Disinformation using artificial intelligence technologies – a key component of Russian hybrid warfare

Ina VIRTOSU,
PhD in EU Law, University of Macau, SAR Macau, China
yb67199@connect.um.edu.mo, ivirtosu3@gmail.com

Mihai GOIAN,
PhD Student, National School of Political and Administrative Studies SNSPA, Romania
mihai.goian.22@drd.snspa.ro

Abstract

For over a decade, the Russian Federation has been at the forefront of hybrid warfare, deploying highly developed capabilities to carry out cyberattacks, political and social subversion, the exploitation of societal tensions, corrupt financial influence, and disinformation operations campaigns worldwide. The primary objective of Russia's disinformation campaign is to distort public opinion formation, undermine confidence in institutions, denigrate political leadership, widen social divide, and meddle with election processes where it has a certain level of interest. The Russian invasion of Ukraine has brought with it a deluge of disinformation and misinformation. Although it became routine for major events to generate fabricated news, images, and video, the war in Ukraine is featuring a whole new set of characteristics and the extensive use of machine learning that will influence how readers react to fake media. Fast advancements in information technology, particularly the use of artificial intelligence (AI), have changed the methods in which information and disinformation may be created and disseminated, making more difficult to differentiate falsehood from real information. In such tensioned environment, websites born to churn out misinformation that operate purely for spreading Kremlin propaganda have launched a real campaign to defend Kremlin right to war. This research paper explores the use of AI in the context of Russian hybrid war, such as user profiling, bot farms, micro-targeting, and deep fakes. The article also examines the ways in which AI can be used to counter such disinformation online and reviews a number of solutions that could help address the spread of AI-powered disinformation for improving the online environment.

Keywords: social media, artificial intelligence, algorithms, fact-check, deep fakes.

1. Introduction

Months before Russian soldiers invaded Ukraine on February 24, 2022, false narratives about Ukraine and its allies began to circulate online, many of which were supported by the Kremlin's disinformation organization. These and hundreds of other charges, ranging from bogus claims of Ukrainian genocide intended towards Russian-speaking Ukrainians to statements that Nazi ideology drives Ukraine's political leadership, have been used to justify Russia's right to war and full-scale invasion of Ukraine.

Only NewsGuard alone has debunked more than 100 false narratives in only one year related to the Russia-Ukraine war and identified more than 350 sites spreading those myths [1]. While most myths disavow Russia's alleged atrocities and other abuses in Ukraine or demonize Ukrainians, NewsGuard has also debunked some pro-Ukrainian and anti-Russian myths, ranging from manipulated images of the mythical Ghost of Kyiv to misleading footage of alleged Russian attacks [2].

A prevalent concern is that modern warfare methods, like as large-scale Russian propaganda efforts, are being utilised to mould the narrative around the war, despite the fact that relevant research is still in its infancy. On the one hand, the Russian government implemented new legislation that gave it control over conventional media sources in order to encourage its citizens to support the war. As a result, domestic news sources are compelled to follow the government narrative [3]. On the other hand, Russian propaganda has been suspected to influence other countries outside Russia, in particular, by using AI, and social media to promote hostility against the West.

AI and its subcomponents, such as algorithms and machine learning, are powerful instruments for creating and disseminating disinformation about the Russia-Ukraine conflict, particularly on social media. Even before the war, there have been extremely provocative reports concerning the emergence of a “weaponized AI propaganda machine”.

2. The concept of hybrid warfare

2.1. Definition and characteristics of hybrid warfare

Hybrid warfare is a concept that describes the use of a combination of military and non-military tactics to achieve political and military objectives [4]. This concept has gained popularity in the context of contemporary conflicts, where actors use their available resources and tools to influence adversaries and promote their own strategic interests. In a hybrid war, military tactics such as special operations, espionage, special forces, and asymmetric operations are combined with non-military methods, such as propaganda, disinformation, manipulation of public opinion, cyber-attacks, and subversion [4]. This combination of diverse and flexible actions makes hybrid warfare difficult to identify and counter by the involved adversaries.

Russia has been one of the key players involved in hybrid warfare. Russia’s strategic objectives in this context include weakening adversaries, destabilizing neighboring regions, and expanding their geopolitical influence [5]. To achieve these objectives, Russia uses a wide range of tactics and strategies in hybrid warfare. Among the tactics used by Russia are propaganda and disinformation. These aim to manipulate public opinion by creating and promoting false and manipulative narratives [6].Propaganda and disinformation are disseminated through state-controlled media, online platforms, and social networks, with the aim of influencing and undermining the position of adversaries. In addition, Russia uses cyber-attacks and hacking operations to distort and manipulate information, as well as to compromise the digital systems and infrastructure of other states [7]. These attacks can target government institutions, non-governmental organizations, private companies, and even critical infrastructure. Russia also uses subversive tactics to destabilize adversaries. These tactics include supporting insurgent and terrorist groups, organizing protests and civil unrest, and supporting separatism in targeted regions [4].

2.2. Disinformation as a key component of hybrid warfare

In the context of hybrid warfare, disinformation plays a crucial role in the strategies used by unfriendly states or actors to achieve their political and military objectives. Disinformation refers to the deliberate dissemination of false or manipulative information, with the aim of creating confusion, influencing public opinion, and undermining trust in legitimate information sources [8] [7].

It is important to understand the definition and characteristics of disinformation in the context of hybrid warfare, as they help us to recognize and counter the disinformation strategies used in this context [8]. By recognizing and understanding how disinformation is used to manipulate public opinion, we can develop effective strategies and tactics to counter this phenomenon [8].

Key characteristics of disinformation in the context of hybrid warfare include:

a) *Information manipulation.* Disinformation involves manipulating and distorting information to support the agenda or interests of the promoter. This can include creating false narratives, intentional omissions, distorted interpretations, or selectively presenting information to create a deceptive perception.

b) *Use of unidentified or false sources.* Disinformation in the context of hybrid warfare may involve the use of unidentified or false sources to support and promote manipulative information. This can make it difficult to verify and authenticate information and can create confusion among the public.

c) *Amplification and rapid spread through technology*. Modern technologies, such as social networks and online platforms, facilitate the rapid and massive spread of disinformation. They enable the propagation of false or manipulative information to a large number of people in a short time, thereby creating a significant impact on public opinion.

d) *Targeting key audiences.* Disinformation in the context of hybrid warfare is often directed towards certain key audiences, such as ethnic communities, interest groups, or specific socio-demographic categories. By tailoring the disinformation messages for these audiences, it aims to influence their perceptions and attitudes in a specific way.

e) *Creating chaos and instability*. Disinformation in the context of hybrid warfare often aims to create chaos, instability, and confusion in the target society. By spreading false and manipulative information, it seeks to undermine trust in institutions and fuel existing tensions and divisions in society.

It is also important to be aware of the impact disinformation can have on society and national security. The propagation of false information and the manipulation of public opinion can weaken social stability, create tensions and divisions, and undermine trust in institutions and democratic processes. Therefore, combating and countering disinformation in hybrid warfare must be a priority for governments, organizations, and civil society.

3. Pillars and composition of Russia disinformation and propaganda ecosystem

3.1. Pillars of Russian disinformation and propaganda

Russia has been using the whole playbook of information manipulation and interference, including disinformation, in an attempt to sow divisions in the societies, denigrate democratic processes and institutions and rally support for its imperialist policies. Russia’s full-scale invasion of Ukraine on February 24, 2022 has shown, again, the wide spectrum of tactics, techniques and behaviour (TTPs) used in the information environment, while building mostly on well-known disinformation narratives.

Ukraine has been the first target of Russia’s foreign information manipulation and interference. The invasion is the culmination of Russia's years-long propaganda campaign and involvement aimed at undermining Ukraine’s sovereignty and territorial integrity. Almost all of the misinformation narratives employed by the Kremlin to explain and mobilise domestic support for the invasion can be traced back to 2013-2014, and the Euromaidan protests, during which the Kremlin attempted to depict Ukraine as a “Nazi state”, a “failed state”, and “not a state at all”. For years, pro-Kremlin media has been setting ground for a military invasion. According to EU foreign policy chief Josep Borrell, “this war is not only conducted on the battlefield by the soldiers, but also waged in the information space trying to win the hearts and minds of the people”, the EU agencies having plenty of evidence that Russia is behind coordinated attempts to manipulate public debates in open societies [9].

There are several pillars of disinformation and propaganda used by Russia to promote its agenda and influence public opinion. These pillars were identified in a report by the U.S. Department of State and include the following:

a) *Strategic disinformation.* Russia utilizes disinformation as a strategic tool to distort facts and create an alternate reality that supports its political and geostrategic objectives (Russia's Pillars of Disinformation and Propaganda Report, 2019). This is a complex process that involves creating false narratives and manipulating information in a coordinated manner.

b) *Emotional Manipulation*. Russia employs techniques of emotional manipulation to influence and evoke intense reactions among the public. This includes the use of shocking imagery, inflammatory speeches, and content that evokes fear, anger, or outrage (Russia's Pillars of Disinformation and Propaganda Report, 2019). The aim is to provoke strong emotional reactions that can impact individuals' judgment and critical reasoning.

c) *Creating diversions*. Russia uses disinformation and propaganda to create diversions and divert public attention from real issues or its own actions. This may include promoting conspiracy theories or spreading false information to create confusion and distract attention from important subjects (Russia’s Pillars of Disinformation and Propaganda Report, 2019).

d) *Manipulation of the information environment*. Russia employs a variety of tactics to manipulate the information environment, including controlling and influencing the mass media, cyber attacks on websites and information platforms, as well as the use of bots and online trolls to amplify disinformation content (Russia's Pillars of Disinformation and Propaganda Report, 2019). These tactics contribute to the spread and amplification of manipulative and disinformation messages.

According to a report of the European External Actions Service, Russia uses a multitude of tactics to control narratives in its disinformation campaigns, which can be broadly divided in so-called 5D tactics:

a) Dismiss tactics is used to push back against criticism, deny allegations and denigrate the source;

b) Distort tactics to change the framing and twist and change the narrative;

c) Distract tactics to turn attention to a different actor or narrative or to shift the blame;

d) Dismay tactics to threaten and scare of opponents;
e) Divide tactics to create conflict and widen divisions within or between communities and groups [10].

According to the same report, in the case of incidents carried out by channels linked to Russia, 42% were intended to distract [10]. The large majority of incidents was used in the context of the Russian invasion of Ukraine, to turn attention to a different actor/narrative or to shift the blame (namely to Ukraine and the EU). Another 35% aimed to distort, twist and frame narratives around the Russian invasion of Ukraine and to deliver attacks against the Ukrainian government and EU officials and institutions (such as the HR/VP) [10]. All incidents related to the energy crisis were also linked to these two objectives. Russia used the divide objective in incidents highlighting the West's alleged Russophobia or promoting Russian worldwide influence in order to create conflict and widen divisions within or between communities and groups. Top-targeted entities by these incidents were the government of Kosovo (in the context of the tensions with Serbia) and Poland [10].

3.2. Disinformation methods used by Russia

3.2.1. Creation and promotion of false narratives

The use of informational technologies in disinformation often involves the creation and promotion of false narratives. Russia uses this tactic to manipulate public opinion and undermine adversaries. By creating false narratives, Russia tries to create and promote distorted versions of events and influence public perception [7]. These narratives can be propagated through state-controlled media, social media platforms, and websites.

A notorious example of using false narratives is Russia's intervention in the conflict in Ukraine. Russia has created and propagated false narratives that present Russia as a "protector" of the Russian-speaking population in Ukraine and claim that Ukraine is governed by fascists [5]. These narratives have resulted in an increase in tensions between Russia and Ukraine and have contributed to the escalation of the conflict.

3.2.2. Using bots and online trolls to amplify disinformation

Russia often uses bots and online trolls to amplify disinformation and influence public discourse. Bots are computer programs that mimic human behavior on online platforms, and trolls are individuals who deliberately post and promote false or manipulative information [11]. These bots and trolls are used to create the illusion that certain opinions or narratives are more popular and supported than they actually are. They can disseminate disinformative content en masse and generate a large volume of activity on social media platforms, making false information seem credible and have a greater impact on the public [12]. A notorious example of the use of bots and online trolls is Russia's intervention in the 2016 US presidential election. Russia used bots and trolls to amplify disinformative messages and influence public opinion, thus creating an atmosphere of disinformation and confusion during the election campaign [6].

3.2.3. Cyberattacks and hacks to distort and manipulate information

Russia uses cyberattacks and hacks to distort and manipulate information. These attacks can target the systems and digital infrastructure of other states, government institutions, and non-governmental organizations, as well as individuals. Through cyberattacks, Russia can gain unauthorized access to sensitive information and modify or distort them for disinformative purposes [7]. It can also compromise systems and digital infrastructure to create chaos and disrupt the normal functioning of institutions and organizations. A significant example is the cyberattack on the Democratic National Committee (DNC) in the US in 2016, attributed to Russia. Through this cybercrime, the party's confidential information was compromised and revealed to the public, having a significant impact on the election campaign and the party's image [8].

3.2.4. Use of personal information and targeted disinformation to influence public opinion

Russia uses personal information and targeted disinformation to influence public opinion and manipulate individuals. By collecting and analyzing personal data, Russia can gain a deeper understanding of individual interests and preferences and can tailor disinformative messages to attract and manipulate them. Also, through targeted disinformation, Russia can create and distribute personalized content to specific groups of people. This approach allows for more efficient and precise influence over these groups, increasing the chances of achieving the desired results. A notorious example is the use of personal data stolen during the cyberattack on Facebook by Cambridge Analytica, a political consulting firm associated with Russia [13]. This data was used to create psychological profiles of users and deliver personalized messages to influence voting and manipulate public opinion in elections.

In conclusion, Russia uses a variety of methods in disinformation. These methods have a significant impact on public opinion, national security, and electoral processes, contributing to the spread of disinformation and undermining trust in information and institutions.

3.3. State and non-state Russian actors involved in disinformation

According to a report published by the Global Engagement Center (GEC) at the US Department of State, Russian propaganda ecosystem is composed by different networks of online and offline channels attributed and/or connected to the official infrastructure of the government or ruling party [14]. The array of disinformation channels is very diverse and the degree of connections with Kremlin range from visible to obscure and denied. These online channels, including websites, YouTube channels, groups, and profiles on social media, have been attributed according to high-confidence level indicators and can be classified in several groups:

1. Official communication channels.
    a) Official Government communication channels officially used by Kremlin and ministries, agencies, and its representatives to deliver statements, which includes official websites of a state or social media accounts;
    b) State-controlled media channels with an official affiliation to a state-actor. They are majority-owned by a state or ruling party, managed by government appointed bodies and they follow an editorial line imposed by state authoritie;
    c) Statement or quotes used by Russian officials.

2. State-linked global channels.
    a) State funded foreign audience media, such as Sputnik Mundo or RT Arabic;
    b) State funded national audience media, such as Pervîi Canal, RIA Novosti or Rossiya 24;
    c) State funded, foreign based media, such as Novosti Moldova or Newsmaker.md;
    d) International Russian socio-cultural institutions, such as different NGOs, think tanks and foundations that either it has funded by the government or Russian-government-associated oligarchs to spread false messages and propaganda.
3. channels facing with no transparent links nor an official affiliation to a state actor, but their attribution has been confirmed by organisations with access to privileged backend data sources, such as digital platforms, intelligence and cyber security entities, or by governments or military services based on classified information. These channels are also called proxy sources.
    a) Russia aligned outlet with global reach, such as Global Reach and News Front;
    b) Local language specific outlet, such as Compact Magazine (Germany);
    c) Witting proliferators of Russian narrative;
    d) Unwitting proliferators of Russian narrative;
    e) Foreign state narrative amplification.
    f) Weaponization of social media.
    g) Infiltration of domestic conversation;
    h) Standing campaign to undermine faith in institutions;
    i) Amplification of protests or civil discord.
    j) Cyber enabled disinformation
    k) Hack and release;
    l) Site capture;
    m) Cloned website;
    n) Forgeries;
    o) Disruption of official sources or objective media [15].

This ecosystem strategy is also well-suited to promote Russia’s overarching goals of undermining democratic institutions and undermining the international legitimacy and cohesiveness of other democratic countries. Because some pillars of this ecosystem generate their own momentum rather than relying on specific orders from the Kremlin on every occasion, they can be responsive to specific policy goals or developing situations, and then pivot back to their status quo of generally slamming Russia’s perceived adversaries.

Internet Research Agency (IRA) is a well-known Russian company engaged in online influence operations on behalf of Russian business and political interests. According to the Mueller Report released in 2017 [15], the IRA was using social media to subvert the political process and is at the core of the federal prosecution. They have been gathering sensitive information on thousands of Americans, such as names, addresses, phone numbers, email addresses, and other important data. The IRA campaign during the 2016

US presidential election was deeply sophisticated and mainly aimed at instigating the social media users against each other on socio-economic grounds, political sentiments, and voting perceptions [16].

As was stated previously, disinformation in the context of hybrid warfare is not limited to spreading false and manipulative information but also involves cyber activities and hackers supporting these disinformation efforts. Russia has a rich history of utilizing hacker groups to advance its political agenda and achieve strategic objectives. One notable example is the hacker group known as Fancy Bear or APT28. This group is believed to be associated with Russian intelligence agencies and is accused of numerous cyber attacks and disinformation operations. Fancy Bear has been involved in various incidents, including attacks on foreign government organizations and institutions, as well as hacking campaigns and cyber espionage [17].

Another hacker group with ties to Russia in Cozy Bear or APT29. This group has been associated with Russia's Foreign Intelligence Service (SVR) and is involved in sophisticated cyber attacks and global espionage operations. Cozy Bear has been identified as responsible for cyber attacks on government organizations, companies, and institutions in the United States, Europe, and other parts of the world [17].

These Russian hacker groups have played a significant role in supporting disinformation efforts through unauthorized acquisition of information, infiltration and manipulation of information systems, as well as the spread of false or manipulative content. Their activities are often coordinated and supported by state institutions, posing a serious threat to cybersecurity and informational integrity. Monitoring and countering Russian hacker groups are essential components in the efforts to combat disinformation in hybrid warfare. Enhancing cybersecurity capabilities, international collaboration in detecting and responding to cyber attacks, and implementing appropriate preventive measures are critical aspects in protecting infrastructure and sensitive information against these threats.

Pro-Kremlin outlets have also been instrumental in justifying and obfuscating war crimes and atrocities committed by Russian soldiers in Ukraine. Further enhanced by the Russian losses on the battlefield, hate speech and incitement to genocide became a regular occurrence in Russian outlets, both offline and online. Narratives supporting the war permeate not just political life and news, but also entertainment content. Print and TV media are the most frequent targets of Moscow’s impersonation, in particular when targeting Ukraine. Despite Russian state-controlled media being banned by some social platforms, such as Meta, Russian disinformation operations have continued on social networks with thousands of unsophisticated accounts flooding the online environment with Russian views on the invasion. They also can use the news anchor deep fakes on platforms including Facebook, Twitter and YouTube while monitoring disinformation operations that the research firm has dubbed “spamouflage”.

Spamouflage refers to an extensive network of Moscow linked accounts that disseminate pro-Russia propaganda. After Russia invaded Ukraine in February 2022, the EU moved to block RT and Sputnik, two of the Kremlin’s top channels for spreading propaganda and misinformation about the war. Nearly six months later, the number of sites pushing that same content has exploded as Russia found ways to evade the ban. They have rebranded their work to disguise it. They have shifted some propaganda duties to diplomats. Indeed, Russia's diplomatic corps serves as a worldwide propaganda network, with hundreds of social media profiles on every continent, where the same statements may be rehashed and adjusted for various audiences in different countries.

Kremlin propaganda machine also has cut and pasted much of the content on new websites – ones that until now had no obvious ties to Russia. Some of the sites pose as independent think tanks or news outlets. About half are English language, while others are in French, German or Italian. Many were set up long before the war and were not obviously tied to the Russian government until they suddenly began parroting Kremlin talking points. For instance, Yala News claims to offer impartial news, but BBC analysis has shown most of its content directly mirrors stories on Russian state-backed media sites – and that it actually operates out of Syria [18].

According to Crovitz, NewsGuard co-CEO, Russian propaganda actors may establish sleeper sites, which are created for a disinformation campaign that lay largely dormant, slowly building an audience through innocuous or unrelated posts, and then switching to propaganda or disinformation at an appointed time [19]. The proliferation of sites spreading disinformation about the war in Ukraine shows that Russia had a plan in case governments or tech companies tried to restrict RT and Sputnik. Russian influence operations use third parties' news companies and their sites, which acts as a "Kremlin loudspeaker" in different parts of the world. These can be noticed by analysing the timeframe and similarities in their stories. This process is called "information laundering" and it is defined as the dissemination of news, whether true or misleading, from unconfirmed sources into the mainstream [20]. Meleshevich and Schafer compare such process with money laundering by stating that in advancing disinformation in such a way that makes it accepted as ostensibly legitimate information into ostensibly legitimate funds [21]. This process is done through occurrence of three phases: placement, layering and integration [21].

The initial publication of false material on a website or social media account is referred to as placement. Disinformation operations, like financial criminals, rely on specific sorts of social media accounts that may distribute information in a way that conceals both its goal and its source. Social media accounts offer disinformation campaigns free access to a platform from which information can be spread, and in many cases the public account ownership need not have any link to the true owner [21].

Layering describes how misinformation flows from one source to other plausible sources, acquiring credibility through reposts, likes, and shares. The phrase refers to the use of intermediate firms, banks, and people to transmit money across borders and through various types of financial instruments in order to break the link between origin and beneficiary. As proven by recent law enforcement instances highlighting how intermediaries transfer cash previously deposited into the financial system in order to offer more distance from the source, third-party money launderers play a critical role in successful money laundering schemes [21]. Layering has two forms in misinformation efforts. The first is the employment of middlemen who appear to have no connection to the source of the material. The second sort of stacking occurs via indirect citations (also known as cascading citations) from unverified social media posts to seemingly reputable news sites.

Integration refers to the point at which purposely deceptive material is embraced by credible news sources or extensively distributed by genuine people on social media platforms. For misinformation peddlers, success translates to influence, and extensive political and social impact occurs when falsehood is woven into public discourse. Disinformation that has been successfully assimilated is difficult to spot. Once a false rumour enters the mainstream, it is nearly tough to disprove, even if it is later discredited [21]. To achieve their goal, propagandists just need to penetrate the permeable layer between doubtful and legitimate news sites – whether the content itself bears up to deeper investigation is practically immaterial.

Russia's ability to successfully conduct hybrid warfare is predicated on the creation of a fog of ambiguity between the Kremlin's actions and the Kremlin itself. By conducting operations through an ad hoc network of proxies, carve-outs, and cutouts – whose connection to the Kremlin is difficult to definitively establish – the Russian government is able to maintain plausible deniability and thus lower the diplomatic and military costs of its actions.

4. The use of information technology in disinformation

4.1. Impact of technology on the spread of misinformation

It has been already demonstrated that online fake news spreads much more quickly and more widely than real news [22]. According to Susarla, online posts with fake information get more views, comments and likes than those with accurate content [23]. In an online when viewers' attention spans are limited and material options abound, it appears that false information is more interesting or engaging to viewers. The problem is getting worse: nowadays are more fake news than real information. This could bring about a phenomenon that researchers have dubbed "reality vertigo" – in which computers can generate such convincing content that regular people may have a hard time figuring out what's true anymore [23].

Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and contrary to conventional wisdom, robots accelerated the spread of true and false news at the same rate, implying that false news spreads more than the truth because humans, not robots, are more likely to spread it. Advancements with generative AI tools have sparked concerns about the technology's capacity to create and disseminate disinformation at an unprecedented scale. These technological advancements come with concerns that Russia are taking to try to control narratives about everything from foreign policy to the war in Ukraine using AI. The skill and efficiency with which AI can generate disinformation is particularly worrying as is more difficult to detect and make difference between fake and real news.

Thus, the use of information technologies in disinformation has broad-reaching societal effects, which can be followings:

a) *Weakening public trust in information sources and democratic institutions.* The use of information technology in disinformation significantly impacts public trust in information sources and democratic institutions. Propaganda and disinformation, fueled by information technology, undermine trust in the media, government organizations, and other institutions essential to a democratic society. For instance, studies show that disinformation and information manipulation on social networks and online platforms have led to a decrease in public trust in traditional media and government institutions [24] [25]. This severely affects the functioning of democracy, as trust in democratic information and institutions is crucial for informed decision-making and civic participation.

b) *Dividing and polarizing society through propagation of conflicts and tensions.* The use of information technology in disinformation contributes to dividing and polarizing society by propagating conflicts and tensions. Targeted disinformation and information manipulation can fuel ideological, ethnic, political, and social disputes and conflicts. Studies show that the use of information technology in disinformation on social networks and online platforms has led to an increase in polarization and tensions among the population [24] [25]. This is because targeted disinformation and information manipulation contribute to the creation of information bubbles and the promotion of extremist beliefs and perspectives among users. This phenomenon can lead to increased divisions and tensions in society, weakening social cohesion and affecting the ability to reach consensus and act collectively to address common problems.

c) *Undermining electoral processes and influencing results.* The use of information technology in disinformation can undermine electoral processes and influence their outcomes. Through targeted disinformation and propaganda, Russia and other actors can manipulate public opinion and affect voters' decisions. Recent studies and investigations have highlighted the influence of disinformation on electoral processes [6]. For example, in the 2016 US presidential elections, disinformation and the spread of false content significantly impacted public opinion and the election outcome. Russia was accused of using disinformation, bots, and online trolls to amplify disinformation messages and influence public opinion in favor of certain candidates [25]. This type of intervention in electoral processes undermines democratic integrity and can compromise the representativeness and legitimacy of the results.

d) *Increasing confusion and uncertainty in society.* The use of information technology in disinformation contributes to increasing confusion and uncertainty in society. Information manipulation and disinformation spread can lead to the emergence of conspiracy theories, multiple versions of events, and misinterpretations of reality. This creates a climate of uncertainty and difficulty in determining the truth. People are exposed to a wide variety of information, and disinformation and information manipulation make the process of verifying and evaluating these more difficult [24]. The increase in confusion and uncertainty can have serious consequences for society, as people may be influenced to make uninformed decisions, adopt extreme attitudes, or fail to act appropriately in the face of real threats.

e) *Impact on national security and international relations.* The use of information technology in disinformation also has an impact on national security and international relations. By propagating disinformation and manipulating information, Russia and other states can create tensions and conflicts between states and undermine international trust and cooperation. Disinformation can be used to fuel anti-Western, anti-EU, or anti-NATO sentiments in certain countries, which can weaken existing security alliances and relations. It can also be used to justify aggressive actions or military interventions, undermining regional stability and security. In international relations, the use of information technology in disinformation can create tensions and conflicts betweenstates, leading to the deterioration of diplomatic relations and possibly even escalation of hostilities. An example of this was seen in the conflict between Ukraine and Russia, where disinformation played a significant role in exacerbating tensions and conflicts [26].

f) *Manipulating public opinion through mass media and social networks.* The use of information technology in disinformation also involves manipulating public opinion through mass media and social networks. Russia and other states utilize media channels to disseminate false and manipulative information strategically, with the aim of influencing public perceptions and attitudes [27]. This manipulation of public opinion can have significant consequences in society. It can create divisions, fuel tensions, and contribute to the polarization of public opinion. Furthermore, it can undermine public trust in the media and presented information, thus affecting the functioning of democracy and the process of informed decision-making.

5. Russia’s information war is being waged mostly on social media platforms

Throughout the Russia-Ukraine war, concerns regarding misinformation have risen across a range of social platforms. The variety of social media platforms in use, as well as the differences in their availability between countries, makes it difficult to coordinate efforts to counteract disinformation while generating various information ecosystems across different geographic territories. The narratives about the ongoing war that are developing on social media vary based on the platform and location, including within Russia and Ukraine. Although Facebook and Twitter are both banned within Russia, Russian propaganda and misinformation directed at external audiences continues to flourish and expand on both platforms. YouTube and TikTok are still available to ordinary residents in Russia, but with significant censorship. Nowadays the most popular social media platform used within Russia is VKontakte (VK), which hosts 90% of internet users in Russia, followed by Odnoklasniki (ok.ru) [28]. It was previously available and widely used in Ukraine until 2017, when the Ukrainian government blocked access to VK and other Russian social media in an effort to combat online Russian propaganda [29]. Also, on September 26, 2022, the VK application (as well as other applications of the holding services) was removed from the Apple App Store due to international sanctions.

Across Telegram – a social media network that is now a battleground between pro-Ukrainian and pro-Russian camps – channels with tens of thousands of subscribers have posted grainy war footage of alleged atrocities or of military vehicles heading toward Kyiv, though much of this content was either taken out of context or reused from previous conflicts. Nowadays, Telegram is the main social media platform accessible to both Russians and Ukrainians. Telegram is an encrypted messaging app that is being used in the war for everything from linking Ukrainian refugees to safe passage options to giving near-real-time footage of military events. However, Telegram has no official policies to censor or remove content of any nature for fighting disinformation. While some channels on Telegram have been shut down, the company releases official statements only occasionally and it generally allows the majority of content posted by users to continue circulating, regardless of its nature [30]. The company claims that it cannot act against individual or group chats, which are "private amongst their participants", however it can respond to requests in relation to sticker sets, channels and bots which are publicly available.

Telegram may be a fairly marginal social media channel in the West, but -unlike Twitter, Facebook, and YouTube -it is one free of restrictions for state-backed propaganda campaigns in Russia, where it remains popular. The Russian state broadcaster RT, for example, has more than 200,000 followers on the platform. Nevertheless, Telegram barred Kremlin-backed media outlets from using its platform within the EU, including RT, since Europe's sanctions require such propaganda to be removed from television broadcasts, video-sharing platforms, internet service providers and other digital networks in an effort to prevent disinformation peddled by Moscow from reaching a large, mostly online, audience.

As the invasion began a handful of channels such as "Donbass Insider" and "Bellum Acta" started pumping out pro-Russian propaganda. Within minutes of explosions being reported in Donetsk, Odessa, and Kyiv, the channels supplied details, images, and video of the war in real time, in Russian, English, Spanish, and French. They showed Russian soldiers heading to war and the missiles landing just outside major Ukrainian cities. On other Telegram channels that trade in far-right memes, images were shared of Putin brandishing a handgun and promising to "crush those filthy Ukrainian", earning heart emoji from followers. As result, the Ukrainian National Security and Defense Council issued a statement labeling which accounts are Russian backed. Ukrainian officials, in potential violation of the Geneva Convention, also have shared imagery of dead and captured Russian soldiers on the platform.

On February 27, 2022, Telegram CEO, Pavel Durov, posted that channels were becoming a source of unverified information and that the company lacks the ability to check on their veracity. He urged users to be mistrustful of the things shared on channels, and initially threatened to block the feature in the countries involved for the length of the war, however, he walks back this plan when it became clear that they had also become a vital communications tool for Ukrainian officials and citizens to help coordinate their resistance and evacuations. Telegram can thus serve as a mostly unfiltered source of disinformation within Russia and Ukraine, reaching audiences that Western social media platforms aren't able to reach. While Telegram does not filter content like many other platforms, it also does not use an algorithm to boost certain posts, and it relies on direct messaging between users. This design makes it difficult for AI tools to effectively boost disinformation.

In contrast, on other platforms such as Twitter and Facebook, AI is further enabling the rapid spread of disinformation about the war. Even before the war, there was much debate over how these platforms prioritized and monitored material on a wide range of political and social themes. In recent years, regulators in the US and the EU [31] have criticized Facebook and YouTube for prioritizing extremist material and failing to appropriately remove disinformation despite some changes to automated and human-led mechanisms. Similar concerns have arisen across a range of platforms. For example, according to an analysis by anti-misinformation site NewsGuard, a new TikTok account may be presented lies about the Ukraine war within minutes of signing up for the app [32]. The organization, which assesses the credibility of news providers across the web, conducted two tests to see how the video-sharing app handled conflict-related content. It discovered that a new account that did nothing but scroll around the app's algorithmically curated For You Page watching war movies would be directed to inaccurate or misleading material within 40 minutes. The streams were nearly entirely filled with both true and fraudulent news on the Ukrainian war, with little difference made between disinformation and genuine sources. None of those videos contained any information about the trustworthiness of the source, warnings, fact-checks, or additional information that could empower users with reliable information. Users on TikTok were shown videos claiming that a "photoshopped" press conference given by Vladimir Putin in March 2020, false allegations about US bioweapon facilities in Ukraine, movies purporting to show the "Ghost of Kyiv" shooting down Russian jets were stolen from a video game, while real footage from the battle was slammed as fake by pro-Russian accounts. TikTok is also abundant with multiple Russian-language influencers parroted the same script in defense of Russia's invasion.

Although, Facebook claims that is continuously monitoring and take down separate multipronged disinformation operations, its algorithms routinely promoted disinformation about the war, including the conspiracy theory that the US is funding bioweapons in Ukraine. A study by the Center for Countering Digital Hate (CCDH) found that Facebook failed to label 80 % of posts spreading this conspiracy theory about US-funded bioweapons as disinformation [33].

Days after Russia invaded Ukraine, multiple social media platforms – including Facebook, Twitter and YouTube – announced they had dismantled coordinated networks of accounts spreading disinformation. Yet as Google, Facebook, Twitter and TikTok actively removed, or demoted, content associated with Moscow, new strategies have begun to bubble to the surface. These networks, which were comprised of fabricated accounts disguised with fake names and AI-generated profile images or hacked accounts, were sharing suspiciously similar anti-Ukraine talking points, suggesting they were being controlled by centralized sources linked to Russia. The Russian government creates accounts on Facebook, Telegram, Twitter, and other social networks that pretend to be organizations, phony social justice initiatives, anti-immigrant activists, and ordinary persons. Account types range from hired trolls using false online identities to bots and cyborg accounts, which are human-operated and employ automation to enhance their messaging. Account identities on Twitter were linked to the Kremlin-funded Internet Research Agency.

According to a research, such accounts can use generic Western names, a purported political leaning, an alleged affiliation with a news organization, a cultural reference, including fandom, or a single name followed by a string of numbers that sequentially change from one account to the next [30]. Many of these accounts post similar or almost identical content, frequently using bots engineered to boost and promote certain ideas.

Digital platforms are especially vulnerable to information laundering because faked videos (deepfakes) and images (photograph manipulation), for instance, can create media moments and spread disinformation bots, fake accounts and click farms “pretend to be people they’re not and create a false sense of consensus” and commercial platforms “designed to keep users online to be served ads, end up prioritising engagement over truth or the public interest” [34].

6. The role of social AI in Russian disinformation and propaganda campaigns

AI has the potential to be used primarily for creating photo, audio, and video fakes, as well as for bot farms. AI can replace a significant part of the personnel in Russian “troll factories,” Internet warriors who provoke conflicts on social networks and create the illusion of mass support for Kremlin narratives by users. Instead of trolls who write comments according to certain guides, this can be done by AI using keywords and the vocabulary offered to it. The influencers mentioned above (politicians, propagandists, bloggers, conspiracy theorists, etc.) have a decisive influence on the loyal audience, and not nameless bots and Internet trolls. However, with the help of AI, the weight of the latter can be increased by quantitative growth and “fine-tuning” for different target audiences.

If fake news served as the foundation for this new automated political propaganda and disinformation machine, social bots, or fake social media profiles, functioned as its foot soldiers – an army of political robots employed to manage social media discussions, intimidate, and misinform people.

A social bot, also described as a social AI, is a software agent that communicates autonomously on social media [35]. The messages that a social bot distributes can be simple and operate in groups and various configurations in hybrid mode, with partial human control via algorithm. Using AI and machine learning, social bots may think and act like people on social media sites, including expressing thoughts in more natural human discourse. While certain bots, such as auto-moderators and chat bots, are meant to provide better service/management, they may be abused by extremist organisations [36].

The challenge is that AI-powered bots can also engage in many harmful actions, such as escalating disputes, inflate influence and promote extreme ideologies performing scams and disseminating fake news. Such bots post using a fake account, with photo, posts, and a good number of followers or so-called friends. Thus, the account is created only to distribute its disinformation messages or political statements. This can be done via likes, sharing and retweets or in the form of posts or comments. Using a programming interface (API), a social bot can access social networks and receive and send data.

The presence of pro-Russian bots on social media is not a new phenomenon. The best-known case concerns Internet Research Agency nicknamed “troll factory”, that in 2016 influenced the US presidential election by favoring candidate Donald Trump in the interest of the Kremlin. In 2018, Twitter found that the phenomenon involved over 50,000 Russia-linked bot accounts [30]. The Oxford Internet Institute analised how Twitter bots were used during the Brexit debate, and it was found that while many were used to spread messages about the Leave campaign, the vast majority of the automated accounts were very simple [37]. They were made to alter online conversation with bots that had been built simply to boost likes and follows, to spread links, to game trends, or to troll opposition. According to Woolley it was controlled by small groups of human users who acknowledged the power of memes and virality, of spreading conspiracy theories online and watching them proliferate [38]. Simple bot-generated spam interrupted conversations by being purposefully attached to key hashtags in order to demobilise online debates. Links to news articles that portrayed a politician negatively were pushed by phoney or proxy accounts set up to publish and republish the same garbage over and again. These advertisements were employed bluntly: these bots were never meant to be conversational. Political bots play a key role in disinformation as smart AI tools allowed computers to pose as humans and help manipulate public conversation. On March 4, 2022, Twitter banned about 100 accounts that had relaunched the hashtag "#IstandwithPutin" for participating in "coordinated inauthentic behavior." However, the data we hold suggests that the phenomenon in Europe affects a much larger number of accounts, which avoided such a ban. European Digital Media Observatory (EDMO) established that since the beginning of the war in Ukraine, on February 24, 2022, a large number of accounts, whose main goal was to spread pro-Russian disinformation, were detected on Twitter [39]. Many of these profiles were suspected to be bots, but a large part could also be managed by actual human beings that act coordinately to spread false or misleading narratives about the conflict. Disinformation about the war in Ukraine started circulating in Europe immediately after Russia invaded the country and in a few weeks the topic became extremely popular among conspiracy theorists and their followers. Russian bots on Twitter typically operate by following and retweeting accounts that support their agenda, as well as posting their own content. This content can include false or misleading information, inflammatory language, and divisive political messages. The goal of these bots is to create the appearance of widespread support for a particular political view or candidate and to manipulate public opinion in their favour.

The use of bots on Twitter allows Russia to amplify its messages and create the appearance of widespread support for its views. Russian bots may give the impression that there is a huge and loud group of people who share their ideas by following and retweeting accounts that support their agenda and by uploading their own material. This has the potential to influence public opinion and shift the political environment in their favour [40]. In addition, the use of bots on Twitter allows Russia to bypass traditional forms of media and communicate directly with the public. By using Twitter, they can reach a large audience quickly and easily without needing to go through the editorial process of traditional news outlets. This allows them to spread their messages quickly and effectively, and to avoid being held accountable for the accuracy or veracity of their content.

A research from January 2023, realised by the analytical centre NewsGuard discovered that the popular chatbot ChatGPT is capable of generating texts that extend current conspiracy theories and integrate genuine events in their context [41]. This programme has the ability to automate the delivery of numerous messages (via bot farms), the topic and tone of which will be selected by a person and the direct text created by AI. NewsGuard analysts directed the chatbot to reply to a series of leading inquiries pertaining to a sample of 100 misleading narratives from NewsGuard's proprietary database of 1,131 top disinformation narratives in the press and their debunkings, published before 2022 [41]. The findings support concerns, especially those highlighted by OpenAI, about how the technology may be weaponized in the wrong hands. ChatGPT developed fake narratives for 80 of the 100 previously detected false narratives, including complete news items, essays, and TV scripts [41]. For those who are inexperienced with the issues or themes addressed, the results may appear legitimate, even authoritative. For 80% of the prompts, ChatGPT provided answers that could have appeared on the worst fringe conspiracy websites or been advanced on social media by Russian government bots.

Another analysis regarding ChatGPT and its use in spreading of Russian propaganda was done by Centre for Democracy and Rule of Law. According to their research, although ChatGPT is not a propaganda platform, it displays information based on what data it has "learned" [42]. If the model is trained on texts that contain propaganda, there is a risk that ChatGPT will provide misleading and distorted information. After all, it will be relying on a data set it learned. The Centre discovered that ChatGPT provides incorrect information in its answers. For instance, "several cases of nuclear weapons use". Obviously, ChatGPT discovered references to nuclear weapons on the Internet but misinterpreted them [42]. ChatGPT is an interesting artificial intelligence model, however it lacks the capacity to filter out information more effectively. Thus, on the one hand, ChatGPT simplifies the work of journalists, writers and everyone involved in content creation, but on the other hand, it can become a generator of false news and distorted facts. To reduce the amount of disinformation in ChatGPT, developers have many ways to do this: a) Improve the Russian propaganda identification system with trigger words; b) Program ChatGPT to refuse to generate texts that could potentially contain Russian propaganda or assessment of the Russian-Ukrainian war; c) Integrate a system into ChatGPT that checks the generated text for the presence of pro-Russian narratives. It should be remembered that ChatGPT is only an algorithm that learns from the texts of human authors but it is not its creator. That is why the texts received from ChatGPT should be treated just as critically as anything else you read on the Internet.

7. Deepfakes and their role in disinformation

The 21st century's answer to photoshopping, deepfakes, use a form of AI called deep learning to make images of fake events. For decades it has been possible to alter video footage, however doing it took time, highly skilled artists, and a lot of money. Deepfake technology changed the game. As it develops and proliferates, anyone could have the ability to make a convincing fake video, including some those who might seek to "weaponize" it for political or hybrid war purposes.

Deepfakes are only possible because of recent breakthroughs in machine learning. Driven by the widespread availability of multimodal data, such as news articles, social media, audio, imagery, and video, as well as the dramatic reduction in costs of high-performance central processing unit (CPU) and graphics processing unit (GPU) computing clusters, machine learning techniques are now ubiquitous [43].

Deepfakes are defined as synthetic media that have been digitally manipulated to replace one person's likeness convincingly with that of another. The term describes both the technology and the resulting bogus content and is a portmanteau of deep learning and fake.

The main machine learning methods used to create deepfakes are based on deep learning and involve training generative neural network architectures, such as autoencoders, or generative adversarial networks (GANs) [42]. For computer vision, or the field of AI that enables computers to interpret and react to visual images, a GAN consists of two key components: a generator algorithm that tries to generate a fake image and a discriminator algorithm that tries to distinguish between real and fake images.

Audio can be deepfaked too, to create “voice skins” or “voice clones” of public figures. Deepfake technology also can create convincing but entirely fictional photos from scratch. Russia has been already creating AI – generated personas with full profiles and a human face. NBC News journalist Ben Collins discovered two specific people who are spreading disinformation from the city of Kyiv [44]. But not everything is what it seems, both of these profiles are not recognized by any system as real people. As it turns out, they were both created by a Russian troll farm in order to spread fake news about Kyiv. The first one Collins introduced is Vladimir Bondarenko, a blogger from Kyiv who despises the Ukrainian Government. Watching his artificially created face is downright scary when you see how real his picture is. On the Ukraine Today website, Vladimir has an antire backstory as if he was a real human being. He studied to become an aviation engineer, but he was later forced to become a blogger when the Ukraine aviation infrastructure collapsed. Russia also created an AI profile of a woman, Irina Kerimova from Kharkiv. She used to be a private guitar teacher, but she eventually became chief editor of this Russia propaganda website that is presumably founded by the RT company (the Kremlin) [44]. She also has a strage mismatch on her earrings. Facebook revealed to Collins that these two profiles are part of Russia's new propaganda operation that was identified by the State Department back in 2020. They are called News Front and South Front and were both created by Alexander Malkevich, the same man who ran the St. Petersburg troll farm after 2016.

On March 2, 2022, shortly after Russia started its full-scale invasion of neighboring Ukraine, a video message featuring Ukrainian President Volodymyr Zelensky emerged temporarily on the news website Ukraine 24 [45]. Dressed in his iconic olive shirt, Zelensky’s tone and outfit mirrored his other statements of that period. However, the message itself was rather different: rather than encouraging Ukrainians to continue fighting, Zelensky urged them to lay down their arms and surrender. The video then shortly spread on VKontakte, Telegram, and other social media sites, where it was picked up and reported on by international media. 2 Zelensky’s office instantly denied its validity, pointing out that it was the type of “deepfake” they had warned of before the conflict. Nonetheless, the episode was a watershed moment in information operations since it was the first high-profile use of a deepfake during an armed conflict.

Deceit and media manipulation have always been a part of wartime communications, but never before it has been possible for nearly any actor in a conflict to generate realistic audio, video, and text of their opponent’s political officials and military leaders. As AI grows more sophisticated and the cost of computing continues to drop, the challenge deepfakes poses to online information environments will only grow. Policymakers and government officials will need to develop robust systems for monitoring and authenticating both public and private messages in real time, while also evaluating when – if at all – to leverage the technology themselves. Deepfakes can be leveraged for a wide range of purposes, including falsifying orders from military leaders, sowing confusion among the public and armed forces, and lending legitimacy to wars and uprisings. While these tactics can and often will fail, their potential to impact an adversary's communications and messaging means that security and intelligence officials will inevitably use them in a wide range of operations.

Beyond deepfakes, experts have expressed concern that AI could be leveraged for more sophisticated disinformation techniques. These include using AI to better identify targets for disinformation campaigns, as well as using techniques such as Natural Language Processing (NLP), which allows AI to produce fake social media posts, articles, and documents that are nearly indistinguishable from those by human posters. It gets harder to make difference from real video as the technology improves. In 2018, US researchers discovered that deepfake faces do not blink normally [8]. No surprise there: most images show people with their eyes open, so the algorithms never really learn about blinking. At first, it seemed like a silver bullet for the detection problem. But no sooner had the research been published, than deepfakes appeared with blinking. Such is the nature of the game: as soon as a weakness is revealed, it is fixed.

For policymakers and officials in democratic states, deepfakes pose a particularly difficult challenge. Given the importance of a trusted information environment to democratic societies, democratic governments should generally be wary of deepfakes, which threaten to undermine that trust. This is particularly true when it comes to military and intelligence operations. Going forward, militaries and security services will need to assume that rival state and nonstate actors alike will have access to deepfake capabilities that can generate compelling audio and video of any state official, leader, or soldier. As a result, they will need to develop the kind of robust authentication mechanisms and “pre-bunking” strategies that Ukraine has already pioneered. Moreover, they will need to understand how deepfake technology adds further complexity to the communications challenges that militaries and insurgent groups already face. Democratic governments will need to develop strategies for how to operate in such an environment without undermining the integrity of their communications or key values and norms. Almost every day, neural networks display an advance in their ability to create graphic, textual, and audiovisual information. Its quality will improve as machine learning skills improve. Popular neural networks are being utilised by Internet users as a toy rather than a tool for manufacturing fakes. However, there are already examples of neural network-generated pictures that not only went viral but were also considered as real by users. Ironically, AI may be the answer to fight deepfakes. AI already helps to spot fake videos, but many existing detection systems have a serious weakness: they work best for celebrities, because they can train on hours of freely available footage. Tech firms are now working on detection systems that aim to flag up fakes whenever they appear. Another strategy focuses on the provenance of the media. Digital watermarks are not foolproof, but a blockchain online ledger system could hold a tamper-proof record of videos, pictures and audio so their origins and any manipulations can always be checked.

8. Countering disinformation in hybrid warfare

The use of disinformation for political and strategic reasons by countries such as Russia, both domestically and globally, increases instability and poses risks in the framework of hybrid warfare. Disinformation operations are intended to manipulate emotions, spread distrust, and cause chaos in order to sway public and political opinion. To avoid and minimize the consequences of disinformation there are necessary the following measures:

a) *Media education and literacy, raising awareness and public engagement.* Media education and literacy are key tools for countering disinformation. Developing critical skills to evaluate information and identify reliable sources can help people recognize and reject disinformation [46]. Promoting media education in schools and communities can contribute to increasing resilience to information manipulation. Media literacy can help individuals become more aware of the tactics used in disinformation and make informed decisions regarding the consumption and distribution of information. Raising awareness and engaging the public in countering disinformation is essential. Through education, information campaigns, and promoting individual responsibility in information consumption and distribution, the public can become more vigilant and resilient against disinformation [47]. Support and active participation of citizens in identifying and reporting disinformation can contribute to reducing its impact.

b) *Fact-checking and information verification.* Another important strategy is promoting fact-checking and information transparency. Fact-checking organizations play a crucial role in identifying and exposing false or manipulative information [46]. Governments and online platforms should support and collaborate with these organizations to promote verified information and expose disinformation. Additionally, transparency regarding the origin and funding of online content can help identify and better understand the sources of disinformation.

c) *Regulations and transparency in the online environment.* Effective regulation of the online environment is necessary to counter disinformation. Governments should implement policies and regulations that promote transparency and accountability of online platforms in combating disinformation [8] This may include requirements for disclosing content distribution algorithms, clear labeling of verified or unverified content, and sanctions for platforms that enable the widespread dissemination of disinformation [46].

d) Online platforms play a crucial role in combating disinformation. By promoting transparency in algorithmic operations, content selection, and user and data policies, they can enhance user trust and limit the impact of disinformation [24]. They can also enforce stricter measures to prevent and remove disinformation content.

e) *Collaboration among governments, organizations, and online platforms.* Combatting disinformation in hybrid warfare requires also strong international collaboration, among governments, organizations, and online platforms. Governments can implement policies and regulations to address disinformation, while organizations and online platforms can take measures to limit the spread of manipulative content and promote transparency and authenticity of information. Governments, international organizations, and relevant actors should enhance their cooperation to exchange information, identify and counter disinformation campaigns, and promote common norms and standards in the information domain.[6] The exchange of best practices and experiences can contribute to the development of more effective strategies in combating disinformation.

f) *Public-private cooperation and civil society engagement.* Cooperation between the public and private sectors is essential in combating disinformation in hybrid warfare. Governments should work together with online platforms and other private entities to develop technological solutions and share relevant information regarding disinformation campaigns.[6] At the same time, civil society and non-governmental organizations can play a crucial role in monitoring and reporting cases of disinformation, promoting media literacy, and mobilizing public opinion to counter the phenomenon [8].

g) *Monitoring and data analysis.* Monitoring and analyzing data on the spread and impact of disinformation can provide valuable insights for countering disinformation. By monitoring online activities and analyzing patterns of disinformation spread, sources, tactics, and distribution networks can be identified. This information can be used to develop more effective strategies to counter disinformation.

h) *Investments in research and technological development*. Investments in research and technological development can contribute to the development of advanced tools and technologies for identifying and combating disinformation [6]. Artificial intelligence and automated content analysis can be used to detect and filter false or manipulative information, thus helping to reduce their spread in the online environment.

i) *Using AI tools to fight disinformation.* While AI is contributing to the spread of disinformation across social media, AI tools also show promise for combating it. The sheer volume of information uploaded to social media daily makes developing AI tools that can accurately identify and remove disinformation essential. For example, Twitter users upload over 500,000 posts per minute, well beyond what human censors can monitor. Social media platforms are beginning to combine human censors with AI, to monitor false information more effectively. Facebook developed an AI tool called SimSearchNet at the start of the COVID-19 pandemic to identify and remove false posts. SimSearchNet relies on human monitors to first identify false posts, and then uses AI to identify similar posts across the platform. AI tools are significantly more effective than human content moderators alone. According to Facebook, 99.5% of terrorist-related content removals and 98.5 % of fake accounts are identified and removed primarily using AI trained with data from their content-moderation teams.

AI offers enormous potential for content creation and processing. For instance, the Centre for Strategic Communication and Information Security in Ukraine monitors the media landscape and analyses a variety of online articles using AI capabilities, specifically automated tools, such as the SemanticForce and Attack Index platforms. Attack Index uses machine learning, cluster analysis, computer linguistics, formation, clustering, and visualization of semantic networks and correlation and wavelet analysis. SemanticForce, using AI, helps identify information trends, track changes in the response of users of social networks to information events, identify hate speech, analyse in details the image to detect inappropriate or harmful content. Using AI, accessible solutions allow differentiating between natural and coordinated content distribution, detecting automated spam distribution systems, assessing the influence of different social network user accounts on the audience, separating bots from actual users, and so on.

Currently, AI aimed at combatting disinformation on social media still relies on both human and computer elements. This limits AI's ability to detect novel pieces of mis- and disinformation, and means that false posts routinely reach large audiences before they are identified and removed using AI. The current technical limitations on being able to proactively identify and remove false information, combined with the scale of information uploaded online, pose a continuing challenge for limiting disinformation on social media in the Russia-Ukraine war and beyond. Researchers are already planning to employ AI to detect these AI-generated fakes. Techniques for video magnification, for example, can detect variations in human pulse to determine if a person in a video is genuine or computer-generated. However, both fakers and detectors will improve. Some fakes may grow so complex that they are difficult to reject or dismiss, in contrast to previous generations of fakes, which utilised basic language and made readily debunked assertions. The best way to combat the spread of fake news may be to depend on people. Since the societal consequences of fake news are significant, eople shall be more wary of information and investigate it, rather than sharing it immediately.

9. Conclusions

Hybrid warfare represents a complex and flexible approach to conflicts, involving the combination of military and non-military tactics to achieve political and military objectives. Russia, as a key-actor in hybrid warfare, uses a wide range of tactics and strategies to promote its interests and extend its influence.

Disinformation in hybrid warfare poses a serious threat to society, national security, and democratic processes. Its impact is felt in various fields, including politics, public health, and electoral processes. Countering disinformation requires concerted efforts globally, involving governments, organizations, online platforms, and the general public. Education, fact-checking, collaboration, and monitoring are just a few of the strategies used to counter disinformation in hybrid warfare. By effectively addressing this phenomenon, we can protect society and democratic values, promoting authentic and verified communication and information.

The Kremlin tends to employ a full spectrum model of propaganda and disinformation. In some aspects, the present Russian approach to propaganda builds on Soviet Cold War-era methods, with a focus on disinformation and convincing targets to behave in the propagandist's interests without realising it. But from other perspectives, it is entirely new and driven by the peculiarities of today's information world. Russia has utilised technology and accessible media in ways that would have been unthinkable during the Cold War. The Internet, AI, social media, and the expanding landscape of professional and amateur journalism and media outlets are among its tools and channels.

Russian propaganda is generated in enormous volumes and transmitted via a wide range of outlets. This propaganda encompasses text, video, audio, and still imagery distributed over the Internet, social media, satellite television, and traditional radio and television transmission. A significant group of hired Internet "trolls" also frequently attacks or undermines opinions or material that runs opposite to Russian themes. But the most challenging to combat is propaganda built with AI. As AI grows more sophisticated and the cost of computing continues to drop, the challenge social bots and deepfakes pose to online information environments during armed conflict will only grow.

AI is going to be the most powerful tool for distributing disinformation on the Internet, assisting in the creation of a new false narrative on a massive scale and much more frequently. Personalised, real-time chatbots could disseminate conspiracy theories in increasingly believable and compelling ways, smoothing out human errors such as bad grammar and mistranslations and progressing beyond immediately detectable copy-paste operations. To navigate that challenge, security officials and policymakers need a far greater understanding of how the technology works and the myriad ways it can be used in international armed conflict. For decades, machine learning algorithms, a type of AI, have been successful in fighting spam email by analysing message text and determining how likely it is that a particular message is a genuine communication from an actual person or company. However, such solutions imply that individuals who distribute fake information will not modify their tactics. They often change strategies, changing the substance of phony postings to make them appear more trustworthy. The biggest challenge, however, of using AI to detect fake news is that it puts technology in an arms race with itself.

Responding to Russia's hybrid challenge requires a comprehensive strategic strategy that considers every element of hybrid warfare. The goal of any such approach would be to contain Russian hybrid initiatives by limiting and counteracting them. The strategy's fundamental components should be successful defence efforts, resilience in the face of Russian operations, and, if necessary, cost-imposing measures. In practice, combating disinformation and increasing awareness about disinformation may necessitate more strong and well publicised attempts to identify Russian propaganda sources and the nature of their activities. It might also take the form of sanctions, fines, or other barriers to the practice of propaganda disguised as journalism. Another suggestion is "to find ways to help put raincoats on those at whom the firehose of falsehood is being directed" [48]. Of course, a plan to tackle hybrid issues is not the sum of a Western policy for Russia. An overall strategic strategy would incorporate conventional and nuclear deterrence, as well as diplomacy to identify whether areas of possible collaboration exist.

Social media sites must also participate by correctly labelling their content and indicating whether an item claiming to be news has been validated by a reliable source. They could train AI through collaborations with news organisations and volunteers, continually improving the system to respond to propagandists' shifting concerns and techniques. Certainly, it is impossible to detect every piece of fake news released online, but it may make it simpler for a large number of individuals to make difference between true and false news.

References

[1] NewsGuard, “100 Myths: NewsGuard has identified more than 100 false narratives about the war in Ukraine,” 2023.

[2] M. Roache and al, “Russia-Ukraine Disinformation Tracking Center, 457 websites spreading war disinformation and the top myths they publish,,” 2023.

[3] M. Alyukov, “Propaganda, authoritarianism and Russia’s invasion of Ukraine,” *National Human Behaviour,* vol. 6, p. 763–765, 2022.

[4] . B. C. Boatwright and al, “Troll factories: The internet research agency and state-sponsored agenda building, Resource Centre on Media Freedom in Europe.,” 2018.

[5] F. D. Kramer and L. M. Speranza, “Meeting the Russian hybrid challenge: A comprehensive strategic framework,” *Atlantic Council,* 2017.

[6] E. Rumer and R. Sokolsky, “Front Matter. In thirty years of US policy toward Russia: Can the vicious circle be broken?,” *Carnegie Endowment for International Peace,* pp. [i]-[iii], 2019.

[7] C. Wardle, “Fake news. It’s complicated, First Draft,” 2017.

[8] T. Rid, Active Measures: The secret history of disinformation and political warfare, New York: Farrar, Straus and Giroux., 2020.

[9] R. Chesney and C. Citron, “Deepfakes and the new disinformation war: The coming age of post-truth geopolitics,” 2018.

[10] J. Borrell, “Disinformation: Opening speech, by High Representative/Vice-President at the EEAS,” 2022.

[11] E. A. S. European Union, “1st EEAS Report on foreign information manipulation and interference threats towards a framework for networked defence, EEAS.,” 2023.

[12] E. e. a. Ferrara, “The rise of social bots, Communications of the ACM,” vol. 59, no. 7, pp. 96-104, 2016.

[13] P. N. e. a. Howard, “The IRA, social media and political polarization in the United States, 2012–2018,” 2018.

[14] C. Cadwalladr, “Facebook’s role in Brexit – and the threat to democracy,” *TED2019. TED.,* 2019.

[15] US Department of State, “GEC Special Report: August 2020 Pillars of Russia’s Disinformation and Propaganda Ecosystem,” Washington, 2020.

[16] R. Mueller, “Report on the investigation into Russian interference in the 2016 presidential election, vol. 1,” US Department of Justice Washington, DC., 2019.

[17] G. M. Graff, “Russian trolls are still playing both sides – Even with the Mueller probe,” 2018.

[18] CrowdStrike, “Global Threat Report 2016,” 2016.

[19] H. Gelbart, “The UK company spreading Russian fake news to millions, BBC,” 2022. [Online]. Available: https://www.bbc.com/news/world-65150030.

[20] ABC News, “Despite bans, Russian disinformation spreading in new ways,” 2022.

[21] European Parliamentary Research Service, “Automated tackling of disinformation, Panel for the Future of Science and Technology European Science-Media Hub,” 2019.

[22] K. Meleshevich and B. Schafer, “Online information laundering: The role of social media, Alliance for Securing Democracy, Policy brief, no. 2.,” 2018.

[23] S. Vosoughi, S and et al, “The spread of true and false news online, Science,” vol. 359, pp. 1146-1151, 2018.

[24] A. Susarla, “How artificial intelligence can detect – and create – fake news,” The Conversation, 2018.

[25] A. M. Guess and et al, “A digital media literacy intervention increases discernment between mainstream and false news in the United States and India,” vol. 117, no. 27, 2020.

[26] S. Woolley and P. Howard, “Automation, algorithms, and politics. Political communication, computational propaganda, and autonomous agents – Introduction,” *International Journal of Communication,* vol. 10, no. 9, 2016.

[27] K. Giles, “Russia’s new tools for confronting the West continuity and innovation in Moscow’s exercise of power, , .,” *The Royal Institute of Foreign Affairs,* 2016.

[28] CSCE, “Scourge of Russian disinformation,” 2020.

[29] President of Ukraine, “Decree of the President of Ukraine №133/2017: On the decision of the National Security and Defense Council of Ukraine dated April 28, 2017 “On the application of personal special economic and other restrictive measures,” 2017. [Online].

[30] C. N. A. Perez, “Information warfare in Russia’s war in Ukraine,,” 2022.

[31] "Digital Services Act: agreement for a transparent and safe online environment,” Press Releases IMCO, 2022.

[32] A. Cadier and et al, “WarTok: TikTok is feeding war disinformation to new users within minutes – even if they don’t search for Ukraine-related content,,” 2022.

[33] B. Collins and J. L. Kent, “Facebook, Twitter remove disinformation accounts targeting Ukrainians,” NBC News, 2022.

[34] K. Kornbluh and E. Goodman, “To fight online disinformation, reinvigorate media policy,” 2019.

[35] P. Efthimion, S. Payne and N. Proferes, “Supervised machine learning bot detection techniques to identify social Twitter bots,” *SMU Data Science Review,* vol. 1, no. 2, 2018.

[36] N. Hajli and et al, “Social bots and the spread of disinformation in social media: The challenges of artificial intelligence,” *British Journal of Management,* vol. 33, no. 3, 2022.

[37] P. N. Howard and B. Kollanyi, “Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU Referendum,” 2016. [Online]. Available: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2016/06/COMPROP-2016-1.pdf.

[38] S. C. Woolley, “Automating power: Social bot interference in global politics,” *First Monday,* vol. 21, no. 4, 2016.

[39] F. Di Blasi, “A pro-Russian bot network in the EU amplifies disinformation about the war in Ukraine,” *EDMO,* 2022.

[40] G. Allison, “Russian bots on Twitter – the point, UK Defense Journal,” 2023.

[41] NewsGuard, “The next great misinformation superspreader: How ChatGPT could spread toxic misinformation at unprecedented scale, Misinformation Monitor,,” 2023.

[42] O. Petriv , “ChatGPT is a Russian propaganda “consumer”: How do we fight it? Centre for Democracy and Rule of Law,,” 2023.

[43] D. L. Byman and et al, “Deepfakes and international conflict, Foreign Policy,” *The Brookings Institution,* 2023.

[44] Marca, „ Russia uses AI to spread disinformation about invasion on Ukraine,” 2023.

[45] The Telegraph, “Deepfake video of Volodymyr Zelensky surrendering surfaces on social media,” YouTube, 2022.

[46] G. Pennycook and D. G. Rand, “Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning,” *Cognition,* vol. 188, pp. 39-50, 2019.

[47] S. Lewandowsky and et al, “Science by social media: Attitudes towards climate change are mediated by perceived social consensus, Memory & Cognition, vol. 47, pp.,,” vol. 47, pp. 1445-1456, 2019.

[48] C. Paul and M. Matthews, “The Russian “Firehose of falsehood” propaganda model: Why it might work and options to counter it,” *CA: RAND Corporation,* 2016.