POLITICS & SECURITY
ISSN 2815-3324 Online, ISSN 2535-0358 Print

THE ALGORITHM OF INSURGENCY. THE UNDERSTANDING OF THE EFFECT OF AI-MEDIATED DISINFORMATION ON COGNITIVE SECURITY OF NATIONS.

Herasym Dei
Department of Public Administration and Project Management
University of Educational Management
Kyiv, Ukraine
https://orcid.org/0009-0009-5358-3047
herasym_dei@edu.cn.ua

Abstract. *The spread of sophisticated artificial intelligence (AI) programs, in particular Large Language Models (LLMs) and generative adversarial networks that can create synthetic media (deepfakes), has opened a new and dangerous era in the history of information conflict. The main issue that this technological inflection point is raising, discussed in this article, is the development of AI as a force multiplier of disinformation campaigns that are high-frequency, hyper-personalized, and are increasingly indistinguishable to real human communication, bypassing the traditional safeguards of epistemology, and posing a national stability threat. This article presents two main points by using a multidisciplinary approach that combines the theories of security studies, behavioral psychology, and computational linguistics. To begin with, AI-based disinformation implies a qualitative shift in the paradigm of information warfare towards a more insidious one, whereby the strategic object of interest is not physical or cybernetic infrastructure but the perception of reality and epistemological underpinnings of the democratic society, by the individual citizen. Second, existing national security principles, which are based on the notions of territorial sovereignty and the deterrence of kinetic threats, are essentially unprepared to combat this automated, scalable, and to a large extent attributionless insurgency against the popular mind. The article concludes with the suggestion of an elaborate framework of Cognitive Defense, and states that national resilience in the twenty-first century would be achieved through systemic integration of technological countermeasures, the rise of media and information literacy to the rank of core security imperative, and the creation of binding international norms governing the use of AI in the information space.*

Keywords Cognitive Security, Artificial Intelligence, Disinformation, Hybrid Warfare, National Security Strategy, Computational Propaganda and Information integrity.

1. INTRODUCTION

Take, as an example, a possible near-future situation. A few days before a national election in one of the largest European democracies, a very realistic audio recording is posted in the social networks. It seems that the tape was made during a personal phone call of the presidential candidate with one of the foreign oligarchs, who were negotiating about the funds transfer in return of the favorable energy policy. Over two million shares of the clip have been made within ninety minutes. Already shaky on account of geopolitical tensions, financial markets note a sharp decline. Demonstrations take place in two large cities. The campaign of the candidate makes a desperate denial, which in its turn is another piece of information in a swirling maverick of assertion and refutation. Two days later, a forensic examination of the audio carried out by a university research lab verifies that the audio is a high-quality deepfake that was most likely created by an AI model. However, at that point, it is too late. It is the doubt that it has planted that has

irreparably tainted the integrity of the election, not the falsity of the recording. This is not a speculative fiction, but a logical extension based on the capabilities, which are already being realized today (Chesney and Citron, 2019).

The phenomenon of manipulating the information to enhance the strategic position is naturally as ancient as the conflict itself. The weaponization of narrative has been a steady quality of statecraft and insurgency since the forged Protocols of the Elders of Zion, up to the detailed dezinformatsiya of the Soviet KGB (Rid, 2020). These methods were perfected into a black art in the twentieth century, and state actors created elaborate bureaucracies focused on the creation and distribution of propaganda and disinformation. What has evolved though, is not the purpose but the tool. The shift in the digital era of social media platforms and shortwave radio over the analog era of pamphlets and shortwave radio was a major quantitative jump, and it dramatically boosted the pace, range, and the amount of the information operations. This change was extensively documented by scholars such as P.W. Singer and Emerson T. Brooking (2018), who demonstrated how social connection platforms were turned into battlefields in the case of likewar. The algorithms that dominated these sites, which are engineered to promote engagement over truth, provided a conducive atmosphere of viral dissemination of false and inflammatory data.

However, the introduction of sophisticated generative AI is qualitatively different to the social media revolution. It is not just another amplifier; it is a whole new form of production of informational weaponry. Where earlier campaigns used teams of human operatives to write content, create fake personas, and operate networks of bot accounts, AI is now able to automate these tasks at scale, velocity and personalization that is too great to be possible by humans (Goldstein et al., 2023). Thousands of pieces of text that are unique, contextually appropriate and linguistically fluent can be produced by an LLM in a single hour. A generative adversarial network is able to generate photorealistic images and video of events that did not happen. The intersection of these technologies forms an industrial capacity of falsehood, which radically changes the information environment.

This paper presents and defends the idea of Cognitive Security as a different and important field of national security. Here, Cognitive Security is understood to refer to the safeguarding of the overall cognitive functions of a nation, i.e. the ability of its citizens to engage in rational deliberation, the faith that they have in collective institutions and their obedience to a common and evidence-based perception of reality, against intentional, technologically-assisted manipulation. It is different but connected to cybersecurity (protection of digital infrastructure) and information security (protection of data integrity). Cognitive Security deals with the integrity itself of the human mind as a strategic resource. The key point is that AI magnifies the disinformation threat to the extent of the current human and institutional defenses. Traditional fact-checking organizations, however dedicated, cannot match the output of an LLM. Electoral commissions are not able to check the authenticity of all the media during a campaign cycle. The intelligence agencies find it hard to assign campaigns that are created and propagated via anonymous and decentralized networks.

This article claims that AI-driven disinformation is an inherent change in the security environment, and the transition to a reactive fact-checking model needs to be changed to an active, structural restoration of the national cognitive resilience. This is analyzed in the following manner. Section 2 creates a taxonomy of AI-motivated threats, analyzing the particular capabilities of deepfakes, LLMs, and automated micro-targeting. Part 3 shifts to the psychological level, examining how these tools take advantage of established cognitive weaknesses and weaken social unity, which is presented as cognitive warfare. Section 4 covers the rough policy and legal issues, such as the issues of attribution and the conflict between regulation and innovation. Section 5 suggests a compromising approach to Cognitive Defense, which is based on technological, educational, and international normative solutions. Last but not least, Section 6 presents some final thoughts and suggests future research directions.

ISSN 2815-3324 Online, ISSN 2535-0358 Print

2. THE TAXONOMY OF AI-DRIVEN THREATS

2.1 Generative A.I. and the Death of Seeing is Believing.

Visual and auditory media had been effective epistemic anchors over centuries. A documentary evidence was captured as a photograph; a voice recording would lead to a conviction in a court of law. The evidentiary weight given to sensory media forms a part of the legal and social systems of the modern democracies. The advent of the so-called deepfake technology, which is a term that became widely popular due to the application of deep learning neural networks to the production of synthetic media, has opened a direct attack on this key trust (Chesney and Citron, 2019). The initial deepfakes, which were created with the help of autoencoders and generative adversarial networks (GANs), could be easily spotted thanks to the visual artifacts: unnatural blinking, uneven lighting, or blurred edges around the face. The technology has however been developing exponentially. The latest models are capable of creating synthetic video and audio that are virtually impossible to distinguish by the untrained eye and ear, and more and more so even with specialized forensic equipment (Toews, 2023).

Its national security implications are immense. Take the sphere of military intelligence. A deep fake video that purports a field commander giving an order to retreat or a head of state declaring a surrender would create chaos on a battlefield or even within a civilian population within a few minutes. A fake recording of a diplomatic negotiation may destroy shaky peace talks or cause an international incident in the geopolitical arena. In March 2022, a crude deepfake video of Ukrainian President Volodymyr Zelenskyy surfaced on the internet in which Zelenskyy allegedly addressed Ukrainian soldiers to surrender their weapons. Although the quality of that specific fake was quite poor, making it rather ineffective, it was also a strong demonstration of a concept, a precursor to what more advanced actors with more resources could do (Sasse, 2023). The technology is also gaining accessibility radically. Open-source tools and easily accessible applications have reduced the hurdle to entry, in that the ability to generate believable deepfakes is no longer the prerogative of state intelligence services or high-budget research institutes. The same person on a consumer-grade laptop and with access to the freely available tools can now create content that only a Hollywood studio would have been able to generate only a decade ago.

The destruction of trust in visual evidence has far-reaching ramifications that go way beyond a single fake clip. It contaminates the whole information well. In a situation where all videos are potentially a deepfake, all genuine videos can be falsely claimed. This relationship is referred to as the Liar Dividend and is elaborated in Section 3. In most ways it is a worse consequence than any single deception, in that it assaults the very notion of common evidentiary reality. The courts as well as newsrooms, which depend on media evidence, have their authority diminished not necessarily because they were misled, but because they are subjected to the risk of being misled any time.

2.2 Hyper-Personalized Persuasion and Large Language Models.

While “deepfakes” undermine the visual foundations of the public sphere, large language models (LLMs) pose an equally serious threat to its textual and discursive structures. OpenAI GPT-4, Claude by Anthropic, and LLaMA by Meta among other models have been shown to be able to produce human-like text that is not only fluent and coherent but also context-appropriate across an unprecedented variety of tones, including informal social media posts and academic writing (Bubeck et al., 2023). They are not only relevant to disinformation because they are capable of generating vast amounts of text, but because they can be instructed to do so with particular persuasive aims and be shaped to particular audiences. This is a paradigm shift of broadcast propaganda to precision-guided narrative manipulation.

The past information operations were based on a relatively crude method: a message was created and sent out to a large audience, with the knowledge that only a small percentage of the audience would be responsive. Combining LLMs with the specifics of psychographics, the type of data that digital advertising platforms and data brokers gather every day, will allow employing an entirely different approach. In theory, an AI system might study the history of social media of a target person, their political views, emotional triggers, and social network contacts and create a persuasive message that is uniquely

adjusted to the individual psychological profile of this person (Bontridder & Poullet, 2021). This will not be a science fiction, but the logical continuation of the micro-targeting methods already used by political campaigns and commercial advertisers, but boosted by the generative power of LLMs. This potential was already seen in the early, pre-AI, Cambridge Analytica scandal, which revealed how psychographic profiling might be applied to scale political messaging (Zuboff, 2019). Such operations are made much more efficient and sophisticated by LLMs.

Moreover, it is possible to utilize LLMs to generate and operate entire networks of artificial personas, i.e., fake social media accounts with regular and realistic posting histories, subtle views, and lifelike personal information. The networks of sockpuppets can be used to promote certain stories, give the impression of grassroots support (astroturfing), bully and silence critics, or infiltrate social groups to divide them internally. The only weakness of such operations previously was the human resource needed to sustain the credibility of each persona over time. Hundreds of these personas can be handled by an LLM and a different voice with a consistent and evolving online identity at an insignificant marginal cost (Goldstein et al., 2023). This puts the process of social manipulation into an industrialized format and makes it exponentially more difficult to detect.

2.3 Automated Micro-Targeting: The Synergy of Surveillance and Generation.

These dangers of deepfakes and LLMs are magnified manifold by the sheer fact that they are compounded with the giant surveillance and data-governing systems of the contemporary digital economy. The idea proposed by Shoshana Zuboff (2019) is called surveillance capitalism: it is an economic logic where predicting and altering human behavior is the main product. The comprehensive behavioral descriptions that data brokers create, including buying patterns, location records, web history, search terms, and social circles, are an outstandingly rich resource to any party interested in targeting individuals with customized disinformation. The synergy is strong: targeting intelligence is gathered through surveillance data, and ammunition is generated with the help of generative AI.

This convergence allows an informational warfare that is precision directed at the individual citizen level. Instead of saturating the zone with just one narrative, an enemy can target particular demographic groups or even single targets with their own disinformation tailored to change the beliefs and behaviors of those particular individuals, such as journalists, community leaders, elected officials, military personnel, etc. In the case of a military audience, this may be in the form of faked intelligence reports. To a community activist it may be a bogus news report on a local problem aimed at causing outrage and political action in a wanted direction. To a financial analyst, it can be a false report of earnings that is aimed at selling a particular stock. This targeting ability granularity is a fundamental redefinition of the calculus of information warfare.

2.4 The Velocity of Infection: Asymmetric Velocity.

Another important aspect of the AI disinformation threat that is often underrated is its speed. The empirical evidence on the transmission of information in social media has proved to be incredibly asymmetrical: fake news spread faster, further, and wider than fact-based information (Vosoughi et al., 2018). False claims are emotionally salient and new, which makes this phenomenon dramatic; however, AI contributes to it tremendously. A generative AI system is able to create disinformation content within a fraction of a second. It can saturate online communities of interest within minutes by being spread out via automated bot networks. Response cycles The delay between the release of a bit of disinformation and the reporters, fact-checkers, or government agencies realizing it is false and generating a counter-narrative is measured in hours to days.

This asymmetry in velocity gives the attacker an edge in terms of structure. Although a bit of AI-generated fake news is ultimately debunked, the debunking will often reach a much smaller audience than the initial fake news, and it will come too late when the original misinformation has already taken root in people's minds and possibly triggered real-life response (Walter et al., 2020). This issue is further complicated by the psychological literature on the persistence of beliefs and the so-called continued

ISSN 2815-3324 Online, ISSN 2535-0358 Print

influence effect: despite the exposure of people to a correction, the misinformation they received initially still affects their reasoning and judgments (Lewandowsky et al., 2012). Essentially, AI enables the attacker to get ahead of the defender, and the very construction of the human brain will make the harm caused during the head start partially irreversible. Such asymmetry of speed is not only a tactical, but a strategic advantage since it implies that the sufficiently fast and voluminous disinformation campaign can produce the situation of the faits accomplis in the information space, forming the attitude of people and political results before even the truth can surface.

3. COGNITIVE WARFARE: THE PSYCHOLOGICAL BATTLEFIELD

3.1 Confirmation Bias and Algorithms Echo Chambers.

The effectiveness of AI-induced disinformation is not merely a factor that depends on the sophistication of the technology, but also a factor that depends on the established weaknesses of human thinking. Several decades of studies in the realms of cognitive and social psychology have pinpointed a set of systematic biases that define the ways in which people process information. Primarily, in disinformation terms, is confirmation bias: the long-known phenomenon where people tend to look for, interpret, like, and remember information that supports their already held beliefs and hypotheses, and pay disproportionately less attention to the other possibilities (Nickerson, 1998). The AI-based disinformation is the only one that can be engineered, at a large scale, to take advantage of this particular vulnerability.

The algorithmic structure of social media platforms mediates and magnifies this exploitation. The recommendation algorithms that organize the news feed of users are made in such a way as to maximize the engagement, which is highly correlated with emotional arousal and, by proxy, with the content that validates and reinforces existing beliefs (Bail, 2021). The outcome is what is increasingly being documented as the so-called echo chambers or filter bubbles, where users become increasingly exposed to a set of narrowing set of views that affirm their worldview and protect against dissonant information (Pariser, 2011). AI produced disinformation, carefully tuned to appeal to the existing moods in a particular echo chamber, faces few cognitive challenges. It does not seem to be a foreign propaganda but rather an organic, natural continuation of the discourse of the community that the user belongs to. A predisposition to internal cognitive (confirmation bias) and an external structural (algorithmic curation) amplifier combine to provide an almost flawless medium through which targeted fake information is delivered.

Additionally, AI has the capability of changing its message on the basis of real-time feedback. Digital marketing A/B test can be used in disinformation campaigns: several versions of a story can be published at the same time, and the one that results in the greatest engagement the most shares, the most emotional responses will be automatically scaled, and the less effective ones will be eliminated. It forms an evolutionary algorithm of propaganda, a process which continually improves its misleading output to have the greatest possible psychological effect. This pace, magnitude and adaptive accuracy could not be practiced by any human propagandist.

3.2 The Liar Dividend: Weaponizing Doubt Itself.

The most strategically important effect of the deepfake age is, perhaps, a phenomenon that legal theorist Robert Chesney and Danielle Citron (2019) have called the Liar's Dividend. As explained in this concept, it is a paradoxical effect of the spread of synthetic media: since everyone learns that one can be convinced of fakes, then any actor, a politician caught on tape, a company caught doing something bad, a government faced with evidence of atrocities, can use it as a scapegoat and deny the true, authentic evidence as a fake one. The very presence of deepfake technology offers an easy-to-use way to get an alibi to avoid responsibility.

The Liar is a scathing power of a dividend. It is not just creating falsehoods to the information environment; it is depriving the information environment of truth. It turns all the genuine facts into a refutable argument, thus compromising the epistemological basis of social responsibility, legal action, and

democratic dialogue. The concept of evidence-based public discourse itself is threatened in the world where they no longer see but believe. This ambiguity can be used by political leaders. When faced with a harmful audio or video, a high-profile individual can merely claim that a recording is an AI-generated fake and a large percentage of his/her supporters, already conditioned to accept the influence of partisan media and confirmation bias, will take this at face value. The onus of proving is practically passed upon the accuser, who now has to prove that the evidence has not only what but is actually genuine.

This relationship has immense implications on military and intelligence scenes. The state that is committing war crimes can discount authentic satellite imagery of troop movements, intercepted communications, or video evidence of war crimes as AI fabrications. It was already a pre-AI era playbook, the Russian government was used to ignoring evidence of its role in the downing of Malaysian airplane MH17 and in the poisoning of Skripals, calling it western fabrication, but deepfake technology offers a much more believable shell to wrap such denials in. The Liar's Dividend is not only a domestic political problem, but also an instrument that can be used to weaken the international rules-based order and avoid state accountability on breaches of international law.

3.3 Social Cohesion As A Security Asset.

Classical approach of security has concentrated on physical property: land, infrastructure, weaponry, and economy. In this article, it is argued that social cohesion, the level of trust, mutual recognition, and shared normative agreement in a society, needs to be redefined as an essential national security asset. The overall strategic goal of the AI-led disinformation campaigns be it by a state or non-state agent is not to advance a particular falsehood but to undermine this social fabric beneath it. It is not about attempting to persuade a target audience of a certain falsehood, but about establishing a state of widespread epistemic insecurity and distrust towards each other and the social fabric, which makes a democratic government ineffective and makes a society susceptible to foreign influence or internal breakdown (Pomerantsev, 2019).

This realization alters the threat. The win of the opponent is not the victory of a particular deepfake when it becomes viral or a certain conspiracy theory receives its followers. It is facilitated when the number of citizens who are able to agree on the most fundamental of facts has become too small, when trust in all institutions, including the government, the media, science, the judiciary, has been eroded sufficiently, and when the sphere of political discourse has become mutually hostile and hermetically sealed information ecosystems. Under this condition democratic processes turn into performative rituals, which have no substantive deliberative content, and the polity is vulnerable to authoritarian capture or paralysis. The information operations by Russia, which have been widely discussed by researchers such as Thomas Rid (2020) and Peter Pomerantsev (2019), have long been interpreted as being directed at exactly this effect, not to triumph in an argument, but to eliminate the possibility of argument.

3.4 Case Studies: Information Operations Digital in Action.

A number of real-life examples explain the above discussed mechanisms and highlight the increasing intensity of the threat. To begin with, the 2016 presidential election in the United States can be discussed as a historical case study of the interference in the election by the Internet Research Agency (IRA). Although the activity of the IRA coincided with the present generation of generative AI, it served as an excellent example of the sort of campaign that can now be automated and scaled by AI. The IRA used hundreds of followers to set up and operate fake social media accounts on Facebook, Twitter, Instagram, and YouTube, where controversial content was shared on race, immigration, guns, and religion. They were not aimed at promoting one candidate but increased the existing social divisions of the American society (Mueller, 2019). More importantly, it was a human operation, which was labor intensive. A LLM was able to duplicate and out produce on a small percentage of the staff and budget.

Second, the COVID-19 crisis provided the information space, also known as an infodemic, where false information about the virus, its causes, and the safety and effectiveness of vaccines went viral on digital platforms (Wardle and Derakhshan, 2017; World Health Organization, 2020). State actors, such as Russia

and China, were reported to be spreading false and misleading health-related information to Western audiences, which is another tactic of weakening the trust of the population in health authorities, as well as contributing to social unrest. The pandemic revealed the potential, direct, quantifiable effects of disinformation on the health and safety of the population, portraying digital manipulation as physical damage. Although much of this disinformation in the pandemic was created by humans, it proved that the society of democracies could be easily targeted by coordinated information attacks in the time of crisis, which is now becoming infinitely easier to do using AI.

Third, the current conflict in Ukraine has been used as a live test area of the information war in the age of nascent AI capabilities. Both parties in the conflict have been using the most extensive information operations, yet the use of the said crude Zelenskyy deepfake in March 2022 was a landmark. It was soon recognized as a hoax, but the event showed that the desire to make use of synthetic media would be a weapon of psychological warfare in the present time and gave a glimpse of the future where such tricks will become much more advanced and, by extension, much more dangerous (Sasse, 2023). It also demonstrated how fast synthetic media can spread and how hard it is to respond quickly and authoritatively in a war time situation.

4. POLICY AND LEGAL PROBLEMS

4.1 The Attribution Problem

One of the principles of international security and deterrence is the capability to assign the hostile actions to a particular actor. Deterrence, be it nuclear or conventional, is based on the assumption that an action of aggression would be identified and sent back to the aggressor where he or she would respond accordingly. This reasoning is essentially disrupted by AI-based disinformation. Synthetic text and media generation tools are highly accessible and can be implemented anywhere on the planet using anonymizing technologies like virtual private networks (VPNs) and the Tor network. The material that is produced by a machine, is devoid of any stylistic imprint that can be ascribed to a particular state intelligence agency or non-state organization. What could have required a team of fifty intelligence agents to carry out now could be done by a small team or even by an individual advanced actor, and making the attribution calculus even more complicated (Rid and Buchanan, 2015).

This attribution gap has a number of harmful implications. It reduces the standard of state-sponsored information aggression, since the perceived likelihood of detection and being held to account is reduced. It also gives the non-state actors such as terrorist groups and criminal groups the power to launch information campaigns that are now as sophisticated as those that could only be launched by state intelligence services. And it poses a political problem: despite having an intelligence agency that can confidently attribute a campaign, as the U.S. intelligence community did in the case of Russian interference in 2016, the ambiguity inherent in it means that even the accused state (but also its domestic supporters) can be able to deny the attribution as politically motivated. The absence of a transparent, globally accepted system of assigning and reacting to information attacks is one of the biggest loopholes in global security system.

4.2 Regulation and Innovation: The Democratic Dilemma.

Democratic regimes are in a deep dilemma of trying to control AI-based disinformation. On the one hand, the danger to the national cognitive security is a fact and requires an answer. Conversely, excessive government control over speech and information technology presents risks of trampling the basic civil rights of the people, specifically, the freedom of speech and the freedom of the press, and a lack of innovation, which is the economic and strategic powerhouse of democratic countries. This dilemma is not experienced in authoritarian regimes that regulate their domestic information environment by censoring and monopolizing state media. Such asymmetry is a strategic drawback to democracies in itself (Diamond, 2019).

One of the approaches adopted by the European Union, including the Digital Services Act (DSA) and the AI Act, is a reflection of a general regulatory framework, which places duties on technology platforms on the matters of content moderation, transparency of algorithms, and risk-assessment. Critics, though, claim that these regulations are slow to keep up with the rate of technological change, and may be excessively cumbersome to smaller firms, and may result in a patchwork global regulation system. The United States, in its turn, has mostly been relying on market-driven strategy, where the federal regulation of online speech has been limited, in part, by expansive protections of the First Amendment and Section 230 of the Communications Decency Act. This policy maintains the greatest freedom and creativity but, arguably, it exposes the information space to high levels of danger. The search of sustainable balance, the balancing that would be able to tackle the real security threat and still provide the openness and pluralism that characterize the democratic societies, is one of the classic policy challenges of the age (Fukuyama et al., 2021).

4.3 International Law and the Issue of Sovereignty.

An AI disinformation campaign across borders, that targets a sovereign state, poses some basic questions within the international law. Is such a campaign a breach of the principle of non-intervention in internal affairs as stated in article 2(7) of the United Nations Charter? Would a sufficiently harmful campaign, such as one that, say, causes a state of civil disturbance, or hobbles the capacity of a government to operate, be sufficient to meet the threshold of a use of force under Article 2(4), or even an armed attack warranting self-defense under Article 51? The current legal framework, which was formed in the period of kinetic struggle, has no definitive solutions (Schmitt, 2017).

The Tallinn Manual on the International Law Applicable to Cyber Operations offers some insight into the matter, proposing that cyber operations with the same effect as a kinetic attack can be defined as such under current use-of-force constructs. Nevertheless, disinformation campaigns do not necessarily disrupt and destroy systems directly, but persuade and manipulate. Their impacts, such as undermining trust, polarizing society, manipulating elections, are diffuse, cumulative, and hard to measure in similar terms as material harm. The global community has not yet come up with the legal agreement regarding the classification of such operations, the determination of the norms of the proportionate response, and the imposition of the norms of behavior in the information space. The formation of such a consensus is a matter of immediate concern, and it is undermined by the fact that there are profound disagreements between democratic and authoritarian states concerning the very sense of information freedom and state sovereignty over the digital realm.

5. RECONCILIATORY FRAMEWORK CONSTRUCTION: COGNITIVE DEFENSE

5.1 Technological Redlines: Watermarking and Counter-AI Detection.

The initial component of a Cognitive Defense architecture should be technological. Technology has brought the problem, but it has to be a key component of the solution. One of the proposed solutions is the mandatory watermarking of AI-generated content. Technical standards like the Coalition for Content Provenance and Authenticity (C2PA) are creating technical methods to insert cryptographic metadata into digital content during creation and offer a verifiable provenance history (C2PA, 2023). With the adoption of such standards, users and platforms could validate the authenticity, AI-generated content, and intervention of a piece of content. Nonetheless, there are still major challenges. Sophisticated adversaries can remove or spoof watermarking and it needs to be universal to all AI developers to work, which is complicated by the fact that many of the most popular AI models are open-source. In addition, watermarking is a very reactive control; it assists in detecting the synthetic content once it has been produced, but not in its production.

Along with watermarking, it is necessary to invest in AI-driven detection devices. The arms race in detection technology is similar to the arms race in computer viruses; as the technology of deepfakes and AI-generated text advances, new antivirus programs and similar software are being developed. Deep

learning models that learn to recognize the delicate statistical artifacts separating synthetic media and real media and other models are being developed to detect AI-generated text by analyzing linguistic patterns, perplexity scores, and other computational features (Mitchell et al., 2023). But this is essentially an arms race, a contest between better detection systems and better generation systems to counteract the detection systems. The use of Cognitive Defense strategy cannot be based on detection technology alone, but must form a portion of a layered defense.

5.2 Nordic Model of Societal Resilience.

Technological defense against cognitive threats is not as strong and can be sustained as societal. In particular, the Nordic economies, especially Finland, Sweden, and Estonia, have led the way in a wholeness-of-society or total defense strategy to information resilience, which presents a strong model to other democracies. Finland, as one example, has made critical thinking and media literacy part of the national school curriculum starting in primary school, and educates students on how to detect the manipulative tricks, how to assess the sources, and how propaganda works (Mackintosh, 2019). This is also supplemented by the high level of the public trust in the institutions, the high and autonomous public media, and the cultural principle of civil society and the openness of the institutions.

This strategy acknowledges that the real aim of cognitive warfare is the individual citizen and the best defense mechanism is to have an informed, critical thinking citizenry. In this context, media literacy is not an educational side effect; it is part of the national security infrastructure, like air defense or border security, which is indispensable. Naturally, the Nordic model is hard to translate into larger, more diverse, and more polarized societies, but the principle behind it is universal: the investment into the cognitive resilience of the population is the only effective defense against information manipulation, no matter how sophisticated the tools used are.

5.3 The Platforms Rule: Cognitive Gatekeepers and Democratic Accountability.

The operators of the social media exchange and the creators of the generative AI models technology companies hold a rather special and highly disputed position in Cognitive Defense arena. They are both the developers of the means by which disinformation is propagated by AI, the proprietors of the infrastructure by which they spread it, and the holders of the data and technical skills most likely to counteract it. Their de facto position of being cognitive gatekeepers gives them a duty that is not limited to their duty to their shareholders. This role includes open algorithmic design, which is not systematically biased to engagement instead of informational integrity, effective and proactive content moderation policies that extend to synthetic media generated by AI, substantive and authentic cooperation with governments, civil society, and academic researchers on threat intelligence, and sincere investment in safety research in accordance with their investment in capability development (Persily and Tucker, 2020).

The difficulty is to have in place systems of good governance among these actors, which are not authoritarian, and yet not so liberal as to be meaningless. It needs a system of democratic responsibility, such as independent auditing of algorithms, obligatory transparency reporting of threats related to AI, and consequences that are meaningful in case of failures of due diligence. The creation of industry-wide norms and standards, similar to those already in place within other areas with high implications to public safety (such as aviation or pharmaceuticals), should be actively sought. The principle is simple, the commercial growth and implementation of strong AI systems must have an implicit burden to reduce the expected harms that such systems can facilitate.

6. CONCLUSION

This article has contended that the emergence of the advanced generative artificial intelligence has marked a new and qualitatively different era in the history of the information conflict. The ability of Large Language Models to generate hyper-personalized persuasive text at scale and the ability of deepfake technology to generate believable visual and auditory evidence has generated an industrial capability of falsehood that

radically transforms the strategic environment. It is not just a numerical improvement of the disinformation problem that arose with the social media revolution, but a qualitative shift that requires a qualitative change in our defensive stance.

The analysis has been carried out on various dimensions. Section 2 has shown the taxonomy of threats that reveal the exact capabilities and mechanisms of weaponization of AI tools to informational aggression. Section 3 examined the psychological battlefield, demonstrating how these instruments play on superbly documented cognitive weaknesses, such as confirmation bias, the Liar's Dividend, and how its ultimate strategic object is not the personal conviction but the very social glue and common epistemological premises on which democratic governance relies. Section 4 had to address the daunting policy and legal obstacles, such as the issue of attribution, which is almost intractable, the dilemma of democracy of regulation and freedom, and insufficiency of the current international law. Lastly, Section 5 suggested a multi-layered Cognitive Defense model which incorporates technological technology, resilience building of the society, and responsibility of the platform.

The main idea of the given article, according to which AI-driven disinformation represents a paradigm shift in the security paradigm, and the reaction to these challenges is to shift the mode of response to active, systemic support of national cognitive resilience, has an urgent connotation. Cognitive security should be promoted as a niche academic issue to the national security strategy alongside physical border security, cybersecurity, and economic security. Mental frontiers of the citizens are not a complete defense of the borders of a nation. The security of the digital infrastructure of a country is irrelevant when the contents passing through these systems are used to shatter the ability of the people to think rationally and agree on a common language.

Future research possibilities are manifold and dire. Additional empirical research is required on how AI-generated disinformation may affect the human psyche specifically, but more specifically, on its long-term consequences on interpersonal trust and democratic values. The possibility of AI itself presenting a means of truth-checking, via real-time provenance, automatic counter-narrative generation, and large-scale epistemic surveillance, is something that should be taken very seriously but such technology should be created with a keen awareness of the dangers of algorithmic censorship. The convergence between the AI disinformation and the emergent issues of quantum computing (potentially rupturing the existing cryptographic standards of watermarking) and the metaverse (potentially a whole new dimension of immersion to be exploited) necessitates proactive speculation. Lastly, the post-truth politics of politics requires further investigation regarding the circumstances in which democratic societies can continue to enjoy rational governance within an information environment that is becoming more hostile to a shared truth. The insurgency algorithm is operational. Whether democratic societies can acquire the cognitive immunity to resist it remains an issue.

REFERENCES

Bail, C. A. (2021). Breaking the social media prism: How to make our platforms less polarizing. Princeton University Press.

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy, 3, e32. https://doi.org/10.1017/dap.2021.30

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

C2PA. (2023). Coalition for Content Provenance and Authenticity: Technical specification. https://c2pa.org/specifications/

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1819.

Diamond, L. (2019). The road to digital unfreedom: The threat of postmodern totalitarianism. Journal of Democracy, 30(1), 20–24.

Fukuyama, F., Richman, B., & Goel, A. (2021). How to save democracy from technology: Ending Big Tech's information monopoly. Foreign Affairs, 100(1), 98–110.

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. Georgetown University Center for Security and Emerging Technology.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest, 13(3), 106–131.

Mackintosh, E. (2019, May 21). Finland is winning the war on fake news. What it's learned may be crucial to Western democracy. CNN. https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. Proceedings of the 40th International Conference on Machine Learning (ICML).

Mueller, R. S. (2019). Report on the investigation into Russian interference in the 2016 presidential election. U.S. Department of Justice.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin Press.

Persily, N., & Tucker, J. A. (Eds.). (2020). Social media and democracy: The state of the field, prospects for reform. Cambridge University Press.

Pomerantsev, P. (2019). This is not propaganda: Adventures in the war against reality. PublicAffairs.

Rid, T. (2020). Active measures: The secret history of disinformation and political warfare. Farrar, Straus and Giroux.

Rid, T., & Buchanan, B. (2015). Attributing cyber attacks. Journal of Strategic Studies, 38(1–2), 4–37.

Sasse, B. (2023). The era of deepfakes: Advancing technologies and growing threats. Atlantic Council Digital Forensic Research Lab.

Schmitt, M. N. (Ed.). (2017). Tallinn Manual 2.0 on the international law applicable to cyber operations (2nd ed.). Cambridge University Press.

Singer, P. W., & Brooking, E. T. (2018). LikeWar: The weaponization of social media. Eamon Dolan/Houghton Mifflin Harcourt.

Toews, R. (2023, May 22). The next frontier for deepfakes: Real-time face swaps. Forbes.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. Political Communication, 37(3), 350–375.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making (Report DGI(2017)09). Council of Europe.

World Health Organization. (2020). Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.