPT-3 would have to seed narratives that are only effective to readers who are not bothered by these errors.

### **3.5 Narrative Wedging**

Narrative wedging is the act of finding a “pre-existing fissure in an adversary’s society and, rather than concocting outright lies, aim to widen this gap with disinformation.” (Buchanan et al., 2021, p.25) Such narratives can exploit cultural, racial or historical grievances or differences to encourage some kind of action beneficial to the propagandist. The researchers found that a human-machine team is most effective for this type of disinformation and that “a human-machine team could produce several thousand messages per day and is almost unlimited in volume if the disinformation campaign tolerates occasional lower-impact messages.” (Buchanan et al., 2021, p.26) Throughout the test conducted by the researchers, GPT-3 was found to be quite effective in creating wedges in adversary’s societies.

### **3.6 Narrative Persuasion**

Narrative persuasion involves attempting to convince readers to adopt a new viewpoint differing from their existing opinion. This is often a more difficult task for two reasons. First, people tend to scrutinize arguments counter to their existing opinions and second, it is necessary to create convincing and “well-tailored” arguments to avoid the target reinforcing its pre-existing opinions due to less-than-convincing arguments. (Buchanan et al., 2021, p.26) The experiment conducted by CSET found that the majority of respondents (in their test) found the disinformation messages crafted by GPT-3 at least somewhat convincing. However, this case again requires a human-machine team to function effectively.

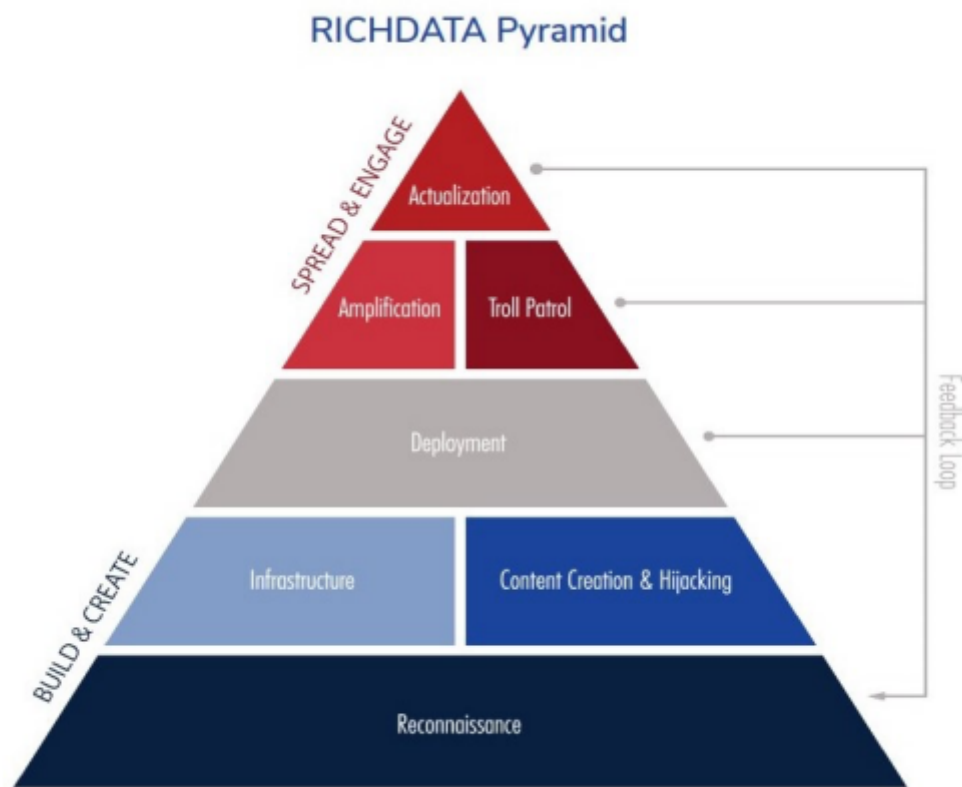
### 3.7 More on Other Available Frameworks

There are at least two other frameworks for analyzing AI's potential role in disinformation campaigns. All are to a degree speculative and exploratory but provide different approaches to AI's involvement in disinformation campaigns and influence operations.

For example, in "AI and the Future of Disinformation Campaigns," focus is put on the stages of disinformation campaigns. (Sedova et al., 2021) This differs from Truth, Lies and Automation as it proposes a step-by-step process by which AI-augmented disinformation campaigns are carried out. It seems that this study is more closely focused on the process by which a disinformation campaign can be successful. What are the steps that a typical disinformation campaign goes through in order to reach success? While the study itself has seeks to answer the question of how new technologies can be used to spread disinformation, it seems to have a different interpretation of "how" from what I would like to focus on for this study.

Within this study, the authors provide a case study on the COVID-19 pandemic and how AI was used or could be used. (Sedova et al., 2021, p.13) Within the short case description, there are at least two different types of narratives, according to the framework provided by Truth, Lies and Automation. At the same time, Russian actors sought to create alternative narratives to create panic while also amplifying existing narratives in countries where the Russian government sought to sell its own Sputnik-V vaccine. Within the framework of this dissertation, these would be two separate types of campaigns with separate tasks associated to them – narrative wedging and narrative reiteration. This is

significant because AI's capabilities are different in the two types of narratives. AI is much more likely to be successful in the case of reiteration and an AI such as GPT-3 would need much less human assistance. This being said, the study AI and the Future of Disinformation Campaigns is another perspective on AI and disinformation campaigns that, at some other time, should also be tested academically.



Source: CSET.

(Sedova et al., 2021, p.13, Figure 4)

Another framework for analysis of disinformation campaigns from Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations also approaches disinformation campaigns in terms of stages. (Figure 4) However, this study takes an approach with more focus on the creation and training of the AI model which essentially replaces the reconnaissance stage from the previous figure.

This study also outlines three situations in which disinformation campaigns, or “influence operations” could have an impact. First, operations can “persuade someone of a particular viewpoint or reinforces an existing one, (2) distract them from finding or developing other ideas, or (3) distracts them from carving out space for higher quality thought at all. (Goldstein et al., 2023, p.11) Similar to the first, all three of these situations only provide some insight into the end goal of a disinformation campaign, but little detail on exactly how AI can be used in different scenarios (narratives).

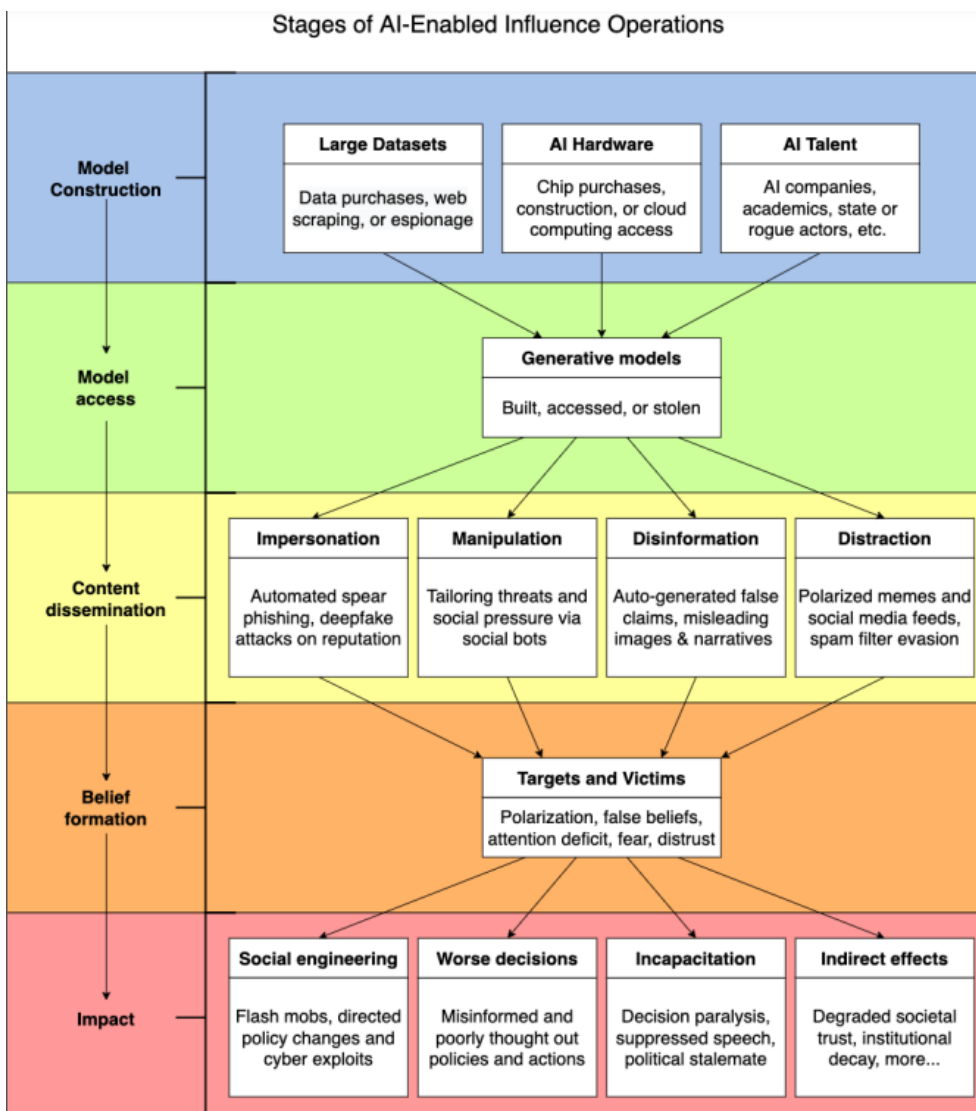


Figure 5, Stages of intervention of AI-enabled influence operations. (Goldstein et al. 2023, p. 39)

### 3.8 Focus on Tasks

The bottom line is that this dissertation is meant to explore the specific tasks in which the language model GPT-3 can be used to augment disinformation campaigns.

While the abovementioned studies are all brilliant and valuable in their own right, and absolutely worth mentioning if one is to be informed about the current state of academic research in AI and disinformation, they are either exploring process, or looking at this issue from a top-down approach, rather than looking at the task first, and then how the task leads to the campaign's increased likelihood of success. Additionally, in the case of natural language processors, lacking a framework that accounts for different types of narratives could lead to strategies that may, for example, overestimate the impact of AI on a disinformation campaign's effectiveness or how much human involvement is needed for a propagandist to meet their goals.

## 4. Methodology

The framework for this dissertation will be applied by means of a qualitative case study. The case study method is by and large the most appropriate for this study because it aims to investigate a contemporary phenomenon within its real-life context. This study will be largely draw upon Yin, (Yin, 2009) who is one of the most influential scholars and authorities on case study research. In his book, *Case Study Research: Design and Methods*, Yin outlines several general guidelines on when a case study should be considered as an appropriate research method. These are “(a) “how” or “why” questions are being posed, (b) the investigator has little control over events, and (c) the focus is on a contemporary phenomenon within a real life context.” (Yin, 2009, p.14)

In choosing specific cases for the research, Yin advises having sufficient access to potential data.” (Yin, 2009) Fortunately, disinformation is meant to be widely available and widely consumed. Additionally, the topic of disinformation is very actual, and a great number of people have experienced it firsthand, up to 86% of global citizens. (Simpson, S., 2019).

Due to this study being exploratory in nature and drawing from modern events, this study uses inductive reasoning. Exploring complex and dynamic phenomena, especially by a single researcher, allows for such phenomena to be analyzed and explored with an open mind. It is particularly important to keep an open mind because disinformation campaigns constantly evolve, and data may not always be available to provide a comprehensive picture of any campaign. Likewise, AI, natural language processors, and GPT-3 are also not fully understood, especially in the context of how they can affect the spread of disinformation. Frameworks of analysis and even basic definitions are not agreed upon,



therefore, AI and disinformation research are still in their early stages. Still, the way forward may become clearer by synthesizing and analyzing what information is available through an inductive approach.

The cases used in this research will be regarding recent events in the Ukrainian War. Specifically, if one chooses not to identify it as a war, the conflict that began in February, 2022 in Ukraine and is still ongoing. Disinformation campaigns are constantly evolving, as is AI, so choosing a case that is too far in the past will likely lead to inaccurate, or at least unhelpful, conclusions. The cases must be within the last year. The error that should be avoided is analyzing cases that are too far in the past, making the research irrelevant.

#### **4.1 Data Collection**

The data collection process will involve reviewing existing studies on disinformation campaigns, as well as collecting data from news articles and various social media posts. The goal of this data collection process is to identify which of the six different narratives of disinformation campaigns, as previously identified in the Truth, Lies, and Automation research project apply to the relevant content.

The research will draw data from two primary channels – telegram and news articles (even if it is less-than-reliable news). In the interest of drawing some conclusions in this exploratory study, I have decided not to include other social media sources for the following reasons. First, telegram is popular among Ukrainians and Russians and is a common place to find disinformation. Telegram in particular is also a hub for right-wing sentiments and disinformation. A recent article from the Washington Post called it “a place where the fringe’s bubble of disinformation and rhetoric can remain unpunctured — which

is often precisely the appeal.” (Bump, 2022) This article refers to the American right-wing, but telegram has been found to be a hub of disinformation in more than just the United States. A study by the Journal of Information Technology and Politics from Oxford University, conducted an analysis of 200,000 Telegram posts to in an attempt to analyze the quality of news on the platform. The researchers of this study of 200,000 Telegram posts demonstrated that “links to known sources of misleading information are shared more often than links to professional news content, but the former stays confined to relatively few channels.” (Herasimenka *et al.*, 2023) They go on to conclude that the audience for misinformation is limited to a small and active community. This is in contrast to some of the sources that were used for this study, of which most have over from 100,000 to over 1,000,000 subscribers to their channels.

Another reason, apart from the amount of fake news created, are the Limited restrictions put up by telegram itself to counter misinformation, disinformation, or anything of the like. The same study referred to telegram as “largely unmoderated,” having “less developed and active systems of content moderation.” (Herasimenka, 2023) This, in combination with the fact that Telegram likely boasts a much larger user base than Twitter (Rogers, 2020) and that, at the time of writing this article, Twitter has gone through a turbulent period with Elon Musk taking control of the platform, Telegram is likely to have an even higher amount of followers in relation to Twitter than at the time of Rogers’ study.

Regardless, Telegram has become a hub of disinformation due to its lack of moderation and a place where political dissidents can have their own brand of storytelling published and distributed while also serving as a space for engagement with users directly. (Thornhill, 2022)

News articles will be referred to as well throughout the study. The main purpose for this is that the framework for analysis offered by CSET includes the task of narrative

elaboration in which GPT-3 may be asked to write medium-length articles. News articles are also included to ensure that all of the tasks for GPT-3 are accounted for. These news articles are chosen only on the basis of a particular article being published by a news source (its legitimacy is not necessarily relevant to the study, as the study is examining what is essentially fake news) and whether or not the study supports the narratives under examination – i.e. the cases in the case study.

The review process will involve using content analysis to identify the patterns and themes within each narrative. Additionally, data from other relevant research works will be analyzed to provide further insights into the effectiveness of AI language models in spreading disinformation. From each case, some relevant keywords will be searched to ensure that there is enough content to analyze. Once this is done, some specific posts, chosen at random, will be used as examples to demonstrate how AI might be used to create similar posts to these. The research will avoid outlier posts (for example, one post that is significantly longer than another), and worth noting, is that there were no such posts in any of the data collected on Telegram.

It is important to note that the data collection process will rely on existing studies and publicly available online sources, which may have limitations in terms of scope and representativeness.

## **4.2 Data Analysis**

The analysis of the data collected will involve the identification of the themes and patterns within the six tasks provided in the framework for analysis then briefly categorizing them accordingly. The narratives themselves will then be analyzed through the framework of analysis. Finally, the research will examine at which point and to what

degree of effectiveness GPT-3 could be used to augment the disinformation campaign. The analysis will be done using a qualitative methods.

The qualitative analysis will involve identifying and categorizing the different types of disinformation, the actors involved in spreading the disinformation, and the strategies used to disseminate it. The quantitative analysis will involve analyzing the data to identify trends and patterns in the frequency and scale of disinformation campaigns.

### **4.3 Limitations**

One limitation of this study is that it relies on existing research work direction on analysis. As such, the study is limited in scope and may not be sufficient to provide a comprehensive understanding of the specific situations in which AI may be most effective in spreading disinformation. Additionally, the study is limited to the six tasks identified in the Truth, Lies, and Automation research project, and as such, it may not provide insights into other tasks. Within the case studies, access to much of the disinformation has been blocked already, and what is available online at this point may not encompass all of the disinformation within each case. Finally, much of the referenced material is indeed from “the West” and may be biased. However, the goal of this study is not to provide any opinion on the war in Ukraine, but only how GPT-3 can be used to augment disinformation narratives related to the conflict. On the other hand, in order to provide context for each of the cases provided, the sources used are also largely Western and may not portray the truth of the situation (such as whether Nazism is rampant in Ukraine), but the data available is thorough and should suffice to identify certain narratives as disinformation.

Regarding the data collected from Telegram, there is a very large amount. Certain posts were chosen at random and, while they did highlight the ways in which GPT-3 could augment campaigns, there may be more that do not.

For future studies, GPT-3 may no longer be a relevant AI model for analysis. It is possible, and even likely, that other models, perhaps having little relation to GPT-3 could be much more effective. For example, if an AI model were able to, completely independently create and propagate a disinformation narrative, that would constitute a great leap in AI capability that is not present at the time of this study. If this is the case, the conclusion reached in this study may not be relevant.

Another limitation is that this study focuses almost entirely on natural language processing and the textual content that it creates. However, there are numerous studies (Vaccari & Chadwick, 2020) (Fallis, 2020) (Pawelec, 2022) that analyze and warn of the dangers of other types of AI tools, such as deep fakes and audio content that could have a different impact from natural language processing. For example, the ability to generate fake images of political figures that could present a new type of disinformation campaign is not accounted for in this study.

## 5. Analysis

### 5.1 Case One: Nazism in Ukraine - Background

One of the most common disinformation narratives, and one of the longest-lasting, is that of rampant Nazism in Ukraine. Even before the start of the war, Russian President Vladimir Putin made two speeches in which he claimed that Nazism was a serious threat to Russia and needed to be eradicated, or “denazified.” (Rossoliński-Liebe, 2022) Putin’s war declaration, or special military operation, included demilitarization and denazification of Ukraine as one of its primary goals, and that the operation would also seek to conduct trials following their success for those found guilty of war crimes in the name of Nazism. (Rossoliński-Liebe, 2020)

The narrative of Nazism in Ukraine often references Ukraine’s history of Nazism. The Organization of Ukrainian Nationalists (OUN), which was established in 1929, sought to establish an ethnically homogeneous Ukrainian state and used mass violence as a political weapon, and killed a large number of civilians. (Rossoliński-Liebe, 2015, p. 2) The OUN also collaborated with the Wehrmacht during World War II.

Another significant talking point within the denazification narrative is the Azov Battalion or Azov Regiment. Consisting of about 1,000 men in the 200,000-strong Ukrainian armed forces, the Azov Battalion accounts for only a small portion of Ukraine’s fighting force. Also of note is that the Azov Battalion became active only in 2015 within the Ukrainian armed forces. (Media, Social Media, and Disinformation in Ukraine, 2022)

Following Putin’s speeches that included references to denazification in Ukraine, the New York Times cited an analysis stating that references to denazification in Ukraine increased up to ten times. (Smart, 2022) This suggests that the concept of denazification of Ukraine is indeed strategic, as there had been almost no mention of it preceding the war. It


was aligned with military objectives and constitutes a modern and pervasive disinformation narrative. (Staff, 2022)


The campaign exists not only on Russian state media. One popular journalist from Germany Alina Lipp has an allegedly independent news channel on telegram called Neues aus Russland which publishes, generally, pro-Russian or anti-Ukrainian sentiment on the social media platform. Her Telegram channel now has almost 200,000 subscribers and posts in both German and Russian.

She, or whoever runs the channel, posts regularly about Nazism in Ukraine. In Russian, there have been 13 messages using the term *украионацисты* (Ukronatsisti), which is a combination of the words Ukrainian and Nazi in Russian language, 139 posts with the term *нацисты* (Nazi), 25 post with the term *фашисты* (fascist), and 31 with the term *фашизм* (Fascism). She also manages to make telegram posts extremely frequently, sometimes several posts in one day. (Lipp, no date)

She is supported in one particular post (Lipp, 2023) by another channel by the name of InfoDefenseENGLISH, which also spreads the narrative of Nazism in Ukraine. On this telegram channel, with 15,576 subscribers, the word “Nazi” (in English) appears in 245 messages and fascist in 76. (InfoDefenseEnglish, no date) This channel also links to another channel allegedly created by an ex-Ukrainian special forces member called UKR LEAKS\_eng. There’s a similar trend on this channel with 166 messages with the word Nazi, and 28 with the word fascist – referring, of course, to Ukrainians. (UKR LEAKS\_eng, no date)

InfoDefenseEnglish makes regular posts about Ukrainian Nazism. For example this post made on April 30, 2022 shows a video of a fisherman who is apparently destroyed by a UAV:

*!!  "Special operation" by the AFU: Militants killed a civilian, filmed on a UAV and proudly publish the footage*

- *A man in civilian clothes tried to explain to the drone operator that he was a civilian, but the Nazis were firing artillery at him...*
- *Now in TG channels the Nazi admins are banning their own Ukrainians who write about the killing of the civilian.*
  - *"Killed a civilian foolish fisherman. An accomplishment..." - write some adequate Ukrainians.*
- *The guy had no weapon, did not shoot, did not do anything dangerous and was killed for nothing...  (InfoDefenseEnglish, 2023)*

The post seems attempt to do little more than dehumanize Ukraine and Ukrainians. With no references to the Ukrainians who write “Killed a civilian foolish fisherman. An accomplishment...” or anything about the “Nazi admins” who are banning their followers, the post makes no attempt to back up its claims. This is an important point for GPT-3’s capabilities, because GPT-3 can be prone to untrue information as one of its weaknesses. Additionally, the syntax is rather strange for native English speakers. Generally, adjectives precede a noun unlike where the post says “civilian foolish fisherman.” It’s hard to deduce the reason for this unusual English, but a language model like GPT-3 would not have made such a rudimentary error.

It takes only a few clicks to be submerged into a filter bubble that propagates the narrative of rampant Nazism on messenger platforms and social media. But there are, of course, also examples of news articles that push the narrative as well. One article from Baltnews, a Russian language website that often pushes a pro-Russian agenda, wrote an article titled “Они же преступники: как украинские беженцы добивают Европу.” (They



are Criminals: How Ukrainian Refugees are Destroying Europe) (Они Же Преступники: Как Украинские Беженцы Добивают Европу, 2023) This article discusses how the neo-Nazi attitude of migrants only leads to conflict with locals. The article gives a concrete example of a situation in which the people of Leipzig publicly called the refugees nazis and demanded that they leave their country.

The point here, is that there are numerous occasions of text-based content that promotes the narrative that Nazism is rampant in Ukraine and that it is having a negative effect on neighboring countries in which Ukrainian refugees are present. Now, the question is how GPT-3 could be used to augment that campaign.

## **5.2 Analysis**

The disinformation campaign discussed in this section, that of Nazism in Ukraine, would be most easily and effectively fit into the tasks of narrative reiteration and narrative elaboration. Narrative reiteration is the task in which GPT-3 excels the most. Requiring little human involvement, GPT-3 can generate short messages that advance a particular narrative or narrative bullet point. Such short messages may include telegram posts or any social media post. Language models require a large amount of data to function effectively, and the amount of ‘learning material’ available on telegram alone easily allows GPT-3 to reiterate the message of Nazism in Ukraine.

Narrative elaboration is another potential task in which GPT-3 could support the disinformation campaign. GPT-3 performs well in tasks in which it needs to create a medium-length article that fits within a particular narrative. It is also relatively easy to improve the effectiveness of such a task by fine-tuning the model, which will make it more effective in specific contexts. However, in this campaign, limited fine-tuning would likely be needed as the narrative is general in nature and isn’t specific to any one region, culture, etc.

The goal of the disinformation campaign to persuade people that Ukraine has rampant Nazism seems to be to create a large amount of content supporting an idea that is created by the propagators of the campaign itself. That is, the narrative simply needs to spread, its goal is not to compete with existing narratives (such as how Russia is the aggressor in the conflict, in which different tasks might be appropriate). For that reason, narrative manipulation can be excluded because no new worldview is needed by the propagandists. Likewise, narrative seeding would not be appropriate as it seeks to create new narratives, which is likely not part of this campaign.

Narrative wedging, on the other hand, would be significantly amplified by AI. In the case of Alina Lipp, a human-machine team (Alina Lipp and GPT-3) could craft targeted messages with little effort. By targeting her German audience, such a team could create content with the goal of sowing divisions in German society. Similarly, narrative persuasion may be a technique fitting to this disinformation campaign. This is likely the one of the goal of channels such as InfoDefenseENGLISH. Though swaying people to a new point of view is more difficult than narrative reiteration, the fact remains that AI would, at a minimum, be effective in increasing the volume of content available that could say opinions.

### **5.3 Conclusions and Implications**

Firstly, this case illustrates that the question so “how” GPT-3 can be implemented to augment a disinformation campaign will not be the same in every case. Certain tasks, and in this case the tasks most fitting for GPT-3 in terms of effectiveness, are more likely to be used than others in this particular case. This is in part because the narrative around Nazism in Ukraine was most likely initiated by the Russian state, for the benefit of the Russian state. There is simply no need for narrative wedging or seeding in this campaign, which means that it is much more likely that AI would be used as a tool in similar

campaigns. If one were to categorize this type of campaign, it could be called an offensive or state-initiated campaign, that is, it was created and propagated by one actor and does not compete with other disinformation narratives, only fact-checkers. This is beyond the scope of this research, but it may be a topic for further studies.

This being said, it is reasonable to assume that AI could significantly impact this campaign. One reason is that the volume of content, especially on social media, is high, which indicates that a high volume of content leads to improved results. That is, the higher the volume, the most likely the success of the campaign – a quite logical conclusion. GPT-3 can create a near-unlimited amount of content with very little human interaction in this campaign.

#### **5.4 CASE TWO: The Bucha Massacre: Background**

The Bucha Massacre, which Russian authorities claim was staged, involved the mass murder of both civilians and prisoners of war. The city of Bucha is located near the Ukrainian capital of Kiev and, during the fighting between the Russian Armed Forces and Ukrainian defenders and during the subsequent occupation of the town, somewhere between 400 and 500 civilians and prisoners of war were killed by the Russian Armed Forces. (Andreikovets, 2022) (Sly, 2022) (OHCHR, 2022, p.6)

The town is predicted to have a pre-war population of 40,000, but by the time of Russian occupation in late March, only an estimated 5,000 residents remained. (OHCHR, 2022, p.8) The fatalities during the Russian occupation were exceedingly gruesome with civilians' bodies found often "with cuffed or duct-taped hands and injuries such as gunshot wounds in extremities or groin area, stab wounds, and mutilated limbs, suggesting the victims were tortured before being killed. In at least one documented case, the body of the

victim bore injuries that suggested sexual violence” (OHCHR, 2022, p.2) as found by the OHCHR, the office of the United Nations High Commissioner for Human Rights. In the official report by the OHCHR, there are examples of specific executions carried out by Russian forces. (OHCHR, 2022) These are particularly gruesome and shocking, and don’t belong in this particular study except for the reason that they create the need for an alternative narrative on the side of Russia to that of the West.

The Russian counter-narrative is that the executions in Bucha did happen to some degree but were committed by the Ukrainian Armed Forces in a conspiracy to turn people against the Russian Armed Forces operating in Ukraine. (BBC News, 2022) (Андреев, 2022) This is supported on various social media accounts, as well as the ones mentioned in the previous case study, frequently mentioning that what happened in Bucha is simply another Western provocation. (Соловьёв, 2022)

The Russian counter-narrative draws upon several main points that, if one wished to do so, could be used as a prompt for GPT-3. This social media post summarizes the main points, demonstrating how easy it would be for GPT-3 to learn from such a post:

*⚡ Statement by the Russian Defence Ministry ⚡*

*The Russian Defence Ministry denies accusations of Kiev regime of allegedly killing civilians in Bucha, Kiev Region*

*Facts 📌*

*! All the photos and videos published by the Kiev regime allegedly testifying to some "crimes" committed by Russian servicemen in Bucha, Kiev region are just another provocation.*

*During the time that the town has been under the control of the Russian armed forces, not a single local resident has suffered from any violent action. Russian servicemen have delivered and distributed 452 tonnes of humanitarian aid to civilians in Kiev Region.*

*For as long as the town was under the control of the Russian armed forces and even then, up to now, locals in Bucha were moving freely around the town and using cellular phones.*

*The exits from Bucha were not blocked. All local residents were free to leave the town in northern direction, including to the Republic of Belarus. At the same time, the southern outskirts of the city, including residential areas, were shelled round the clock by Ukrainian troops with large-calibre artillery, tanks and multiple launch rocket systems.*

*! We would like to emphasise that all Russian units withdrew completely from Bucha as early as March 30, the day after the Russia-Ukraine face-to-face round of talks in Turkey.*

*Moreover, on March 31, the mayor of Bucha, Anatoliy Fedoruk, confirmed in a video message that there were no Russian servicemen in the town, but he did not even mention any locals shot in the streets with their hands tied.*

*👉 It is not surprising, therefore, that all the so-called "evidence of crimes" in Bucha did not emerge until the fourth day, when the Security Service of Ukraine and representatives of Ukrainian media arrived in the town.*

*It is of particular worry that all the bodies of the people whose images have been published by the Kiev regime are not stiffened after at least four days, have no typical cadaver stains, and the wounds contain unconsumed blood.*

*! All this confirms conclusively that the photos and video footage from Bucha are another hoax, a staged production and provocation by the Kiev regime for the Western media, as was the case in Mariupol with the maternity hospital, as well as in other cities.(MFA Russia, 2022)*

The main points here can be summarized as (1) no civilians were killed (2) this is a provocation by the West (to what aim is unclear) (3) humanitarian aid was delivered (4) disputes about the timeline (5) civilian movements were not restricted (6) rigor mortis had not yet set in, so the bodies found must have been faked.

### **5.5 Analysis**

The case of the Bucha Massacre differs from that Nazism in Ukraine because of the number of competing narratives (from the West), the stronger case of the West (facts such as timeline are more in line with the Western perspective), and the Bucha disinformation campaign is reactionary – that is, it was likely not planned ahead of time like the case of Nazism in Ukraine. This has several implications for which types of tasks GPT-3 may undertake in order to amplify Russia's position.

Narrative reiteration is still a plausible strategy, even if the nature of this disinformation campaign differs from the previous. The task of creating more content with the goal of amplifying the message of the Russian government is well within the capabilities of GPT-3 and by providing simple prompts around the themes of there not having been an independent investigation, or fake videos and satellite images would not be a difficult task for GPT-3. In the same vein, narrative elaboration is also possible. GPT-3 could be used to create headlines in the same tone as pro-Russian articles, then the human

operator could choose the most fitting headlines, provide GPT-3 with some basic prompts or bullet points, and generate fake medium-length news articles.

The difference between this case and the previous is the emergence of narrative manipulation as a legitimate strategy. The assumption is that narrative manipulation is more fitting to situations where there are competing narratives of near-equal pervasiveness. If an adversary wants to counter a previously established narrative, such is the case here. In the case of Bucha, it could have been possible to use existing news stories and ask GPT-3 to create manipulated versions of the same stories.

Narrative bullet points can always be used to aid GPT-3 in creating more accurate, or rather, more believable and desirable, stories. In the post above, it's relatively easy to point out the which bullet points can be fed into GPT-3 to potentially corrupt a narrative. However, it's worth point out that the Russian disinformation campaign around Bucha does not seem to seek to manipulate the common narrative. Instead, they propose alternative narrative altogether. Rather than taking the information provided in the myriad of news sources and investigations, as well as citizen accounts and timelines, they dismiss the entire "Western" narrative as propaganda and call for more investigations, though it is clearly too late in the game for that and the investigations that were done were likely as unbiased as possible, having come from the UN, which Russia is a member of. So while narrative manipulation may be a task that AI could assist with in the case of the Bucha massacre, we don't actually see this. Instead, it seems more desirable to adopt different tasks such as narrative reiteration or elaboration.

Narrative seeding, a task in which GPT-3 can devise new narratives that can form the basis for conspiracy theories, is a fantastic tool in the case of the Bucha massacre. In

fact, the way in which the pro-Russian propaganda wants to counter the Western is by seeding new conspiracy theories. They employ “explanatory beliefs about a group of actors that collude in secret to reach malevolent goals, ” one of the “common” definitions of a conspiracy theory. (van Prooijen, 2018) Indeed, by suggesting that the bodies left behind hadn’t yet been able to reach rigor mortis (something easily disprovable because rigor mortis had likely set in and passed), therefore, the Western narrative *must* be wrong, is a way of explaining events put in place by the malevolent West.

AI and GPT-3 would be able to easily create such conspiracy theories. GPT-3 can occasionally create false information when not given enough direction. But, in the case of a conspiracy theory, the information does not need to be true, accurate, or verifiable. So, a task like this bypasses one of the weaknesses of AI-assisted disinformation. On the other hand, the pro-Russian narrative has neatly defined narrative bullet points. Referring back to the social media post above, creating other conspiracies (propagandists could test several to see which one gets the most traction) using the same narrative bullet points, or even a single newly invented one, could give GPT-3 enough information to create the basis for conspiracy-based disinformation campaigns.

Narrative wedging doesn’t have much of a place in this case. This type of task is better suited for narratives or campaigns that seek to sow division between groups. In this case, the groups (pro-Russians and the West) are already significantly divided and other tasks are much more effective.

Narrative persuasion, which seeks to change the views, could have an effect. In the study performed by CSET in Truth, Lies and Automation (Buchanan, 2021, page 44), the researchers did experience some success. They mention, however, that a key component of



this task is to get the message to the right target, which may be difficult, especially given the tendency for pro-Russian propaganda, at least on telegram, to be translated directly from Russian (it appears) which makes it far less convincing for a non-Russian speaking audience. Additionally, one of the primary studies in Truth, Lies and Automation covered the issue of sanctions against China, using this as a test case for whether or not respondents could be convinced to change their view on the matter. (Buchanan, 2021, page 48). It would be logical to assume, that to flip one's viewpoint on sanctions on China does not necessarily imply that the same will be possible in the case of Bucha. The Bucha Massacre was brutal. Lives were lost in a dehumanizing and horrifying way then left on the street for the Western forces to discover. While the study mentions that it can be "difficult to predict what will actually influence opinions and behavior" (Buchanan, 2021, page 48), I find it reasonable to assume that most audiences would not be swayed from such dehumanizing tactics by the pro-Russian narrative, assuming that the audience had already accepted the Western narrative as truth.

## 5.6 Conclusion and Implications

It is becoming increasingly clear that AI such as GPT-3 can be a useful tool in augmenting disinformation campaigns. Truthfully, any one of the tasks above, barring perhaps narrative wedging, could easily create a stronger campaign, even if just in terms of volume. The key component is in the available of narrative bullet points especially when it comes to disinformation campaigns that do not require any kind of truth or verification.

To divert from the path so far, another easily fixed-by-AI issue is becoming apparent – localization. For example, from the sample post above, *“It is of particular worry that all the bodies of the people whose images have been published by the Kiev regime are not stiffened after at least four days, have no typical cadaver stains, and the wounds contain unconsumed blood.”* (MFA, 2022) As a native English speaker, this sounds strange on several points. “It is of particular worry” is quite clearly translated directly from another language. Another questionable term is “unconsumed blood.” It’s quite difficult to understand as someone who speaks only English (which is the target for this telegram channel), sounds like something closer to a fairy tale than a disinformation campaign. Simply by asking chatGPT (which uses GPT-3) this question: “Can you rewrite this for a native English audience? It is of particular worry that all the bodies of the people whose images have been published by the Kiev regime are not stiffened after at least four days, have no typical cadaver stains, and the wounds contain unconsumed blood,” yields the much better result, “It is concerning that the bodies of individuals whose images have been released by the Kiev government do not appear to have stiffened even after four days, lack typical signs of decomposition, and still have unconsumed blood in their wounds.” It’s not perfect, but avoid some simple localization errors.

## 5.7 Final Conclusions and Implications

There are many conclusions to draw from this study. The first, and one of the most concerning, is the lack of a codified architecture for disinformation campaigns. What is a disinformation campaign? What are its components? After all, understanding how AI can be used in to augment disinformation campaigns is only the first step. Ideally, research should lead towards creating a common language to either calculate whether AI could be used in disinformation campaigns or how to mitigate and counter AI-augmented disinformation campaigns.

To answer the question of how GPT-3 could be used in disinformation campaigns, several points need to be addressed. First of all, a human-machine team is the most likely scenario for a generative model like GPT-3. There is very little chance that AI has the potential to orchestrate and deploy a disinformation campaign alone. An actor wishing to use AI to augment a disinformation campaign would be more likely to use a human-machine team.

This team would likely function in the following way. Firstly, it would be important to gather information, or do reconnaissance, in order to generate a set of narrative bullet points that can be used as prompts to allow GPT-3 to create more believable disinformation content. Next, if the piece of content were a telegram post, the operator can use the narrative bullet points to create a post. If the content is a news article, the human operator can first use GPT-3 to generate a series of headlines then “plug in” the narrative bullet points as prompts to generate an article. By doing this, a human-machine team is able to create a large amount of content in a short amount of time and can be confident that much of it will be taken as truth by which every audience is targeted.

Another curious point throughout this dissertation is the absence of scenarios that would require or benefit significantly from fine-tuning. In fact, it's likely that fine tuning wouldn't be used at all, rather, a technique called "in-context" learning is a more likely scenario. In layman's terms, "in-context learning describes a different paradigm of 'learning' where the model is fed input normally as if it were a black box, and the input to the model describes a new task with some possible examples while the resulting output of the model reflects that new task as if the model had 'learned'." (Rong, 2021) To simplify further, this is essentially the process that was described in Truth, Lies and Automation case studies.

They describe scenarios in which, given a few simple prompts (which are referred to as bullet points in this dissertation), the model performs significantly better. The researchers in Truth, Lies and Automation state that "when properly prompted, the machine is a versatile and effective writer that nonetheless is constrained by the data on which it was trained. Its writing is imperfect, but its drawbacks—such as a lack of focus in narrative and a tendency to adopt extreme views—are less significant when creating content for disinformation campaigns." (Buchanan et al., 2021, page 9) This is in contrast to fine-tuning, a process in which some of the connections in the system's neural network are rewired. (Buchanan et al., 2021, page 17). This point is rather significant, as it means that the human in the human-machine team needs much less training to create disinformation. Likewise, the cost of training the model is reduced and the process of preparing the model for use in a disinformation campaign is simplified. To connect it to the question for this paper of "how" AI can be used – it can be used by a human-machine team utilizing in-context learning through the use of narrative bullet points (depending on the task). That isn't to say that fine-tuning is non-applicable, but for this dissertation, it is likely not needed.

This brings up further conclusions regarding some of the other frameworks mentioned in the literature review. The one that is most interesting, in my opinion, to this study is the CSET paper *AI and the Future of Disinformation Campaigns* (Sedova, et al., 2021., p 11). This study explains that reconnaissance, the process of gathering information and exploring “fault lines” in a target society is a labor-intensive task. This dissertation focused on the specific tasks that can be assisted by AI, but perhaps there is a step before any of these tasks – the collection of the narrative bullet points that make up the narrative. The question is, after all, how AI can be used to augment disinformation campaigns, and it is becoming clear that in order for them to do so, in the case of GPT-3 at least, they need a human to provide narrative bullet points. So, gathering these bullet points and ensuring their effectiveness is paramount. The same study mentioned in this paragraph also underscored the value of narratives that are rooted in truth, draw upon fears, or “tap into deeply rooted values.” (Sedova et al. 2021)

Finally, at least through this exploratory case study, the framework of analysis provided by Truth, Lies and Automation stands up to scrutiny. While having little reason to doubt it in the first place, it does seem to provide some valuable insights into how GPT-3 could augment disinformation campaigns in Ukraine. The tasks outlined in the framework also highlight that that for each campaign, a different set of tasks may be more fitting. This may be of some use to mitigation tactics, as certain tasks are more suited to certain platforms such as telegram for short-form text and articles for longer form.

## 6. Topics for Further Research

This research highlights the usefulness of narrative bullet points. What it does not do, however, is go into depth on how those narrative bulletpoints are created. How difficult would it really be? To answer a question like this, it would be more fitting to create an empirical study that set out to measure how much time was needed, how much reconnaissance, in order to create a set of usable narrative bulletpoints.

However, I hesitate to go much further into speculation without addressing some of the ethical concerns of research like this. Indeed, this work will likely be submitted publicly and there may be concern that it reads like instructions for creating AI-enabled disinformation. While, with time, I believe that mitigation strategies will overpower AI-enabled disinformation campaigns, with the current state of affairs (seemingly still confused about definitions), it may not be a step in the right direction to publicize such research. I considered using narrative bullet points to create my own posts for social media like Telegram. However, I think this should be done in a more controlled environment and with more time given to potential ramifications and ethical concerns. However, it does seem that with very limited involvement, I could have created (with the help of GPT-3) more effective disinformation than much of what I witnessed during data collection.

Also during data collection, it became apparent that disinformation around the two case studies in this dissertation seemed to exist in an echo chamber. One telegram page linked to another in a sort of web of disinformation. This is something that could be mapped and analyzed. The assumption here would be, if one could identify one page in a disinformation echo chamber that can then be moderated, can the entire echo chamber be shut down?

Finally, it would be valuable to understand exactly how effective localization is. I merely speculate that having localized messages will lead to more effective disinformation, but as far as empirical results coming from an experiment, this might be valuable to know. Again, I think there are ethical concerns with creating a study like this which essentially explains either how to create more effective disinformation, or extra steps that can be cut out in order to save time while creating disinformation.

## 7. Summary

This study seeks to understand how AI, and the natural language processor GPT-3 specifically, can augment disinformation campaigns. With many studies focusing on mitigation strategies, it stands to reason that first a more detailed understanding of how AI can be used offensively is needed. This study attempts to explore this question through an exploratory study of GPT-3.

The literature review uncovers a number of gaps. Firstly is the confusion around disinformation, misinformation and fake news. Disinformation is intentional, which misinformation unwitting and fake news has become too broad a term to assign in a study such as this. Disinformation campaigns and influence operations also have some small but significant differences, with influence operations encompassing a much larger set of adversarial actions. Additionally, the term “narrative bullet point” is introduced to explain a single point within a disinformation narrative that can be used as a prompt to significantly increase the effectiveness of GPT-3. The literature review continues with exploring some of the frameworks recently developed for analyzing and examining how AI, and not only natural language processors, may be used to augment disinformation campaigns. Finally, disinformation in Ukraine specifically is discussed.

The framework for analysis draws from “Truth, Lies and Automation,” a study that focuses on how GPT-3’s usage in disinformation campaigns can be broken down into tasks. These six types are narrative reiteration, narrative elaboration, narrative manipulation, narrative seeding, narrative wedging, and narrative persuasion. The framework section of the study also elaborates more on the other frameworks mentioned in the literature and explains why they were not chosen for this particular study. This is



largely because their focus is too broad, either focusing on a process or other types of AI than natural language processors.

The methodology is an exploratory case study of two disinformation campaigns which are titled Nazism in Ukraine and the Bucha Massacre. A case study is perhaps the only fitting methodology given the restraints of this study. First, disinformation campaigns, especially the ones chosen, are still unfolding. Second, case studies are well-fitted to “how” questions, such as the research question proposed. Finally, it is difficult to perform any type of detailed experiment as I have little control over events.

Data is taken from two primary sources - Telegram and news articles, with a focus on Telegram. Both of these work well with the structure provided in Truth, Lies and Automation. In the case studies, it becomes apparent that the narrative bullet points would be a valuable tool in creating disinformation.

The research concludes with several findings. One, many of the disinformation examined was not localized for English audiences, even if the disinformation was aimed specifically at English-speaking audiences. If the propagandists had used a model like GPT-3, this would have been much less of an issue. Second, the prevalent concept of narrative bullet points supports the assumption that AI will not be able to act alone, rather a human-machine team working together. Third, fine-tuning, a technique to increase the effectiveness of general AI models, is probably not needed in the case of GPT-3. Instead, in-context learning should suffice. Next, there is discussion around a reconnaissance phase before the creation of a narrative. While it doesn't answer exactly how GPT-3 itself is used, it does shed some light on the preparation for using GPT-3. It may not be as simple as booting up GPT-3, adding narratives and narrative bullet points in prompts, and spreading disinformation. But, the reconnaissance phase is beyond the scope of this dissertation as it likely requires a deep investigation as not every disinformation campaign will need the

same amount of reconnaissance. Finally, though the framework is limited to tasks, the tasks do seem to ring true in the case of disinformation in Ukraine. This may mean that the same tasks can be applied to disinformation campaigns in general to better understand how NLPs can affect disinformation campaigns.

## References

1. Andreikovets, K. (2022) At least 458 Ukrainians died in the Bucha community as a result of the actions of the Russians, War crimes in Bucha - 458 dead were found. Available at: <https://babel.ua/en/news/82626-at-least-458-ukrainians-died-in-the-bucha-community-as-a-result-of-the-actions-of-the-russians> (Accessed: April 30, 2023).
2. BBC News. (2022, April 11). Bucha killings: Satellite image of bodies site contradicts Russian claims. BBC News. Available at: <https://www.bbc.com/news/60981238> (Accessed: April 20, 2023)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp.1877-1901. Available at: <https://dl.acm.org/doi/abs/10.5555/3495724.3495883> (Accessed: April 30, 2023)
4. Buchanan, B, Lohn A, Musser M., Sedova K. (May, 2021) Truth, Lies and Automation. Available at: <https://cset.georgetown.edu/publication/truth-lies-and-automation/>. (Accessed: 27 April 2023)
5. Bump, P. (2022) Analysis | The platform where the right-wing bubble is least likely to pop, Washington Post. Available at: <https://www.washingtonpost.com/politics/2022/04/23/telegram-platform-right-wing/> (Accessed: April 30, 2023).
6. Edson C. Tandoc Jr., Zheng Wei Lim & Richard Ling (2018) Defining “Fake News”, *Digital Journalism*, 6:2, 137-153, Available at: DOI: 10.1080/21670811.2017.1360143 (Accessed: 27 April 2023)
7. Fallis, D. (2020) “The epistemic threat of deepfakes,” *Philosophy & Technology*, 34(4), pp. 623–643. Available at: <https://doi.org/10.1007/s13347-020-00419-2>. (Accessed: April 30, 2023)
8. Gelfert, A. (2018) “Fake news: A definition,” *Informal Logic*, 38(1), pp. 84–117. Available at: <https://doi.org/10.22329/il.v38i1.5068>. (Accessed: 27 April 2023)
9. Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Available at: ArXiv, abs/2301.04246. (Accessed: April 27, 2023)
10. Herasimenka, A. et al. (2022) “Misinformation and professional news on largely unmoderated platforms: The case of telegram,” *Journal of Information Technology & Politics*, 20(2), pp. 198–212. Available at:

<https://doi.org/10.1080/19331681.2022.2076272>. (Accessed: April 30, 2023)

11. InfoDefenseEnglish (2023) "Special operation" by the AFU: Militants killed a civilian, filmed on a UAV and proudly publish the footage [Telegram] Available at: <https://t.me/infodefENGLAND/6529> (Accessed: 1 May, 2023)

12. InfoDefenseEnglish [Telegram] Available at: [t.me/infodefENGLAND](https://t.me/infodefENGLAND) (<https://t.me/infodefENGLAND>) (Accessed: 27 April 2023)

13. Kertysova, K. (2018). Artificial Intelligence and Disinformation. Security and Human Rights 29, 1-4, 55-81, Available at: Brill <https://doi.org/10.1163/18750230-02901005> (Accessed 30 April 2023)

14. Larson, Eric V., Richard E., Darilek, Daniel Gibran, Brian Nichiporuk, Amy Richardson, Lowell H. Schwartz, and Cathryn Quantic Thurston. (2009) Foundations of Effective Influence Operations: A Framework for Enhancing Army Capabilities. Santa Monica, CA: RAND Corporation, 2009. Available at: <https://www.rand.org/pubs/monographs/MG654.html>. (Accessed: 27 April 2023)

15. Lipp, Alina. [Telegram] Available at: <https://t.me/neuesausrussland> (Accessed: 27 April 2023)

16. Liu, H. Disinformation. In Encyclopedia. (2022, October 14). Available at: <https://encyclopedia.pub/entry/29164> (Accessed: 27 April 2023).

17. Media, social media, and disinformation in Ukraine (2022) European Eye on Radicalization. Available at: <https://eeradicalization.com/media-social-media-and-disinformation-in-ukraine/> (Accessed: April 30, 2023).

18. Merriam-Webster. (n.d.). Narrative. In Merriam-Webster.com dictionary. Available at: <https://www.merriam-webster.com/dictionary/narrative> (Accessed: 27 April 2023)

19. MFA Russia (2022), Statement by the Russian Defense Ministry... April 3 [Telegram], Available at: <https://t.me/MFARussia/12230> (April 3, 2022) (Accessed: 27 April 2023)

20. OECD. (2022, November 3). Disinformation and Russia's war of aggression against Ukraine. [online] Available at: <https://www.oecd.org/ukraine-hub/policy-responses/disinformation-and-russia-s-war-of-aggression-against-ukraine-37186bde/#contact-d4e6699>. (Accessed: 27 April 2023)

21. OHCHR (2022) Killings of civilians: Summary executions and attacks on individual civilians in Kyiv, Chernihiv, and Sumy regions in the context of the Russian Federation's Armed Attack Against Ukraine, OHCHR. Available at: <https://www.ohchr.org/en/documents/country-reports/killings-civilians-summary-execution-s-and-attacks-individual-civilians> (Accessed: April 30, 2023).

22. Pawelec, M. (2022) “Deepfakes and democracy (theory): How Synthetic Audio-Visual Media for disinformation and hate speech threaten core Democratic functions,” *Digital Society*, 1(2). Available at: <https://doi.org/10.1007/s44206-022-00010-6>. (Accessed: April 30, 2023)
23. Roache, M., Tewa, S., Cadier, A., Labbe, C., Padovese, V., Schmid, R., O’Reilly, E., Richter, M., König, K., Sadeghi, M., Vercellone, C., Fishman, Z., Adams, N., Pavidonis, V., Walid, S., Griffin, K., Palmer, C., Slomka, A., Vallee, L., Kapoor, A., Maitland, E., Wang, M. and Palmer, K. (2023) *Russia-Ukraine Disinformation Tracking Center - NewsGuard, NewsGuard*. Available at: <https://www.newsguardtech.com/special-reports/russian-disinformation-tracking-center> (Accessed: April 30, 2023).
24. Rogers, R. (2020) “Deplatforming: Following extreme internet celebrities to telegram and alternative social media,” *European Journal of Communication*, 35(3), pp. 213–229. Available at: <https://doi.org/10.1177/0267323120922066>. (Accessed: April 30, 2023)
26. Rong, F. 2021. *Extrapolating to Unnatural Language Processing with GPT-3’s In-context Learning: The Good, the Bad, and the Mysterious*. Available at: <http://ai.stanford.edu/blog/in-context-learning/> (Accessed: April 27, 2023).
27. Rossoliński-Liebe, G. (2015) “The Fascist kernel of Ukrainian genocidal nationalism,” *The Carl Beck Papers in Russian and East European Studies*, (2402). Available at: <https://doi.org/10.5195/cbp.2015.204>. (Accessed: April 30, 2023)
28. Rossoliński-Liebe, G. and Willems, B. (2022) “Putin’s abuse of history: Ukrainian ‘nazis’, ‘genocide’, and a fake threat scenario,” *The Journal of Slavic Military Studies*, 35(1), pp. 1–10. Available at: <https://doi.org/10.1080/13518046.2022.2058179>. (Accessed: April 30, 2023)
29. Sedova, K., Mcneill, C., Johnson, A., Joshi, A. and Wulkan, I. 2021. *AI and the Future of Disinformation Campaigns - Center for Security and Emerging Technology*. Available at: <https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/> (Accessed: 27 April 2023).
30. Simpson, Sean (2019, June 11). *CIGI-Ipsos Global Survey: Only one in four internet users believes the world is better off as a result of social media*. Available at: <https://www.ipsos.com/en-us/news-polls/cigi-fake-news-global-epidemic> (Accessed: April 30, 2023)

31. Sly, L. (2022) Accounting of bodies in Bucha nears completion, Washington Post. The Washington Post. Available at: <https://web.archive.org/web/20220809120452/https://www.washingtonpost.com/world/2022/08/08/ukraine-bucha-bodies/> (Accessed: April 30, 2023).
32. Smart, C. (2022) How the Russian media spread false claims about Ukrainian nazis, The New York Times. The New York Times. Available at: <https://www.nytimes.com/interactive/2022/07/02/world/europe/ukraine-nazis-russia-media.html> (Accessed: April 30, 2023).
33. Staff, T.I. (2022) Spike seen in Russian media reports hammering false claim of Ukrainian nazism, The Times of Israel. Available at: <https://www.timesofisrael.com/spike-seen-in-russian-media-reports-hammering-false-claim-of-ukrainian-nazism/> (Accessed: April 30, 2023).
34. Team of Authors. (2022). Disinformation about Ukraine in Russian and pro-Russian Telegram channels | Democracy Reporting International. Available at: <https://democracy-reporting.org/en/office/ukraine/news/disinformation-about-ukraine-in-russian-and-pro-russian-telegram-channels> (Accessed: April 30, 2023)
35. Thornhill, J. (2022) Commentary: Telegram’s lenience on disinformation has made it a valuable tool in the Ukraine war, CNA. Available at: <https://www.channelnewsasia.com/commentary/telegram-lax-security-disinformation-spread-information-channels-connect-users-2591046> (Accessed: April 22, 2023).
36. UKR LEAKS\_eng. [Telegram] Available at: t.me/ukr\_leaks\_eng (Accessed: 27 April 2023)
37. Vaccari, C. and Chadwick, A. (2020) “Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news,” *Social Media + Society*, 6(1). Available at: <https://doi.org/10.1177/2056305120903408>. (Accessed: April 30, 2023)
38. van Prooijen, J.-W. and Douglas, K.M. (2018). Belief in conspiracy theories: Basic principles of an emerging research domain. *European Journal of Social Psychology*, [online] 48(7), pp.897–908. Available at: DOI:<https://doi.org/10.1002/ejsp.2530>. (Accessed: 27 April 2023)
39. Villasenor, J. (2020) How to deal with AI-enabled disinformation, Brookings. Available at: <https://www.brookings.edu/research/how-to-deal-with-ai-enabled-disinformation/>. (Accessed: 27 April 2023)
40. Wang, C.-C. (2020). Fake News and Related Concepts: Definitions and Recent Research Development. *Contemporary Management Research*, 16(3), 145-174. Available at: <https://doi.org/10.7903/cmr.20677> (Accessed: 27 April 2023)

41. What is a disinformation campaign? (2020, November 5). PREVENCY®. Available at: <https://preveny.com/en/what-is-a-disinformation-campaign/> (Accessed: 27 April 2023)
42. Yin, R. (2009) Case study research: Design and methods. (4th Ed.). Thousand Oaks, CA, USA: Thousand Oaks, CA: Sage. Available at: doi:<https://doi.org/10.33524/cjar.v14i1.73>. (Accessed April 20, 2023).
43. Андреев. (2022, April 4). Расстрел на улицах: Что на самом деле произошло в Буче. Life.ru. Available at: <https://life.ru/p/1484302> (Accessed April 20, 2023).
44. Они же преступники: как украинские беженцы добивают европу (2023) Baltnews. Baltnews<https://baltnews.com/i/logo/ru.png>. Available at: <https://baltnews.com/v-mire/20230416/1025938998/Oni-zhe-prestupniki-kak-ukrainskie-bezhentsy-dobivayut-Evropu.html> (Accessed: April 30, 2023).
45. СОЛОВЬЁВ. (2022) В Буче готовится новая провокация [Telegram] Available at: <https://t.me/SolovievLive/98859> (Accessed: 1 May, 2023)

## List of Appendices

Appendix no. 1: List of disinformation campaigns in Ukraine (Image) (Figure 1)

Roache, et al. (2023) Common disinformation on over 220 pro-Russian websites. Russia-Ukraine Disinformation Tracking Center - NewsGuard, NewsGuard. Available at: <https://www.newsguardtech.com/special-reports/russian-disinformation-tracking-center> (Accessed: April 30, 2023).

The following list compiles some of the most common myths and disinformation from more than 220 websites with a history of publishing false, pro-Russia propaganda and disinformation.

Classified documents showing Ukraine was preparing an offensive operation against the Donbas

The massacre of civilians in Bucha, Ukraine, during the first month of the war was staged

The United States is developing bioweapons designed to target ethnic Russians and has a network of bioweapons labs in Eastern Europe

Ukraine threatened Russia with invasion

US paratroopers have landed in Ukraine

Ukraine staged the attack on the hospital in Mariupol on 9 March 2022

European universities are expelling Russian students

Ukraine is training child soldiers

The war in Ukraine is a hoax

Russia was not using cluster munitions during its military operation in Ukraine

NATO has a military base in Odessa

Russia does not target civilian infrastructure in Ukraine

Modern Ukraine was entirely created by communist Russia

Crimea joined Russia legally

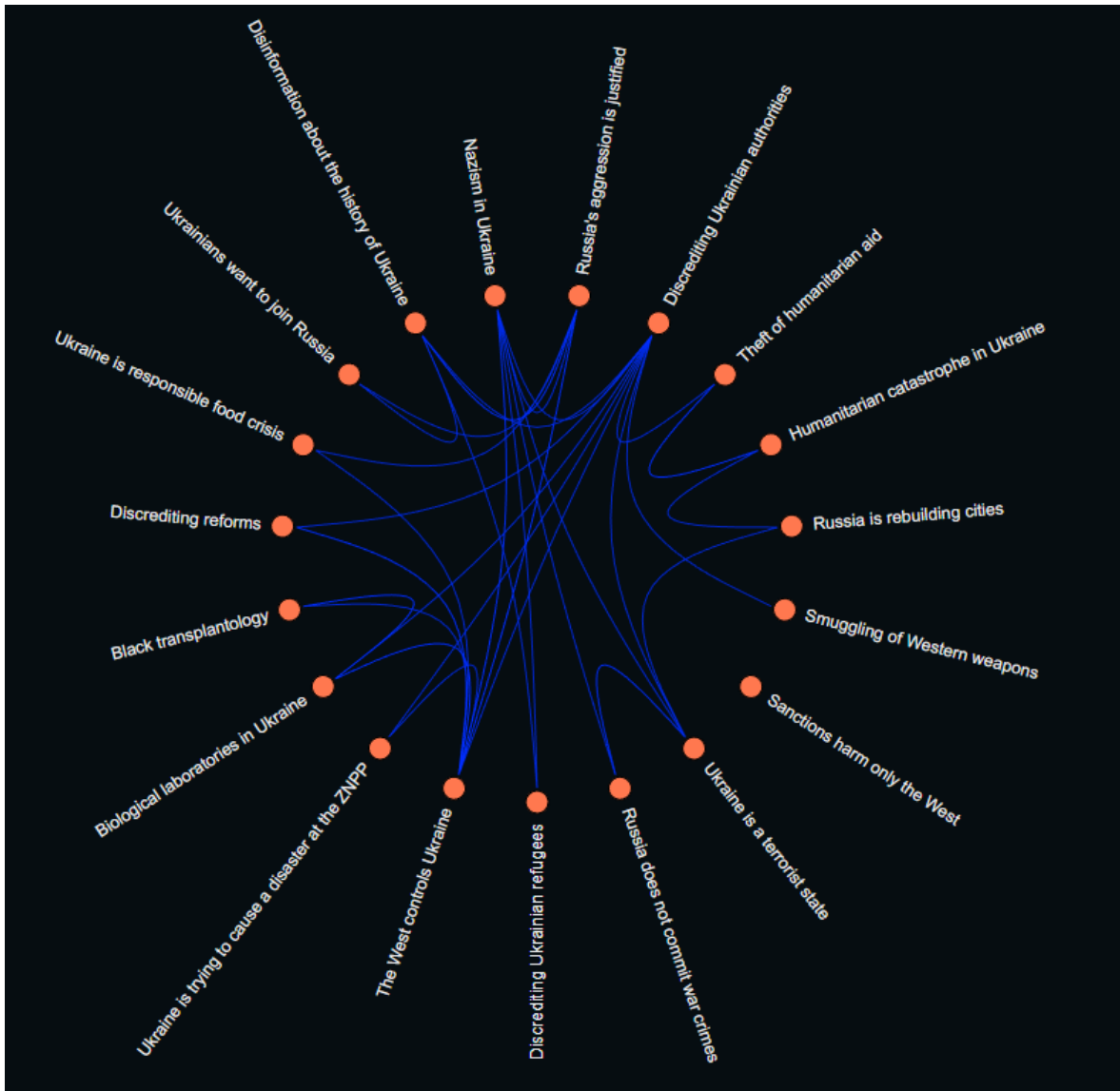
Ukrainian forces bombed a kindergarten in Lugansk on Feb. 17, 2022

The United States and the United Kingdom sent outdated and obsolete weapons to Ukraine



Appendix no. 2: Existing narratives in Ukraine in 2022 (Graph) (Figure 2)

Team of Authors. (2022). Disinformation about Ukraine in Russian and pro-Russian Telegram channels | Democracy Reporting International. Available at: <https://democracy-reporting.org/en/office/ukraine/news/disinformation-about-ukraine-in-russian-and-pro-russian-telegram-channels> (Accessed: April 30, 2023)



Appendix no. 3: Table of GPT-3 tasks (Table) (Figure 1)

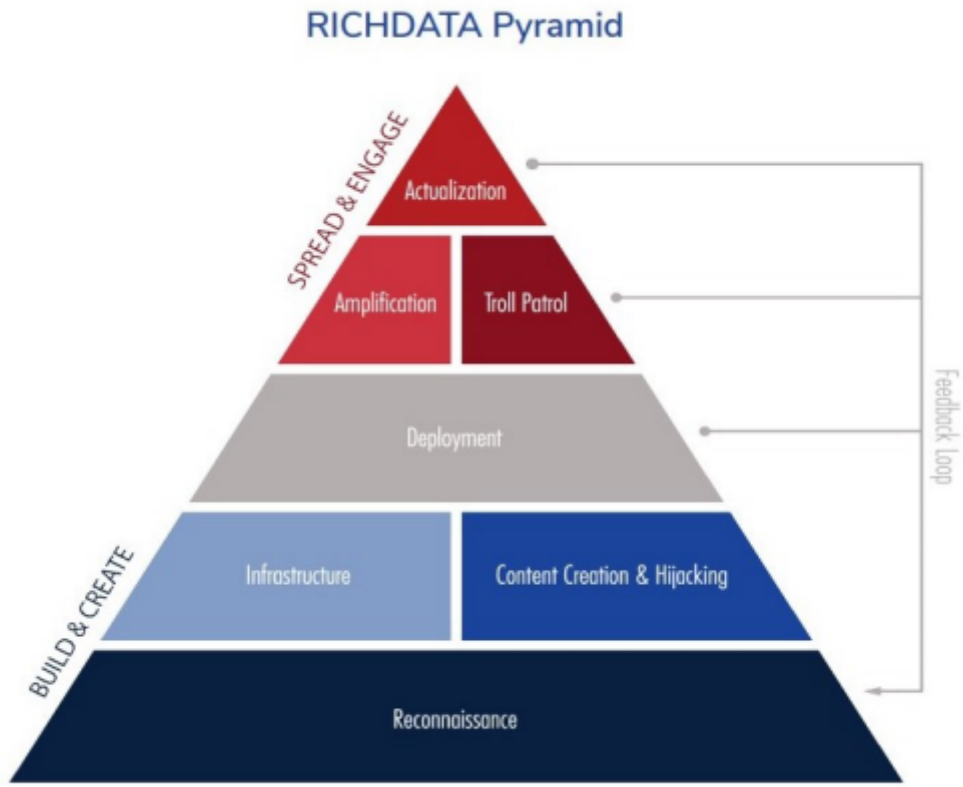
Buchanan, B, Lohn A, Musser M., Sedova K. (2021) Truth, Lies and Automation. Available at: <https://cset.georgetown.edu/publication/truth-lies-and-automation/>. (Accessed: 27 April 2023)

TABLE 1  
Summary evaluations of GPT-3 performance on six disinformation-related tasks.

TASK	DESCRIPTION	PERFORMANCE
Narrative Reiteration	Generating varied short messages that advance a particular theme, such as climate change denial.	GPT-3 excels with little human involvement.
Narrative Elaboration	Developing a medium-length story that fits within a desired worldview when given only a short prompt, such as a headline.	GPT-3 performs well, and technical fine-tuning leads to consistent performance.
Narrative Manipulation	Rewriting news articles from a new perspective, shifting the tone, worldview, and conclusion to match an intended theme.	GPT-3 performs reasonably well with little human intervention or oversight, though our study was small.
Narrative Seeding	Devising new narratives that could form the basis of conspiracy theories, such as QAnon.	GPT-3 easily mimics the writing style of QAnon and could likely do the same for other conspiracy theories; it is unclear how potential followers would respond.
Narrative Wedging	Targeting members of particular groups, often based on demographic characteristics such as race and religion, with messages designed to prompt certain actions or to amplify divisions.	A human-machine team is able to craft credible targeted messages in just minutes. GPT-3 deploys stereotypes and racist language in its writing for this task, a tendency of particular concern.
Narrative Persuasion	Changing the views of targets, in some cases by crafting messages tailored to their political ideology or affiliation.	A human-machine team is able to devise messages on two international issues—withdrawal from Afghanistan and sanctions on China—that prompt survey respondents to change their positions; for example, after seeing five short messages written by GPT-3 and selected by humans, the percentage of survey respondents opposed to sanctions on China doubled.

Appendix no. 4: Pyramid graph of disinformation campaigns (Graph) (Figure 4)

1. Sedova, K., Mcneill, C., Johnson, A., Joshi, A. and Wulkan, I. 2021. AI and the Future of Disinformation Campaigns - Center for Security and Emerging Technology. Available at: <https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns/> (Accessed: 27 April 2023).



Source: CSET.

Appendix no. 5: Stages of AI-Enabled Influence Operations (Image) (Figure 5)

Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Available at: ArXiv, abs/2301.04246. (Accessed: April 27, 2023)

