CYBERPOLITIK JOURNAL
Siber Politikalar Dergisi

A Peer Review International E-Journal on Cyberpolitics, Cybersecurity and Human Rights

www.cyberpolitikjournal.org

ABOUT THE JOURNAL

Owner/Sahibi
On behalf of Association for Human Rights Education / İnsan Hakları Eğitimi Derneği adına
Prof. Dr. Nezir Akyeşilmen

Issue Editor
Dr. Kamil Tarhan

Peer Review
All articles in this journal have undergone meticulos peer review, based on refereeing by anonymous referees. All peer review is double blind and submission is online. All submitted papers (other than book and article reviews) are peer reviewed.

The Journal
The languages of the Journal are both Turkish and English.

ISSN 2587-1218
Cyberpolitik (CP) aims to publish peer-reviewed scholarly articles and reviews as well as significant developments regarding cyber world, cybersecurity, cyberpolitics and human rights.

Indexing/Endeksler

Cyberpolitik Journal is being indexed by;

* Academia Social Science Index (ASOS),
* Scientific Indexing Services (SIS),
* Eurasian Scientific Journal Index (ESJIndex),
* Index Copernicus International (ICI), (ICV 2017=64.65)
* Directory of Research Journal Indexing (DRJI).
* JournalITOCs.
* Open-Web.info.
* Google Scholars

Issue Referees / Sayı Hakemleri

Prof. Dr. Mehmet Emin Erendor

Assoc. Prof. Dr. Sonny Zulhuda

Assoc. Prof. Dr. Murat Tümay

Assoc. Prof. Dr. Khuram Ikbal

Assoc. Prof. Dr. Emre Çıtak

Asst. Prof. Dr. Ayşegül Sili Kalem

Asst. Prof. Dr. Sadullah Özel

Asst. Prof. Dr. İsmail Yıldız

Asst. Prof. Dr. Abdul Gafur Arifin

Asst. Prof. Dr. Elif Davutoğlu

Dr. Kamil Tarhan

Dr. Gloria Shkurti Özdemir

Dr. Yonah Welker

Dr. Sarkis Karaguezian

Cyberpolitik consists of the following sections:

Research Articles: Each Volume would publish a selection of Articles covering aspects of cyber politics and human rights with a broad universal focus.

Comments: This section would cover recent developments in the field of cybersecurity, cyber politics and human rights.

Book/Article Reviews: Each Volume aims to review books on cyber politics, cybersecurity and human rights.

Cyberpolitik Award: Each year one '*Cyberpolitik*' prize will be awarded, for the best article from material published in the previous year.

CONTENTS / İÇİNDEKİLER

EDITORIAL PREFACE: THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE IN ACADEMIC WRITING AND ETHICS: THE CONDITION OF HYPER-PLAGIARISM

Dear Readers

We are proud to present to you the 20th issue of the *Cyberpolitik Journal*. It is a great honor for all of us to continue our journey that we started nine years ago without interruption. As the digital world grows every day and every second, new developments and new technologies emerge, we are trying to read and understand this domain within our limitations.

In an era dominated by the omnipresence of technology and interconnected digital ecosystems, the role of artificial intelligence cannot be overstated. The articles featured in the volume 20th issue of the *Cyberpolitik Journal* bring forth a compelling narrative, shedding light on diverse facets of cyber landscapes, from ethical considerations and human rights to human rights, from cybersecurity to data protection in digitally enriched environment.

The use of artificial intelligence in academic writing processes has increased at an unprecedented pace in recent years. While artificial intelligence offers technical conveniences and expanded access to academic literature, it simultaneously carries a serious potential to undermine academic production skills and fundamental principles of academic ethics. Academic production is, by its very nature, required to be original, written in an academic style, and grounded in objective and verifiable scientific evidence. Academic texts are the product of deep intellectual labor and are therefore regarded as reliable and credible.

However, the rapid and largely uncontrolled proliferation of generative artificial intelligence tools has led to the subordination of academic ethics to pragmatic considerations, placing academic writing under serious threat. First and foremost, generative artificial intelligence does not produce original ideas or thoughts. At its core, it merely processes and recombines information already present in existing databases. From a classical perspective, such production is entirely secondary in nature and cannot claim originality. Consequently, the direct and uncritical use of texts generated by artificial intelligence clearly falls within the scope of plagiarism.

This situation is more problematic than traditional forms of plagiarism. Generative artificial intelligence is capable of producing inaccurate, incomplete, or entirely fabricated information while using a highly persuasive and fluent language. Such outputs inherently carry risks of disinformation, misinformation, and information manipulation. If the researcher lacks sufficient expertise in the relevant field or fails to critically examine the generated text, the incorporation of entirely fictional content into academic work becomes almost inevitable. Unfortunately, current observations indicate that such erroneous and unverified uses are becoming increasingly widespread.

A third—and perhaps the most critical—problem concerns the dimension of labor. Even in cases of traditional plagiarism, the researcher typically engages in some level of academic effort: conducting research, accessing sources, and exerting at least a minimal degree of cognitive labor. In contrast, with the extensive use of generative artificial intelligence, even this minimal effort is largely eliminated. The production process is almost entirely delegated to the machine, while the researcher's contribution becomes negligible. In such cases, neither the idea nor the text belongs to the researcher, nor is academic responsibility meaningfully assumed. Academic production thus turns into an ethically problematic activity with an ambiguous or absent subject.

Thus, uncontrolled and unlimited use of artificial intelligence in academic work does not merely result in individual ethical violations; it constitutes a structural threat to the credibility of academia and to the epistemological value of knowledge itself. Although traditional plagiarism and artificial intelligence–based plagiarism are fundamentally different, placing them on the same level is neither accurate nor fair. Artificial intelligence increasingly assumes the role of the first author. Therefore, describing such works simply as plagiarism is insufficient. Instead, it is more appropriate to define these labor-free, error-prone, and ready-made artificial intelligence products as **hyper-plagiarism.** This form of digital plagiarism far exceeds traditional plagiarism in scale and severity and clearly constitutes an unethical practice.

At its core, as argued by many liberal thinkers—most notably John Locke—property is fundamentally grounded in labor. Through labor, individuals mix their efforts with nature and thereby transform something into property. Intellectual property is similarly shaped and constructed through labor. Excessive use of artificial intelligence, however, eliminates this

labor. Under such circumstances, whatever is produced can scarcely be regarded as property at all.

Moreover, artificial intelligence use is largely devoid of the critical thinking that lies at the heart of academic writing. Critical thinking is a driving force of intellectual advancement. Generative artificial intelligence not only undermines critical thinking but effectively eliminates thinking altogether. Researchers increasingly resort to the comfort of artificial intelligence, allowing it to "think" on their behalf instead of engaging in intellectual and especially critical reflection themselves. As this reliance deepens, individuals gradually distance themselves from thinking, resulting in intellectual stagnation and mental passivity. In this sense, artificial intelligence contributes to a form of cognitive regression, contradicting the long-held assumption that human evolution proceeds toward greater intellectual capacity.

Furthermore, artificial intelligence undermines the principle of objectivity, one of the most defining features of academic production. Researchers strive—however imperfectly—to interpret available data as objectively and impartially as possible. Yet artificial intelligence algorithms are typically designed in ways that reflect the preferences, interests, and priorities of those who develop or control them. On many controversial issues, artificial intelligence provides one-sided interpretations, thereby obstructing the distinction between right and wrong and undermining the collective pursuit of more accurate and beneficial knowledge.

Artificial intelligence also erodes the reliability of academic writing. While drawing from existing databases, it can simultaneously generate information that is entirely irrelevant or incorrect, thereby producing disinformation. Disinformation, misinformation, and information manipulation ultimately undermine one of the core purposes of scientific production: contributing to solutions for human problems. Rather than serving this goal, such distorted information generates new problems and deepens existing ones.

That said, it would be incorrect to conclude that artificial intelligence is inherently harmful or entirely unsuitable for academic work. When used correctly and responsibly, artificial intelligence can function as a valuable analytical tool. Scholars may consult it for ideas or use it to enrich their existing arguments. If artificial intelligence is purposefully trained, transparently used, and carefully controlled, it can be beneficial. Researchers are ethically obliged to clearly disclose the extent and manner of its use. This includes specifying whether artificial intelligence was employed for language editing, translation, structuring, or idea generation, and to what degree.

When used transparently and honestly, such practices do not constitute ethical violations and may even enhance the reliability and quality of research. However, in many submissions to journals or edited volumes, authors claim that artificial intelligence was used solely for editing or translation. Yet when subjected to artificial intelligence detection tools, careful scrutiny, and comparison with the authors' previous works, substantial portions of the text are revealed to have been generated by artificial intelligence. This not only fosters intellectual laziness but also encourages dishonesty, resulting in a comprehensive ethical failure.

For these reasons, I argue that academic texts in which artificial intelligence functions as the primary author or is used extensively should not be labeled merely as plagiarism. The appropriate term should instead be **hyper-plagiarism**, and such works should be categorically rejected—at least under current technological conditions. Future developments may allow for clearer distinctions between human and artificial intelligence–generated content, and new ethical frameworks may emerge. However, the present argument is grounded in the existing technological context and the current mode of generative artificial intelligence use. While it remains uncertain whether the concept of hyper-plagiarism fully captures the phenomenon, a conceptual tool is clearly needed to distinguish between traditional, limited, labor-based plagiarism and entirely labor-free, thoughtless artificial intelligence–generated texts. In this regard, the concept of **hyper-plagiarism** may serve as a useful analytical framework.

Contents of the New Issue

In recent decades, the rapid evolution of digital technology has fundamentally transformed the way we live, work, and communicate. As the digital domain continues to expand, it brings with it a myriad of opportunities that promise to enhance our global connectedness, increase access to information, and democratize knowledge. However, alongside these benefits, the digital age also presents significant ethical dilemmas that challenge our moral frameworks and societal norms. As the contributors to this issue of *Cyberpolitik Journal* explore, the ethics of the digital domain are multifaceted and require careful consideration from scholars, policymakers, and practitioners alike.[1]

In this context, the first article of the new issue is handled by Muhammet Ali Demir with the title "*Digital Shield: The Protective Role Against Human Rights Violations in Cyber*

[1] ChatGPT has been used for translation and language polishing purpose in this editorial piece.

Interventions. The article explores the potential of cyber humanitarian intervention within the framework of the Responsibility to Protect (R2P) in preventing and halting mass atrocity crimes. Moving beyond long-standing debates on the sovereignty implications and operational risks of traditional military interventions, the article focuses on the emerging opportunities offered by digital technologies and assesses how cyber operations may contribute to the implementation of R2P. Drawing on a normative analytical framework that integrates international law, cybersecurity, and humanitarian intervention scholarship, the study examines the role of cyber measures in safeguarding access to information, protecting communication infrastructures, and constraining the digital capacities of perpetrators. It further identifies key legal, ethical, and practical challenges- such as sovereignty concerns, attribution problems, limited international cooperation, and accountability gaps- arguing that cyber humanitarian intervention functions as a complementary, rather than a standalone, mechanism for advancing R2P.

Carmen-Gabriela Bostan's Study, *"Artificial Intellienge in Education: Regulation, Ethics, and Security*", offers an in-depth examination of the implications of using AI in digital education, with a particular focus on algorithmic transparency, data protection and institutional responsibility. The author analyses how AI systems influence decision making x processes in teaching and learning, highlighting the need for clear public policies to regulate their implementation. Drawing on international best practices and case studies from Finland, Estonia and Romania, the study proposes strategies for the responsible use of AI, including training teachers in digital ethics, developing governance frameworks based on risk assessment and ensuring human oversight of algorithmic decisions. Through an interdisciplinary approach that combines digital pedagogy, ethical standards and legal safeguards, the research argues for aligning technological innovation with democratic values and human rights, so that AI becomes a tool for support and empowerment in education. Overall, the paper provides a coherent framework for the ethical and safe integration of artificial intelligence into educational systems, advocating for a sustainable and inclusive approach to the use of intelligent technologies in learning environments.

The study, titled "*AI-Driven Disinformation as A Global Cybersecurity Threat to Democratic Systems",* written by Murat Emeç, analysiz that AI-generated disinformation has become a new form of cybersecurity threat that targets not technical infrastructure but the cognitive foundations of democracy—public trust, perception, and informed decision-making. It shows how generative AI amplifies false narratives with unprecedented speed, scale, and realism,

weakening electoral processes and institutional credibility. The study concludes that democracies must adopt a holistic security approach that strengthens cognitive security and societal resilience alongside technical measures.

The article "*Dijital Kapitalizm Çağında Yapay Zekâ, Gözetim ve İnsan Hakları: Mahremiyetin ve Özgürlüğün Geleceği"* by Demet Şefika Mangır argues that author addresses the increasing importance of protecting human rights and freedoms. She particularly emphasizes the growing surveillance systems resulting from the increasing integration of technological tools with AI. She analyzes the threats these systems pose to both fundamental rights and freedoms, as well as the right to privacy. In this context, the author presents a theoretical framework that intersects the economic logic of digital capitalism with the social impacts of surveillance technologies, examining how human rights are transformed in the digital age.

The last article "*Yapay Zeka ve Küresel Güvenlik Mimarisi: Güç Dağılımının Mekanizmaları ve Yapısal Dönüşüm*" by Kürşat Kan examines how artificial intelligence is reshaping the global security architecture. It approaches AI not merely as a technical tool but as a capability that accelerates and reconfigures power relations. The study argues that competition is shifting from model success to strategic resources, including digital infrastructure, data, compute capacity, and advanced chip supply chains. It compares emerging coordination challenges across institutions such as NATO, the EU, and the UN. It also explains this transformation through a traceable analytical framework built around specific mechanisms. Ultimately, it maintains that modular governance instruments are more practical when global regimes become gridlocked. In doing so, the article offers concrete policy options for risk reduction and norm-making.

The first opinion was written by Mihai Sebe, Alexandru Georgescu, and Eliza Vaş. Titled "Democracy in the Age of AI: The Fine Line Between the Known and the Unknown," their opinion paper offers a significant global perspective on AI and democracy, making a substantial contribution to both this issue of the journal and the literature as a whole.

In the second commentary, Merve Suna Özel-Özcan offers fascinating insights with his commentary titled *" World-System Hierarchies and AI-Driven Security Competition."* The second commentary is a powerful and original piece that blends world systems theory with offensive realism in the context of artificial intelligence. In her opinion, the author offers the reader a crucial perspective on the transition between the AI world and the classical world.

Finally, an important book reviews provide valuable insights into ethics. Selim Mürsel Yavuz reviews the book "*Humans in the Cyber Loop: Perspectives on Social Cybersecurity* " (Edited by Dorota Domalewska, Aleksandra Gasztold, and Agnieszka Wrońska) (2025). This study offers a comprehensive overview of the concept of cyber loop in cybersecurity studies.

In summary, the articles, commentaries, and book review in this issue contribute to our better understanding of the opportunities and risks presented by the digital age. These contents, prepared with academic depth and visual integrity, aim to open doors to interdisciplinary thought and new areas of discussion. We hope they inspire our readers and open new horizons.

Nezir AKYEŞİLMEN, Ph.D

Editor-in-Chief

RESEARCH ARTICLES / ARAŞTIRMA MAKALELERİ

DIGITAL SHIELD: THE PROTECTIVE ROLE AGAINST HUMAN RIGHTS VIOLATIONS IN CYBER INTERVENTIONS

Muhammet Ali Demir[*]
ORCID ID: 0009-0004-0201-8391

Declaration[*]

Abstract

Atrocity crimes represent some of the most severe violations of international order and are primarily addressed within the framework of humanitarian intervention and the Responsibility to Protect (R2P). Traditional military interventions have been widely criticized due to their potential infringement on state sovereignty and the high risk of operational failure, whereas emerging digital technologies have introduced cyber humanitarian intervention as a possible alternative. The aim of this article is to explore the potential of cyber operations in preventing or halting mass atrocity crimes within the context of R2P and to critically assess the legal, ethical, and practical constraints of this approach.

Methodologically, the study adopts a normative analytical framework, drawing on international law, cybersecurity, and humanitarian intervention scholarship to establish a conceptual and legal basis. Existing literature tends to focus predominantly on military or diplomatic means of intervention, with only limited engagement with the notion of cyber humanitarian intervention. This gap highlights the need for a comprehensive assessment of how cyber measures align with international law, their feasibility, and associated risks.

The findings suggest that cyber interventions may support the implementation of R2P by safeguarding access to information, protecting communication infrastructures, and limiting the digital capacities of perpetrators. Nevertheless, the approach also entails significant limitations, particularly concerning state sovereignty, attribution challenges, the lack of international cooperation, and ethical accountability. In conclusion, while cyber humanitarian intervention does not constitute a definitive solution on its own, it can be considered a complementary tool for enhancing the effective realization of the R2P principle.

Keywords: Cyber, Humanitarian Response, Responsibility to Protect, R2P

[*] Lecturer, Turkish National Police Academy, Karaman, Türkiye, malidemir1501@gmail.com
[*] This article has been prepared without the use of any Artificial Intelligence (AI) tools or assistance.

Introduction

Atrocity crimes are crises that threaten individuals right to life, lead to widespread human rights violations and mass victimisation, and necessitate solutions from the international community. Such crimes, particularly in situations such as wars, genocides and internal conflicts, make the protection of civilians a moral and legal responsibility. In this context, the approaches developed by the international community through the principles of *humanitarian intervention* and *R2P* play a significant role in preventing and resolving victimisation. Humanitarian intervention aims to establish a rapid and effective intervention mechanism for crisis areas by balancing the sovereign rights of states with the fundamental rights of individuals. However, the political, legal, and ethical dimensions of these interventions give rise to international debates.

In recent years, alongside technological developments, transformations have been occurring in the dynamics of conflict and crisis, with cyber technologies reshaping the concepts of war and intervention. In this context, cyberspace has emerged as a new arena of struggle for individuals, institutions and states through information and communication technologies. The growing influence of cyberspace has brought the concept of humanitarian intervention into the digital realm. At this point, the concept of *cyber humanitarian intervention* refers to an innovative approach developed to prevent human rights violations and protect civilians through digital technologies. Methods such as information operations, digital surveillance, and the protection or manipulation of communication networks in crisis areas are considered within the scope of cyber humanitarian intervention. However, this new paradigm raises questions about how it will align with the principles of sovereignty and intervention in international law and how it will be ethically grounded.

The role of cyber humanitarian intervention in preventing human rights violations is of critical importance, particularly in conflict zones, in areas such as protecting communication infrastructure, preventing disinformation, and ensuring the safety of victims. However, fundamental challenges encountered in this process include interventions that conflict with states sovereign rights, the risk of misuse of technological tools, and technical and political obstacles that limit the effectiveness of digital interventions. Therefore, cyber humanitarian intervention stands out as a multidimensional phenomenon that presents both opportunities and risks.

This article will first examine the theoretical foundations of the concepts of humanitarian intervention and the responsibility to protect. It will then discuss the characteristics of cyberspace and the concept of cyber humanitarian intervention. Finally, it will explore the role of cyber humanitarian intervention in preventing human rights violations and evaluate the opportunities and limiting factors in this field. In this context, a critical analysis will be presented on how cyber technologies provide advantages in humanitarian intervention processes, as well as how international law and ethical values will be shaped. The article aims to discuss the potential of cyber humanitarian intervention to offer an innovative solution to humanitarian crises.

The Concept of Humanitarian Intervention and the Responsibility to Protect (R2P)

Over the past thirty years, the concept of *humanitarian intervention* has been one of the most frequently debated topics among both academics and practitioners. Questions such as what the challenging role of the concept is in relation to state sovereignty, or what the minimum level of crisis should be for intervention in humanitarian crises caused by the sovereign itself, have formed the basis of this debate. Given that humanitarian intervention is a concept that is controversial in essence, it is important to define it in order to express its scope (Gulati and Khosa, 2013: 398).

Şaban Kardaş (2003: 21) defines humanitarian intervention as coercive action taken by a state or states or international organisations against a target state that seriously and flagrantly violates human rights, with the aim of protecting the target state's citizens, through the use of armed force or the threat of force, regardless of the target state's consent. This definition, on which Kardaş bases his argument, is actually in line with the definition adopted by NATO in November 1999. The fundamental elements of this definition are focused on *sovereignty* and *human rights*. Firstly, for an action to be considered humanitarian intervention, there must be a violation of the sovereignty of the target state. Secondly, the fundamental trigger for the intervention must be the aim of resolving human rights violations (Roberts, 2000: 1).

International law did not consider any intervention on the territory of a state without the consent of that state to be legitimate, even for urgent humanitarian purposes agreed upon by the entire international community, until the Second World War. In 1945, however, the United Nations (UN) prohibited intervention, banning the use of force or the threat of force against the territorial integrity of a state, and also prevented any state from providing military support or intervention to either side in another state's civil war. Serious efforts to develop a

form of collective intervention began under the leadership of the UN Security Council (UNSC) with the end of the Cold War. In 1991 and 1992, interventions took place in Iraq and Somalia, not primarily justified on humanitarian grounds – a term not found in the UN Charter – but fundamentally due to mass human rights violations (Helkin, 1999: 824).

In 1999, NATO bombed Yugoslavia to protect the Albanian population in Kosovo from ethnic cleansing. Although this military operation was considered morally justified, it was criticised for violating international law for the sake of interests, and indeed the UN Security Council did not express a favourable opinion on the military intervention in question. (Gilligan, 2013: 22). The Kosovo intervention and crises such as those in Rwanda, Burundi and Bosnia and Herzegovina, which led to mass killings in political history, have revealed a normative deficiency agreed upon by both states and international organisations (Coady, Dobos and Sanyal, 2018: 18-19).

The Canadian government established an independent commission called the *International Commission on Intervention and State Sovereignty (ICISS)* in 2000 in an effort to overcome the humanitarian intervention crisis. In 2001, the commission published a 90-page report entitled *The Responsibility to Protect (R2P)*, accompanied by a 400-page book detailing the report. The most significant development for the concept came in 2005 when heads of state endorsed R2P in the Outcome Document of the UN World Summit. In subsequent years, the UNSC referred to the R2P concept and published a report entitled *Implementing the Responsibility to Protect* in 2009 (Badescu, 2011: 3).

The ICISS report essentially consists of three main sections. These are *prevention*, *response* and *reconstruction*. *Prevention* is the first section placed at the centre of R2P. This stage involves a shift from the habit of responding after a crisis has occurred to the habit of taking preventive measures before a crisis occurs (ICISS, 2001: 39). The prevention phase is itself divided into three sub-headings. The first part is the *Early Warning and Analysis* section, where data and information on human rights violations are collected, the reality is clearly revealed, and the aim is to take swift political action based on the data collected. The second part is *Efforts to Prevent the Root Causes of Crises*, which aims to transform the main causes of conflict, such as income inequality, underdevelopment or political oppression, through various reforms, effective governance, the protection of fundamental rights and freedoms, the rule of law and the development of welfare. The final section is *Direct Prevention Efforts*. This section includes various sanctions such as providing direct assistance to the violated

community, imposing political sanctions on the violating state, diplomatic isolation, and the threat of force (ICISS: 21-24).

Reaction is the second and most controversial section of the R2P report. There are two main reasons why it is controversial: firstly, reacting poses a threat to state sovereignty; secondly, the question of who has the authority to react. At this point, the ICISS has established six fundamental criteria for the legalisation of a military response to mass human rights violations. These are: *proper authority*, *just cause*, *right intention*, *last resort*, *proportionate means*, and *reasonable expectations*. *Proper authority* lies with the most appropriate international body, the UN Security Council. *Just cause* refers to situations involving large-scale loss of life or ethnic cleansing. *Right intention* is to stop and prevent human suffering. *Last resort* means that all possible diplomatic or non-military means must be considered before resorting to military force. *Proportional means* that the level, duration and intensity of the intervention must be kept to a minimum, taking into account humanitarian safeguards. *Reasonable expectations* should be pursued if there is a likelihood of success in preventing a humanitarian crisis following the intervention (ICISS: 32-37).

Reconstruction is the final section of the ICISS report. This section actually addresses the question of how to emerge better from the state that has been intervened in after the intervention and focuses on the post-intervention period. The fundamental aim is to ensure *lasting peace*. The objective here is not to provide humanitarian aid or achieve development goals, but to create the right conditions for genuine reconciliation that will eliminate the possibility of renewed conflict. The reconstruction process is divided into three sub-sections: security, justice, reconciliation and development. The coordination established between local and international actors facilitates reconstruction efforts (ICISS: 39-45).

The 2009 UN Secretary-General's R2P report examined the extent to which the Responsibility to Protect implementation strategy assisted the organisation's efforts to fulfil its commitment to protect communities from atrocity crimes, highlighted shortcomings, and stated that R2P should be understood as a programme based on three pillars. The first pillar is the *state's responsibility to protect*. A state is required to protect its own society from serious human rights violations such as genocide, ethnic cleansing or crimes against humanity, and this is an international obligation. The second pillar is *international assistance and capacity building*. Unlike the responsibility placed on the state in the first pillar, this pillar places a responsibility on the international community, including supporting states in protecting their populations

from these crimes, including meeting the urgent needs of communities at risk. The third and final pillar is *the international community's timely and resolute response through the UN*. This means that, when peaceful options prove insufficient and the relevant community cannot be clearly protected from atrocity crimes, member states must take collective action through the UN Security Council (United Nations General Assembly, 2009: 2).

Nicholas J. Wheeler (2000, 34-35) argues that four fundamental thresholds must be met for an international military intervention to be considered a legitimate humanitarian intervention. First, the existence of an urgent humanitarian situation, such as mass deaths or ethnic cleansing; second, the exhaustion of all diplomatic and economic avenues, making military force the last resort; third, the violence used must be proportionate and not exceed the humanitarian objective; and finally, the intervention must have a reasonable prospect of producing a positive outcome in improving the humanitarian situation in the region.

Sovereignty in the modern international system functions not merely as a protective shield against external interventions, but rather imposes a positive responsibility upon states to ensure the welfare and security of their own populations (Deng, 2010, 354-355). The R2P doctrine has redefined state sovereignty by shifting it away from Jean Bodin's classical interpretation of absolute and inviolable authority, reconceptualizing it instead as a sphere of responsibility inherently linked to the duty to protect the population. Gareth Evans, the former co-chair of the ICISS and one of the primary architects of the doctrine, articulates this transformation in the following terms:

The issue is not the right of states to intervene, but rather the responsibility of states to protect their own people from crimes of mass atrocity and the responsibility of the international community to assist them in this regard. This shift is a transformation from sovereignty as control to sovereignty as responsibility. (Evans, 2008: 42).

The R2P concept is criticised from many angles. The first criticism concerns the fact that, although it is referred to in UN reports or documents, it is not a binding international legal norm. The absence of an international agreement that explicitly refers to R2P and its conflict with certain customary international law principles, such as sovereign equality among states and non-interference in internal affairs, has been the focus of criticism regarding this lack of legal norms (Borgia, 2015: 228). Another point of contention is that R2P has yet to establish a standard of success or implementation in humanitarian crises. The fact that the UNSC acts within the framework of national interests in international humanitarian crises and does not

grant authorisation to intervene in one crisis while refusing to do so in another crisis of similar severity forms the main argument of such critical studies (Alexander, 2024). Furthermore, the fact that the concept of R2P is discussed more than humanitarian values in situations where humanitarian crises occur is another criticism levelled at the doctrine (Illingworth, 2024: 185). The final point of criticism is that the doctrine legitimises the use of force by citing humanitarian objectives. These criticisms emphasise that the doctrine is used in the same sense as military intervention, which it envisages as a last resort in the response phase (Massingham, 2009: 804).

Cyber, Cyberspace and Cyber Humanitarian Intervention

The concept of cyberspace has been expressed in many different ways in the literature, such as anything related to computers/the internet or a virtual reality, but no common definition has been agreed upon. The concept's limitless and multi-layered structure has led to it being called *cyberspace*. Nezir Akyeşilmen (2018a:54-55) has stated that in order to conceptualise cyberspace, it is necessary to identify all its elements. According to Akyeşilmen, cyberspace essentially consists of four elements. These are: the actor *human* who uses the internet/computer, which is the environment of virtual space, and who creates, destroys or disseminates the information/data found there; *information*, which contains elements such as images, videos or text developed within the virtual framework; the virtual language, i.e. the *logical framework* (software) created with code prepared according to a specific protocol, and the *physical infrastructure* (hardware), from computers to cables or other service providers, which enables the formation of this logical framework.

Cyberspace is frequently discussed in International Relations (IR) literature alongside concepts such as *cyber attack*, *cyber warfare*, or *cyber security*. When examining the main arguments of these studies, the focus is generally on whether reciprocal cyber attacks can be labelled as warfare, and if cyber warfare exists, whether it is similar to or different from traditional warfare. Consequently, within the discipline, one can observe either a reductionist approach or an approach that attaches excessive importance to concepts with the prefix “cyber”. For example, Thomas Rid (2011: 5-7) defines cyber attacks not as warfare but as actions that can be used for destruction, espionage, and sabotage. focusing on the deadly nature of war and its character as a means to political ends, as described in Clausewitz's On War, and considers it unlikely that cyber will acquire the nature of war in the past, present or future. In the literature, there are views that the idea of cyber capabilities being used as an

absolute weapon is pessimistic, that very few cyber attack outcomes translate into political impact, and therefore the use of cyber capabilities will not be widespread (Liff, 2012: 426). John Stone (2012: 106-107), however, responds to Rid's arguments by emphasising that war involves power and violence but does not necessarily result in death, and states that cyber warfare is possible, referring to cyber as an unusual phenomenon.

While the ontological status of cyber warfare as a distinct phenomenon of armed conflict remains a subject of scholarly contention within the discipline, the strategic significance accorded to cyberspace by sovereign states continues to intensify. The fundamental reason for this is that cyberspace essentially encompasses *information*. Andrey Kokoshin, former Deputy Defence Minister of Russia, defined cyberspace as a way to render the opponent's command and control systems ineffective through misinformation, highlighting its strategic and operational aspects (Thomas, 2014: 103). It can be said that today's states are *information-based actors*. They analyse and attempt to solve problems related to their governance by gathering information. Individuals also need information, or data, from states, ranging from social security rights to justice, agriculture to weather data (Balkin, 2012: 4).

Rapid developments in information and communication technologies have integrated the internet, computers, smartphones and social media into every aspect of life. While these developments have significantly facilitated access to information, they have also brought about certain negative consequences. Particularly in digital and chaotic environments where individuals' rational and instinctive thinking abilities weaken in the face of complex situations, and where excessive and diverse information flows prevail, mental shortcuts aimed at reducing cognitive load have begun to be used. This situation makes it easier to change or direct the perceptions of individuals and societies. Regardless of their objectives, various actors can exploit this vulnerability to wage a kind of ‘information war’ through self-serving propaganda or false content (Lin, 2019: 189).

The boundless and largely anarchic nature of cyberspace makes the principles of cyber governance more essential than ever today. Cyber governance emphasises that cyberspace is not merely a technical infrastructure domain; it is also an integral part of the global governance paradigm that encompasses strategic objectives such as respect for human rights, the rule of law, and the establishment of online democracy. In this context, cyber governance serves as an effective safeguard and refuge for the protection of fundamental rights and freedoms (Akyeşilmen, 2018b, 2-5).

However, the digital age has also generated new threat domains that facilitate interference in democratic processes by both state and non-state actors and challenge the fundamental values of democratic societies. Among the most prominent of these threats are the sabotage of democratic electoral processes, the dissemination of violent content, and the manipulation of public opinion. Allegations of Russian interference in the 2016 United States presidential elections, state-sponsored cyber operations such as the Stuxnet and Sony attacks, and the decision of the Australian government to prevent Huawei from participating in the country's 5G infrastructure constitute notable examples of how cyberspace can be exploited by states for malicious purposes (Paterson, 2020: 439–440). In addition to states, hacker groups such as Anonymous—lacking a centralized authority, a coherent ideology, or a fixed objective—also engage in activities within cyberspace that influence states and societies. These groups are particularly known for actions such as releasing leaked materials, gaining unauthorized access to the data of global security firms, and disrupting the websites of multinational corporations (Uitermark, 2017: 403). Moreover, cyberspace is extensively utilized by global terrorist organizations. For instance, the Islamic State of Iraq and Syria (ISIS) has recruited militants from various countries through social media-based propaganda campaigns; in this regard, El-Ravi (2016: 744) notes that the organization increased its global visibility by disseminating positive content in multiple languages that emphasized charitable activities toward the elderly and portrayed everyday life as sustainable in the cities under its control.

Data is as threatening as bullets and bombs (Pellerin, 2011). In an era where bombs are guided by GPS systems and war vehicles are equipped with massive amounts of data, neglecting cyberspace represents a major security vulnerability for the international order in terms of the risks it poses, and a significant loss in terms of opportunities (Roscini, 2014: 2). *Cyber humanitarian intervention (CHI)* is also a concept that is quite important in this regard and should not be neglected. CHI can be defined as interventions using preventive cyberspace to prevent repressive regimes from committing crimes against humanity, such as genocide, ethnic cleansing, and discriminatory violence, against their own societies or against the people of another state (Güler, 2015: 139). Considering the dependence of the perpetrators of such crimes on digital platforms and online networks in directing their actions, planning, or seeking support today, the necessity of CHI becomes apparent.

Although R2P is an important principle in the UN, one of the main reasons it remains ineffective in the face of systematic human rights violations today is that the principle is often associated with a military response. However, if R2P can be implemented without the use of military force, preventing the crisis from escalating would result in a less complex process for both the target country and the intervening states. The UN has emphasised the importance of the prevention phase by publishing a report entitled *Framework of Analysis for Atrocity Crimes: A Tool for Prevention*. The report states that mass human rights violations generally occur in countries experiencing a certain level of instability or crisis, and that preventing the crisis from escalating to the point of requiring military intervention could avert not only loss of life but also physical, psychological and social trauma. On the other hand, the report states that the cost of prevention is lower than the cost of continuing crises and evaluates the limited options for preventive action (UN, 2012: 2). Therefore, one of the most important issues neglected in the literature on R2P is the question of what preventive interventions might be. Considering the negative aspects of technology that facilitates, deepens and covers up the aforementioned human rights violations, it may be appropriate to evaluate CHI as an antidote for preventive purposes.

One of the most fundamental operations of the CHI is undoubtedly to provide uninterrupted digital access to information in crisis areas. The ability to securely transmit and receive data, coordinate actions in real time, and maintain situational awareness in large and complex crisis areas is the cornerstone of the modern digital world. Without secure and resilient communications, even the most advanced autonomous systems and AI-powered platforms become isolated, vulnerable entities. Considering the possibility that perpetrators may deliberately damage communication infrastructure to avoid repercussions for their actions, it is essential that affected communities have access to internet-based communication channels to make their voices heard, demonstrate the depth of the crisis to the global public, and provide evidence of the elements of the actions.

The internal conflicts that took place in Libya in 2011 and ended Muammar Gaddafi's nearly half-century rule with his death are an example of the regime's blocking of communication channels. The regime had always sought to maintain its monopoly over the internet, blocking websites that produced content inconsistent with its policies or that were critical of it, and imposing harsh penalties on individuals who made critical comments. As a result of the

crackdown, internal unrest began in February 2011, leading to an internet blackout that lasted until August 2011. In Libya, only 17 per cent of the population had access to the internet due to high internet costs, while mobile phone ownership was widespread among almost the entire population. Consequently, the regime not only cut off internet access but also restricted access to CHI cards. After the regime's collapse, archives were found containing files on the online activities of Libyan dissidents communicating with foreigners (Freedom House, 2012).

Syria is another country where the cyber domain is controlled by regime leader Bashar al-Assad through public institutions such as the Syrian Telecommunications Establishment (STE). Although the number of Syrian citizens with internet access reached 4 million in 2010, the regime has always monitored user activity and required businesses such as internet cafes to record customer information and online activities. STE has utilised advanced technologies to block the public's phone calls, text messages, emails and internet access. In 2007, the regime introduced a nationwide surveillance system capable of actively monitoring the internet without individuals' knowledge, resulting in the procurement of devices capable of network filtering, blocking, and surveillance (Helwani, 2024: 249).

It is possible to multiply the policies implemented by repressive regimes to prevent their citizens from communicating with the outside world. What is important here is what steps countries that desire peace and wish to prevent crises will take in the face of repressive regimes. To prevent the blocking of the internet and other communication infrastructures in crisis areas by regime interventions and to ensure the healthy exchange of information, *satellites* can be considered within the scope of CHI. Currently, many states use this satellite technology within the scope of national cyber security. For example, the United States uses satellites for observation, communication and mapping. The Russia-Ukraine war has also highlighted the importance of satellites. In the war that began in February 2022 with Russia's invasion of Ukraine, Russian cyber attacks dealt a heavy blow to Ukraine's communications infrastructure, causing serious communication disruptions between army units and rendering military equipment that required network connectivity unusable. At this point, help for the Ukrainian army came from Starlink, the world's largest satellite constellation owned by SpaceX. With more than 20,000 Starlink terminals provided to Ukraine, the satellite became an indispensable communication infrastructure for the army (Abels, 2024: 843). Furthermore, Starlink satellites were utilised during the Los Angeles wildfire that began on 7 January 2025, replacing the damaged internet and communication infrastructure to ensure both firefighting

teams remained in contact and national and international media could broadcast from the disaster area (Conklin, 2025).

The second type of operation related to CHI could be the implementation of applications for *online surveillance* in crisis areas. Surveillance applications can be used to collect data from crisis areas in a digital environment and to identify threats and risks. Such applications include technologies such as advanced cyber intelligence systems, artificial intelligence-powered data analytics, and satellite imaging. Surveillance operations are critical in identifying atrocities faced by civilians, documenting human rights violations to increase the accountability of perpetrators in international courts, and enabling rapid intervention by the international community. In this context, for example, Zhengyang Hou and colleagues (2024: 1) use image processing techniques to detect destruction in civil war zones using satellite imagery, converting image pixels into information with an application they call PtNet and presenting it through a detection scheme called TKDS. The authors emphasise that real-time detection of damage that may occur in current and future countries due to civil unrest, earthquakes or extreme weather events is of vital importance.

During the Cold War, the primary purpose of surveillance satellites, whose importance grew, was to detect and classify rival states' nuclear-tipped missiles or submarines, warplanes, military equipment, and other communication infrastructure. However, with the technology of the previous century, images were exposed and captured on film, and it took days for the film rolls to reach experts and for the films to be developed. With the digital era, film rolls have been replaced by surveillance technologies with digital sensors that continuously capture images. In the following period, imaging radars with higher resolution capabilities, able to focus on a target, detect different radiation levels in the monitored area, and scan a wider area, were developed, such as Germany's SAR-Lupe, Italy's COSMO-SkyMed, Israel's TecSAR, China's YaoGan, and India's Cartosat-2. (Norris, 2011: 44-46).

The third CHI method could involve preventing social media and other internet-based posts containing hate speech and violent content in order to break the perpetrator's digital assault, and ensuring that supportive content for the victim is included. In the digital age, hate speech content increases social polarisation and can be a powerful factor capable of triggering crises of violence in societies. Violent rhetoric spreading through social media applications radicalises individuals, can lead to increased othering of minorities, and can disrupt social harmony. Therefore, blocking such content through cyber intervention is necessary to prevent

crises from escalating. Content blocking within the framework of CHI must be carried out with great care, as it treads a fine line between freedom of expression and the preservation of peace. CHI, which is not a censorship mechanism that restricts freedom of expression, should aim to combat disinformation, identify digital environments that encourage hate speech, and raise awareness both in victimised communities about the crisis and in other communities around the world about victimised communities. Digital literacy enhancement education programmes can also be beneficial in this regard within the scope of CHI.

The effect of violent content spreading rapidly on social media, thereby deepening crises, is clearly evident in the crimes against humanity committed by the Myanmar army against the Rohingya minority in 2017. The Myanmar army launched an ethnic cleansing operation against Rohingya Muslims, while Facebook, a social media application owned by Meta Technologies, encouraged and reinforced this ethnic cleansing with its algorithms. Radical Buddhist nationalist groups and Myanmar army personnel spread a great deal of misinformation on the app, claiming that Muslims would take over Myanmar as invaders in the near future. They shared photos of human rights activists defending the rights of the Rohingya people within the Myanmar population and threatened them with death. In a report published in 2022, Amnesty International acknowledged that Meta contributed to the atrocities in Myanmar with its dangerous algorithms for profit. Rohingya activist Mohammed Showwife accused Facebook of destroying the dreams of the Rohingya people, who aspire to live like everyone else (Amnesty International, 2022).

Violent content is not limited to social media. Looking further back in history, the 1994 Rwandan genocide confronts humanity. In attacks carried out by Hutus against Tutsis, approximately one million Rwandans were killed and two million people were forced to flee their country. In Rwanda, where two ethnic groups had coexisted peacefully in the past, the fact that ordinary civilians attempted to kill each other with any object they could find highlights the role of communication tools in triggering the genocide. After the death of President Habyarimana in a plane crash, the Hutus were gripped by the fear that the Tutsis would seize power and begin discriminatory activities. During this period, the Hutus used the radio to incite and direct the genocide. Radio broadcasts via Radio Télévision Libre des Milles Collines, calling on other Hutus to take action against the Tutsis, were constructed around memories such as Rwanda's colonial history, suggesting that the only way out of this cycle of the past was through genocide, triggering absolute violence between the two ethnic groups (Kellow and Steeves, 1998: 107).

The final alternative type of operation related to CHI may involve targeted cyber interventions aimed at terminating the perpetrators' actions. By incorporating both traditional warfare techniques and modern cyber space elements, it can directly damage the perpetrators' communication channels. As this method resembles a military intervention rather than a preventive measure, it also carries the risk of harming civilians. The detonation of radios used by Hezbollah by Israel on 18 September 2024 (Aljazeera, 2024), the 2010 Stuxnet Operation by the US targeting Iran's nuclear facilities, which damaged one-fifth of the gas centrifuges (Willett, 2024: 69), or Iranian hackers attempting to infiltrate rural water flow and wastewater treatment systems in Israel (Heller, 2020) are examples of direct, targeted cyber actions. Due to the potential for direct or indirect harm to civilians, their implementation is highly challenging.

The most successful example of a targeted cyber operation is Operation Glowing Symphony, conducted in 2016 by the US Cyber Command and the US National Security Agency. The primary objective of the operation was to target ISIS's global media operations and propaganda, destroying materials and disrupting its digital recruitment and financial activities. As part of the operation, ten accounts used by the organisation to spread its propaganda were listed and phishing emails were used. This allowed the operation teams to gain control, enabling them to freely navigate ISIS networks and plant malicious software on servers. First, ISIS networks were mapped, propaganda content was removed, and the organisation's propaganda methods, such as the Amaq Agency app, were blocked. The teams then moved on to creating technical errors and problems that would often appear to be IT issues, creating a psychological effect within the organisation, such as confusion, anger and deception. This forced the organisation's digital managers to use vulnerable and unreliable tools that would reveal their physical locations, making them targets for kinetic attacks (Raston, 2019: Cohen and Bar'el, 2017: 36).

There are fundamental similarities and differences between CHI and Traditional Humanitarian Intervention (THI). Firstly, both CHI and THI are based on international norms and aim to protect humanitarian values, relying on the obligation of states or international actors to intervene in the face of systematic human rights violations or crimes of mass atrocity. Due to the lack of sufficient interest in CHI in the literature on international law and international relations, there is no study that comprehensively outlines the differences and similarities between CHI and THI. However, considering the similarities and differences noted by Kallberg (2016: 84), it can be seen that the cyber warfare-conventional warfare

distinctions frequently studied in the disciplines are in line with the characteristics of CHI and THI.

CHIs, like cyber conflicts or wars, differ from THIs in terms of the environment in which they are created, the techniques and tools used, and their strategies. Firstly, CHIs promise to have a significant impact in weakening the pressure mechanisms of regimes or other criminal actors with their sudden and unexpected operational capacity. Since any operation undertaken using THI methods will require a specific preparation process, perpetrators can take a defensive position in the event of diplomatic signals or military movements. CHI breaks the control mechanisms of perpetrators, particularly those based on communication and information gathering, ultimately undermining their capacity to maintain pressure. Furthermore, physical boundaries are of no significance for CHI (Healey, 2016: 44-45).

Another fundamental characteristic that distinguishes CHI from THI is that human interventions in the digital environment involve much lower costs and risk rates compared to conventional methods. While military or direct kinetic interventions typically require significant economic investment, logistical support, and extensive operations, digital interventions can yield effective results even with limited resources. For example, according to a report by Richard Norton Taylor and Peter Capella (1999) in The Guardian, the cost of NATO's Operation Allied Force intervention in Kosovo exceeded £30 billion. While military operations involve numerous complex processes, such as the logistics of military units, the maintenance of air operations, and ensuring the safety of the civilian population in the region during the operation, CHI generally requires software-based strategies and thus requires fewer material resources. For example, providing a global VPN or encryption tools against the censorship practices of an oppressive regime is much less costly than air operations. Furthermore, since CHI does not require a physical presence in crisis areas, it does not carry the risk of conflict. In THI, intervention forces must be present on the ground and may therefore face situations such as becoming direct targets or being exposed to retaliatory attacks (Li and Liu, 2021: 8183-8184).

One of the fundamental reasons why CHI is less costly than THI is that CHI can be implemented by a much wider range of actors. THI is carried out by large, centralised and fixed actors, based on states sending their troops to military coalitions formed under the umbrella of international organisations. In CHI, however, in addition to the states that will implement digital interventions, there is the possibility that individuals, civil society

organisations, private companies and hacker groups may take part under the supervision of international and regional organisations. This diversity gives CHI the characteristics of speed, flexibility and low cost, while at the same time raising questions about who the implementing actor for CHI will be.

Any CHI action undertaken to weaken and eliminate the perpetrators' means of coercion in crisis regions should be the responsibility of states and regional/global organisations. The concentration of power in cyberspace by actors such as private companies or hackers, without state or organisational oversight, carries the risk of drawing states into a conflict zone, raising concerns that such cyber interventions will contribute to international instability rather than peace (Pattison, 2020: 251). However, states or organisations may employ private technology companies or individuals with cyber capabilities to carry out humanitarian interventions on their behalf. For example, the United States obtained surveillance opportunities during the Kosovo War through a contract with DynCorp, a private military contractor, to monitor the withdrawal of Serbian military forces from Kosovo. As can be seen, states or organisations can form a cyber humanitarian intervention team under their own identity, but they can also utilise the private sector and, in some cases, even opt for a hybrid structure (Rhiannon, 2021: 187).

The ethical and moral assessment of CHI's implementing actors within the context of international law is also important. In international law, the principle of jus ad bellum specifies when and under what conditions a state or the international community may legitimately resort to the use of force. Conventionally, the use of force between states is prohibited under Article 2(4) of the UN Charter, and this provision has become a jus cogens norm. The only exceptions to the prohibition on the use of force are situations approved by the UN Security Council and situations involving elements of legitimate defence. However, when considering CHI, it is possible that it could be evaluated within the framework of the jus ad bellum principle, as it does not involve the use of physical force, unlike traditional military interventions, and is carried out with the aim of deterring human rights violations by repressive regimes.

Another fundamental principle of international law is *jus in bello*, meaning that during wartime, the rules governing the conduct of war require that combatants respect human rights and civilians. Even if CHI does not involve kinetic operations but rather activities such as data collection or countering disinformation, the level and form of intervention must be

proportionate. For example, CHI actions that affect the health system or basic infrastructure of the country being intervened in may be considered an unacceptable violation under International Humanitarian Law. In this context, the principle of jus in bello requires careful consideration to ensure that CHI operations only target the perpetrators of atrocities and do not harm civilians.

From an ethical perspective, CHI can serve as a deterrent against human rights violations by oppressive regimes and can provide critical evidence for international justice mechanisms. However, the risk of such interventions being misused should not be overlooked. Cyber operations conducted unilaterally, particularly by certain states or international organisations, can be manipulated for political gain, even if they are claimed to be carried out for humanitarian reasons. Therefore, international oversight mechanisms and transparency principles must be implemented to ensure that CHI maintains its legitimacy within a legal and ethical framework.

Conclusion

This study has examined whether a cybersecurity-based humanitarian intervention can be situated within the framework of the Responsibility to Protect (R2P), addressing the question not merely at the level of technical feasibility or operational effectiveness, but through its normative, legal, and ethical constraints. The central finding of the analysis suggests that while the use of cyber technologies in humanitarian intervention may appear theoretically possible, such an approach remains in significant tension with the normative foundations upon which R2P is built. Consequently, the issue is less about whether a cyber intervention can be conducted, and more about whether such an intervention can be defined as legitimate, constrained, and genuinely protective within the scope of R2P.

The R2P doctrine conceptualizes the prevention of atrocity crimes as a collective responsibility, yet it deliberately leaves unresolved the question of which instruments may be legitimately employed to fulfil this responsibility. Proposals for cyber humanitarian intervention draw upon this ambiguity, presenting cyber operations as a seemingly less intrusive alternative to traditional military intervention and as a means of avoiding the political and humanitarian costs associated with kinetic force. However, this study demonstrates that the inherent characteristics of cyber operations—namely their opacity, difficulties of attribution, and indeterminate scope—risk undermining rather than reinforcing the principles of legitimacy, transparency, and accountability that R2P seeks to uphold.

A critical dimension of the research question concerns whether cyber humanitarian intervention genuinely serves the protection of civilians, or whether it merely transforms intervention into a more invisible and less regulated practice. Cyber operations may indeed contribute to civilian protection through early warning systems, information gathering, or the documentation of violations. Yet the same tools can easily be repurposed for operations that infringe upon state sovereignty, disrupt critical infrastructure, or generate wide-ranging indirect effects on civilian populations. This raises unresolved questions regarding how core R2P principles such as last resort and proportionality can be meaningfully applied in the cyber domain.

The study further reveals that the normative vacuum surrounding the legal status of cyber operations in international law effectively shifts cyber humanitarian intervention from a rule-based framework into a realm of political discretion. Given that R2P practices remain contested even in the context of conventional interventions, their extension into cyberspace - an arena characterized by blurred boundaries and limited accountability- risks facilitating the normalization of intervention under increasingly permissive conditions. In this sense, cyber humanitarian intervention may be interpreted not as an evolution of R2P, but as an indicator of its normative erosion.

Accordingly, the answer to the research question must be cautious and conditional. While cybersecurity-based intervention under R2P may be technically conceivable, it is difficult to argue that such interventions can presently be considered genuinely “humanitarian” within the existing international legal order and prevailing power structures. On the contrary, the discourse of cyber humanitarian intervention may function to lower the threshold for intervention and weaken mechanisms of accountability, particularly in the context of the digital reconfiguration of sovereignty.

In conclusion, rather than framing cyber humanitarian intervention as a normative advancement, this study positions it as a contested domain that exposes the inherent limitations and contradictions of the R2P doctrine. The incorporation of cyber technologies into humanitarian intervention can only be justified under conditions of clearly articulated norms, robust oversight mechanisms, and genuinely collective decision-making processes. Absent these safeguards, cyber humanitarian intervention risks becoming a legitimizing discourse for new, less visible forms of intervention, rather than a meaningful instrument for the protection of civilians.

References

Abels, J. (2024). Private Infrastructure in Geopolitical Conflicts: The Case of Starlink and The War in Ukraine. *European Journal of International Relations*, *30*(4), 842–866.

Akyeşilmen, N. (2018a). *Disiplinlerarası Bir Yaklaşımla Siber Politika & Güvenlik*. Ankara: Orion Kitabevi.

Akyeşilmen, N. (2018b). Cyber Good Governance: A New Challenge in International Power Politics?. *Cyberpolitik Journal*, *3*(5-6), 2-21.

Al Jazzeera. (2024). Hezbollah Walkie-talkies Exploded Too, What to Know about Israel's Attacks. https://www.aljazeera.com/news/2024/9/18/more-devices-exploding-across-lebanon-whats-happening

Alexander, K. (2024). The limits of international law: the Responsibility to Protect (R2P), Israel and the International Court of Justice. https://www.realinstitutoelcano.org/en/commentaries/the-limits-of-international-law-the-responsibility-to-protect-r2p-israel-and-the-international-court-of-justice/

Al-Rawi, A. (2016). Video Games, Terrorism, and ISIS's Jihad 3.0. *Terrorism and Political Violence*, *30*(4), 740–760.

Amnesty International. (2022). *Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – new report*. https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/ (Erişim Tarihi: 17.01.2025)

Badescu, C. G. (2011). *Humanitarian Intervention and the Responsibility to Protect: Security and Human Rights*. Abingdon: Routledge.

Balkin, J. M. (2012). The First Amendment is an Information Policy. *Hofstra Law Review*, *41*(1), 1-41.

BM. (2012). Framework of Analysis for Atrocity Crimes A Tool for Prevention. https://www.globalr2p.org/resources/framework-of-analysis-for-atrocity-crimes-a-tool-for-prevention/

Coady, C. A. J., Dobos, N. and Sanyal, S. (2018). *Challenges for Humanitarian Intervention: Ethical Demand and Political Reality*. Oxford: Oxford University Press.

Cohen, D. and Bar'el, O. (2017). The Use of Cyberwarfare in Influence Operations. *Yuval Ne'eman Workshop for Science, Technology and Security*. https://en-cyber.tau.ac.il/sites/cyberstudies-english.tau.ac.il/files/media_server/cyber%20center/cyber-center/Cyber_Cohen_Barel_ENG.pdf

Conklin, A. (2025). Los Angeles Wildfires: Elon Musk Personally Delivers Starlinks to California First Responders. https://www.foxnews.com/us/los-angeles-wildfires-elon-musk-personally-delivers-starlink-satellites-california-first-responders

Deng, F. (2010). From Sovereignty as Responsibility' to the Responsibility to Protect. *Global Responsibility to Protect*, *2*(4), 353-370.

Evans, G. (2008). *The Responsibility to Protect: Ending Mass Atrocity Crimes Once and For All*. Washington: Brookings Institution Press.

Freedom House. (2012). Freedom on the Net 2012 – Libya. https://www.refworld.org/reference/annualreport/freehou/2012/en/88761

Gilligan, E. (2013). Redefining Humanitarian Intervention: The Historical Challenge of R2P. *Journal of Human Rights*, *12*(1), 21–39.

Gulati, J. and Khosa, I. (2013). Humanitarian Inter Humanitarian Intervention: Tention: To Protect State So otect State Sovereignty. *Denver Journal of International Law & Policy*, *41*(3), 397-416.

Güler, A. (2015). İnsani Müdahale Aracı Olarak Siber Uzay. *Medeniyet Araştırmaları Dergisi*, *2*(4), 139-149.

Healey, J. (2016). Winning and Losing in Cyberspace. N.Pissanidis, H.Rõigas, M.Veenendaal (Eds.). *8th International Conference on Cyber Conflict: Defending the Core*. Talinn: CCD COE.

Heller, A. (2020). Israeli cyber chief: Major attack on water systems thwarted. https://www.aronheller.com/articles/israeli-cyber-chief-major-attack-on-water-systems-thwarted/

Helwani, I. (2024). Cyberactivism in Syria: Emergence, Transformation, Potentials, and Limitations. *Güvenlik Stratejileri Dergisi*, *20*(48), 239-263.

Henkin, L. (1999). Kosovo and The Law of "Humanitarian Intervention". *American Journal of International Law*, *93*(4), 824-828.

Hou, Z., Qu, Y., Zhang, L., Liu, J., Wang, F., Yu, Q., ... and Zhou, C. (2024). War City Profiles Drawn from Satellite Images. *Nature Cities*, *1*(5), 1-11.

ICISS. (2001*). The Responsibility to Protect*. https://www.globalr2p.org/resources/the-responsibility-to-protect-report-of-the-international-commission-on-intervention-and-state-sovereignty-2001/

Illingworth, R. (2024). Not the 'Fairest Norm of Them All' but Still Needed: On Hobson and Criticism of the Responsibility to Protect. *Journal of Intervention and Statebuilding*, *18*(2), 181–190.

Kallber, J. (2016). Humanitarian Cyber Operations. IEEE Tchnology and Society Magazine, https://cyber.army.mil/Portals/3/Documents/publications/external/Humanitarian%20Cyber%20Operations.pdf?ver=2017-09-26-135216-070

Kardaş, Ş. (2003). Humanitarian Intervention: A Conceptual Analysis. *Alternatives: Turkish Journal of International Relations*, *2*(3&4), 21-49.

Kellow, C. L and Steeves, H. L. (1998). The Role of Radio in the Rwandan Genocide. *Journal of Communication*, *48*(3), 107-128.

Li, Y. and Liu, Q. (2021). A Comprehensive Review Study of Cyber-Attacks and Cyber Security; Emerging Trends and Recent Developments. *Energy Reports*, *7*, 8176-8186.

Liff, A. P. (2012). Cyberwar: A New 'Absolute Weapon'? The Proliferation of Cyberwarfare Capabilities and Interstate War. *Journal of Strategic Studies*, *35*(3), 401–428.

Lin, H. (2019). The Existential Threat from Cyber-Enabled Information Warfare. *Bulletin of the Atomic Scientists*, *75*(4), 187–196.

Massingham, E. (2009). Military Intervention for Humanitarian Purposes: Does The Responsibility to Protect Doctrine Advance the Legality of the Use of Force for Humanitarian Ends? *International Review of the Red Cross*, *91*(876), 803-831.

Neilsen, R. (2021). Cyber Humanitarian Interventions: the Viability and Ethics of Using Cyber-operations to Disrupt Perpetrators' Means and Motivations for Atrocities in the Digital Age. PhD Thesis, UNSW Australia.

Norris, P. (2011). Developments in High Resolution Imaging Satellites for the Military. *Space Policy*, *27*, 44-47.

Norton, R. and Capella, P. (1999). *Bill for Kosovo war goes over £30bn*. https://www.theguardian.com/world/1999/oct/15/balkans

Paterson, T. and Hanley, L. (2020). Political Warfare in The Digital Age: Cyber Subversion, Information Operations and 'Deep Fakes.' *Australian Journal of International Affairs*, *74*(4), 439–454.

Pattison J. (2020). From Defence to Offence: the Ethics of Private Cybersecurity. *European Journal of International Security*, *5*(2): 233-254.

Pellerin, C. (2011). *Dod Releases First Strategy for Operating in Cyberspace*. https://www.army.mil/article/61720/dod_releases_first_strategy_for_operating_in_cyberspace

Raston, D. T. (2019). How the U.S. Hacked ISIS. https://www.npr.org/2019/09/26/763545811/how-the-u-s-hacked-isis?t=1582468821601

Rid, T. (2011). Cyber War Will Not Take Place. *Journal of Strategic Studies*, *35*(1), 5–32.

Roberts, G. W. (2000). Humanitarian Intervention: Definitions And Criteria. *CSS Strategic Briefing Papers*, *3*(1), 1-2.

Roscini, M. (2014). *Cyber Operastions and The Use of Force in International Law*. Oxford: Oxford University Press.

Stone, J. (2012). Cyber War Will Take Place! *Journal of Strategic Studies*, *36(*1), 101–108.

Thomas, T. (2014). Russia's Information Warfare Strategy: Can the Nation Cope in Future Conflicts? *The Journal of Slavic Military Studies*, *27*(1), 101–130.

Uitermark, J. (2016). Complex Contention: Analyzing Power Dynamics within Anonymous. *Social Movement Studies*, *16*(4), 403–417.

United Nations General Assembly. (2009). Implementing the Responsibility to Protect. *Report of the Secretary-General*, https://documents.un.org/doc/undoc/gen/n09/206/10/pdf/n0920610.pdf

Wheeler, N. J. (2000). *Saving Strangers: Humanitarian Intervention in International Society*. Oxford: Oxford University Press.

Willett, M. (2024). A Short History of Cyber Operations. *Adelphi Series*, *64*(511–513), 63–104.

ARTIFICIAL INTELLIGENCE IN EDUCATION: REGULATION, ETHICS, AND SECURITY

Carmen-Gabriela BOSTAN [*]

Declaration[*]

Abstract

The integration of artificial intelligence (AI) into e-learning environments is transforming educational practices by enabling personalized learning pathways, automating assessment, and enhancing administrative efficiency. While these innovations offer significant pedagogical benefits, they also raise complex ethical and security concerns. This paper explores the implications of AI use in digital education, focusing on algorithmic transparency, data protection, and institutional accountability. It examines how AI systems influence decision-making processes in teaching and learning, and highlights the need for clear public policies to regulate their deployment. Drawing on international best practices and case studies from countries such as Finland, Estonia, and Romania, the study identifies key strategies for responsible AI implementation. These include teacher training in digital ethics, risk-based governance frameworks, and mechanisms for human oversight of algorithmic decisions. The paper also discusses the importance of transparency policies, such as algorithmic audit protocols and public registries, to ensure that AI systems operate fairly and explainable. By adopting an interdisciplinary approach that combines digital pedagogy, ethical standards, and legal safeguards, the research advocates for a sustainable and inclusive model of AI integration in education. The findings underscore the urgency of aligning technological innovation with democratic values and human rights, ensuring that AI serves as a tool for empowerment rather than control in the learning process.

Keywords: Artificial Intelligence (AI), AIED, Educational Policies, Pedagogical Innovation, Ethics

Introduction

The rapid integration of artificial intelligence (AI) into e-learning platforms marks a significant transformation in the organization of teaching, learning, and assessment. AI-driven systems personalize instructional content, automate evaluation, and optimize administrative

[*] cagabosro@gmail.com, Institute of Educational Sciences, Romania
[*] This article has been prepared without the use of any Artificial Intelligence (AI) tools or assistance.

processes, thereby expanding access to educational resources and enabling more flexible learning environments (Holmes & Tuomi, 2022). However, these algorithmic systems also shape students' learning trajectories and influence teachers' pedagogical decisions, raising questions about transparency, fairness, and accountability.

European policy documents emphasize that algorithmic transparency is essential for maintaining trust in digital education. Teachers, students, and parents must understand how AI systems operate, what data they process, and what limitations they entail (European Commission, 2022). Without such transparency, automated decisions risk becoming opaque, potentially reinforcing biases or undermining pedagogical autonomy. The European Union's AI Act classifies educational AI systems as "high-risk," requiring strict standards of auditability, documentation, and human oversight (European Commission, 2023). Complementing this, the Council of Europe's Framework Convention on AI (2024) introduces legally binding obligations for risk assessment, data protection, and stakeholder participation.

Despite these advances, national regulatory frameworks—particularly in countries such as Romania—remain fragmented. The absence of clear guidelines on algorithmic governance, data protection in educational contexts, and teacher training creates uncertainty for institutions seeking to adopt AI responsibly (Adăscăliței, 2025). This paper addresses these gaps by analyzing international good practices and proposing a coherent framework for ethical and secure AI integration in education.

Methodology

This study employs a qualitative research design grounded in documentary and comparative analysis. The methodological approach integrates multiple sources of evidence to construct a comprehensive understanding of the ethical, legal, and pedagogical implications of AI adoption.

Data Sources

The analysis draws on:

- **Academic literature** on AI in education, algorithmic ethics, and digital pedagogy (Holmes & Tuomi, 2022; Bostan, 2024).

- **Policy documents** including the EU AI Act (European Commission, 2023), the Council of Europe Convention (2024), and UNESCO reports (2019, 2025).
- **Empirical surveys** on teachers' perceptions of AI integration (Sămărescu et al., 2024).
- **Case studies** of AI-driven educational platforms such as Carnegie Learning, eKool, Ariadna, and Școala Nouă.

Analytical Strategy

The analysis identifies recurring themes in policy documents and academic literature, contrasts national approaches to AI governance, and synthesizes insights from international case studies. The comparative dimension highlights divergences between countries with mature digital ecosystems (Finland, Estonia) and those at earlier stages of development (Romania).

Research Objectives

The study is designed to:

- **Map the regulatory landscape** of AI in education at national and international levels.
- **Identify ethical challenges** related to transparency, data protection, and and algorithmic bias in AI-assisted learning environments.
- **Assess cybersecurity risks** associated with the use of AI in e-learning platforms.
- **Formulate policy recommendations** for equitable, responsible, and sustainable AI governance in education.

Legislative context

The regulation of the use of AI in education is at an early stage in Romania, being influenced by the European directives on artificial intelligence and data protection. The draft European Regulation on Artificial Intelligence (AI Act) proposes to classify AI applications according to the risk they pose, and AI-based educational systems are considered "high risk", requiring transparency, auditability and protection of personal data (European Commission, 2023). In parallel, national legislation on digital education and cybersecurity is fragmented, without a unitary framework explicitly regulating the use of AI in schools and universities.

Theoretical framework

The theoretical foundation of this study rests on an interdisciplinary synthesis that integrates educational, ethical, and governance perspectives. Four complementary pillars provide the conceptual scaffolding: *AI-assisted personalized learning*, *algorithmic ethics*, *digital governance in education*, and *digital pedagogy*. Together, these dimensions enable a comprehensive analysis of both the opportunities and risks associated with artificial intelligence in e-learning environments.

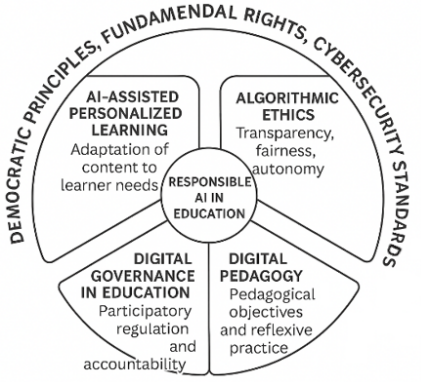

Figure: 1 A visual diagram of this conceptual framework

AI-Assisted Personalized Learning

AI systems can adapt instructional content to learners' needs through recommendation algorithms, predictive analytics, and adaptive feedback, enhancing engagement and supporting differentiated instruction (Holmes & Tuomi, 2022). By leveraging recommendation systems, predictive analytics, and adaptive feedback, AI can foster inclusivity and differentiated instruction. This theoretical perspective highlights the pedagogical potential of AI to enhance engagement and improve learning outcomes.

Algorithmic Ethics

Algorithmic ethics examines the moral implications of automated decision-making, including risks of bias, opacity, and reduced student autonomy. Ethical deployment requires transparency, fairness, and mechanisms for human oversight (European Commission, 2022). Central concerns include the risk of bias embedded in algorithms, the opacity of decision-making processes, and the potential erosion of student autonomy. Ethical analysis in this domain underscores the necessity of transparency, accountability, and fairness in the design and deployment of educational technologies.

Digital Governance in Education

Digital governance highlights the need for participatory regulatory models involving policymakers, educators, and technology developers (Adăscăliței, 2025). Effective governance must integrate democratic principles, accountability mechanisms, and clear institutional responsibilities.

Digital Pedagogy

Digital pedagogy argues that technological innovation must be guided by pedagogical objectives rather than technological enthusiasm. Teachers must critically assess how AI tools influence cognitive development, learning processes, and student autonomy (Bostan, 2024). Digital pedagogy insists that the integration of emerging technologies must be guided by explicit pedagogical objectives rather than enthusiasm for innovation. It calls for a reflexive approach to understanding how AI tools influence cognitive development, learning processes, and student autonomy. Teachers, therefore, must consciously decide when and how to employ AI, considering its impact on both intellectual growth and emotional formation.

Integrated Lens

By combining these four pillars, the study adopts an *interdisciplinary lens* that bridges educational theory, ethical reflection, and governance frameworks. This integrated approach ensures that the analysis captures the full spectrum of implications—pedagogical, ethical, and legal—while promoting a vision of AI in education that respects democratic values, fundamental rights, and cybersecurity standards.

Results: International Good Practices in AI for E-Learning

To understand the concrete impact of artificial intelligence (AI) in e-learning, it is essential to analyze international good practices that highlight both pedagogical benefits and ethical and security challenges. This section presents a comparative analysis of AI implementation in education across various national and international contexts, focusing on policy frameworks, technological innovations, and practical applications.

European Union: Ethical Guidelines for Teachers. The European Union has taken proactive steps to guide educators in the ethical use of AI. The 2022 guidelines emphasize algorithmic transparency, data protection, and the development of digital competencies among teachers (European Commission, 2022. Concrete examples include AI applications in language

learning, personalized task assignment, and automated assessment. These initiatives promote a culture of responsibility and critical reflection, encouraging educators to engage with AI tools ethically and effectively.

United States: Carnegie Learning's Smart Tutoring System. Carnegie Learning exemplifies the use of AI in personalized mathematics instruction. Its smart tutoring system provides real-time feedback by analyzing student responses and behavior. While studies report improved academic performance, concerns persist regarding algorithmic transparency and data privacy (Holmes & Tuomi, 2022). This case underscores the need for balancing technological innovation with ethical safeguards.

Finland: National Strategy for Digital Education. Finland's comprehensive strategy integrates AI through adaptive learning platforms, continuous teacher training, and regular impact assessments. Initiatives such as EduCluster Finland and EduExcellence offer immersive learning environments and professional development programs. These platforms connect education with professional life, fostering holistic learning and safe technology adoption. Finland's approach demonstrates how national policy can align technological advancement with pedagogical integrity (UNESCO, 2025).

Estonia: eKool Platform. Estonia's eKool platform showcases advanced digital infrastructure and AI integration. It facilitates secure data access, predictive analytics for dropout risk, and real-time communication among students, parents, and educators. Despite its effectiveness, critiques highlight risks of stigmatization and surveillance, emphasizing the importance of robust privacy regulations. eKool's interoperability and user-centric design offer valuable insights into scalable AI deployment (Council of Europe, 2024).

Romania: Ariadna and Scoala Noua Platforms

The *Ariadna* experiment, coordinated by Romanian experts in digital education, proposes: using AI for personalized instruction, automated grading, and data-driven decision-making; addressing challenges related to privacy, algorithmic bias, and technological costs, as well as promoting a critical pedagogy that encourages discernment in the use of AI (Istrate, 2022). The Ariadna project, developed within the Digital Education platform, explores the use of AI in automated assessment and instructional personalization. Teachers were involved in testing tools such as ChatGPT and Grammarly in teaching activities. The study highlighted the need

for ethical and digital training, as well as the risks related to excessive reliance on algorithms in educational decision-making

The *Școala Nouă* platform in Romania (https://scoalanoua.ro/offer) is a significant digital tool designed to enhance the educational experience by providing a comprehensive digital catalog that allows real-time tracking of student grades and facilitates seamless communication between parents and teachers. Beyond its core functionalities, *Școala Nouă* integrates artificial intelligence-based features designed to personalize learning paths, identify student strengths and weaknesses, and provide personalized recommendations to support academic development. The platform supports educators in efficiently managing and automating administrative tasks as well as promoting a collaborative environment between students, teachers, and parents. Its user-friendly interface and accessibility contribute to increased transparency and inclusion within the educational ecosystem. Furthermore, *Școala Nouă* emphasizes data privacy and security, aligning with national regulations to protect students' sensitive information. The continuous development of the platform includes incorporating feedback from educators and stakeholders to adapt to evolving educational needs and technological advances.

In addition to *Școala Nouă*, Romania benefits from complementary digital solutions such as *ADMA* and *Edu Apps*, which further support the management of online courses and the development of digital infrastructure. Together, these platforms contribute to building a modern, inclusive and transparent education system that uses artificial intelligence responsibly to improve learning outcomes and stakeholder engagement. The ADMA application, (managed by Edu Apps on the Google Cloud Platform within the European Union) together with Future Class (Clasa Viitorului): Google Workspace for Education and Office 365 A1 (hereinafter referred to as "the Applications"), are available free of charge to students, parents, and teachers. These applications enable each teacher to manage classes and lessons online through a professional account created by the school. Edu Apps (https://www.eduapps.ro/) develops and implements solutions for innovation in education for any level of education such as: applications (school management, teaching and learning), equipping the school space with devices and the Internet, training and consulting services.

Comparative Insights

Across these contexts, several common themes emerge:

- Pedagogical Innovation: AI enhances personalization, feedback, and engagement.
- Ethical Challenges: transparency, bias, and data protection remain critical concerns.
- Policy and Governance: national strategies and international guidelines shape responsible AI use.
- Teacher Empowerment: continuous training and ethical literacy are essential for effective implementation.

Discussions

The documentary analysis highlights several essential dimensions regarding the integration of artificial intelligence (AI) in education. Empirical studies indicate that teachers show increasing openness toward adopting AI, particularly in assessment practices and personalized learning, provided that adequate training frameworks and professional support are available. This underscores the importance of continuous professional development as a prerequisite for effective and responsible AI implementation.

At the same time, significant legislative gaps persist both internationally and nationally. Currently, no specific regulatory framework fully addresses the use of AI in education. In Romania, existing legislation on data protection and cybersecurity does not sufficiently cover the specificities of educational algorithms, generating uncertainty for schools and learners. This situation reinforces the need for coherent public policies that ensure transparency, accountability, and equity in AI-assisted educational environments.

Ethical risks also remain central. The opacity of algorithmic decision-making, the potential for automated bias, and challenges related to equitable access to digital technologies require an interdisciplinary approach that combines technological expertise with ethical and pedagogical principles. Ensuring fairness, explainability, and human oversight is essential for maintaining trust in AI-supported learning processes.

International models provide valuable insights for addressing these challenges. Finland and Estonia have adopted national strategies that explicitly integrate AI into education, emphasizing teacher training, transparent governance, and clear regulation of high-risk applications. By contrast, Romania is at an early stage of development: although interest among teachers and institutions is growing, the absence of a coherent legislative framework limits the potential for responsible and sustainable AI integration. Adapting international good

practices to the national context can support the development of a balanced ecosystem that promotes innovation while safeguarding fundamental rights.

Overall, the successful integration of AI in e-learning requires robust data protection measures, algorithmic transparency, and the active involvement of teachers in evaluating and adapting technologies. These elements are essential for ensuring that AI contributes positively to educational ecosystems and supports a safe, equitable, and sustainable digital transformation.

The Need for Responsible Integration of AI in E-Learning

The case studies analyzed demonstrate that the integration of artificial intelligence in e-learning must be approached responsibly, with attention to both pedagogical value and potential risks. AI systems can enhance personalization, feedback, and learning efficiency, but their benefits depend on transparent design, ethical use of data, and mechanisms that ensure human oversight. Without these safeguards, automated decisions may become opaque, reinforce biases, or undermine trust in digital learning environments. Responsible integration therefore requires clear institutional policies, robust data protection measures, and continuous monitoring of algorithmic impact.

The Romanian educational context illustrates clearly why AI integration must be approached responsibly. Although digitalization has accelerated in recent years, schools continue to face disparities in infrastructure, uneven access to technology, and limited institutional capacity to evaluate AI tools. These conditions make the risks associated with opaque algorithms, data misuse, or unregulated automated decision-making particularly acute. Responsible integration in Romania therefore requires not only technical safeguards but also clear national guidelines, transparent procurement processes, and mechanisms that ensure that AI supports pedagogical goals rather than replacing professional judgment.

Teachers' Openness and Professional Development

Empirical evidence shows that teachers are increasingly open to adopting AI tools, particularly for assessment and personalized instruction. However, this openness is conditional on the availability of adequate training and institutional support. Educators need opportunities to develop both technical competencies and ethical literacy, enabling them to

critically interpret algorithmic outputs and make informed pedagogical decisions. Professional development programs that address digital ethics, data protection, and AI-supported instructional design are essential for ensuring that teachers can use these technologies effectively and responsibly.

Romanian teachers demonstrate growing openness toward AI, especially in assessment, feedback, and personalized learning. However, this openness is strongly conditioned by their confidence in digital skills and the availability of structured training programs. Current professional development opportunities remain fragmented, and many educators report uncertainty about how to evaluate AI outputs or how to use such tools ethically. Strengthening teacher training in Romania—through national programs focused on digital ethics, data protection, and AI literacy—is essential for ensuring that educators can critically and effectively integrate AI into their practice.

Legislative Gaps and Policy Challenges

The ethical implications of AI in education remain a central concern. Key risks include algorithmic opacity, potential bias in automated decisions, and unequal access to digital technologies. Addressing these challenges requires an interdisciplinary approach that integrates technological expertise with ethical, legal, and pedagogical perspectives. Ensuring fairness, explainability, and respect for student autonomy must be core principles guiding the design and deployment of AI systems. Ethical governance frameworks should also include mechanisms for appeal, human review, and transparent communication with stakeholders.

Romania lacks a coherent legislative framework regulating the use of AI in education. Existing laws on data protection and cybersecurity offer only partial coverage and do not address the specificities of educational algorithms, such as transparency requirements, auditability, or the governance of high-risk systems. This regulatory vacuum places schools in a vulnerable position, as they must navigate complex ethical and legal issues without clear national guidance. Developing a unified policy framework—aligned with European standards but adapted to local realities—is crucial for ensuring safe and equitable AI adoption in Romanian education.

Ethical Risks and Interdisciplinary Approaches

Ethical risks remain central to the debate on AI integration in education. The most significant concerns include the opacity of algorithmic decision-making, the potential for automated bias,

and persistent inequalities in access to digital technologies. These risks highlight the need for an interdisciplinary approach that brings together technological expertise, ethical reflection, and pedagogical judgment. Algorithmic ethics emphasizes transparency, fairness, and explainability, while digital pedagogy insists that AI tools must serve clearly defined educational objectives rather than replace human decision-making.

In Romania, ethical challenges are amplified by disparities in digital infrastructure and varying levels of institutional preparedness. Limited teacher training, inconsistent data governance practices, and the absence of national ethical guidelines increase the likelihood that AI systems may reinforce existing inequities or generate opaque automated decisions. Addressing these vulnerabilities requires coordinated efforts to establish ethical standards for AI use in schools, including mechanisms for human oversight, transparent communication with stakeholders, and clear procedures for contesting algorithmic outcomes.

Insights from International Models

International models provide valuable insights into how AI can be responsibly integrated into education.

- Finland has adopted a national strategy that explicitly integrates AI, emphasizing teacher training, adaptive platforms, and equity assessments. Its initiatives, such as Virtual Life Labs, connect education with professional life, ensuring that AI serves both pedagogical and societal goals.
- Estonia has developed advanced digital infrastructure through platforms like eKool, which enable secure data access, predictive analytics, and real-time communication. While effective, Estonia's model also highlights risks of surveillance and stigmatization, underscoring the need for privacy safeguards.
- Romania, by contrast, is at an early stage of development. Projects such as Ariadna and Scoala Noua demonstrate growing interest among teachers and institutions, but the absence of a coherent legislative framework limits the potential for responsible integration.

The following table synthesizes the main differences across the three national contexts.

Table: 1 Comparative Table: AI in Education – Finland, Estonia, Romania

Dimension	Finland	Estonia	Romania
Policy Framework	National Strategy for Digital Education explicitly integrates AI; strong emphasis on teacher training and equity.	Advanced digital infrastructure supported by national e-governance strategy; AI embedded in platforms like eKool.	No coherent national AI strategy in education; reliance on general data protection and cybersecurity laws.
Teacher Training	Continuous professional development in AI and digital ethics; immersive programs (EduCluster, EduExcellence).	Training focused on digital literacy and platform use; less emphasis on ethics.	Limited training; pilot projects (Ariadna) highlight need for ethical and digital literacy development.
Platforms & Tools	Adaptive learning platforms, Virtual Life Labs, immersive simulations; strong link to professional readiness.	eKool platform integrates AI for monitoring progress, dropout prediction, and communication.	Scoala Noua (digital catalog), ADMA/Edu Apps for class management; Ariadna project tests AI tools like ChatGPT.
Ethical Safeguards	Regular assessment of AI's impact on equity and inclusion; clear guidelines for responsible use.	Privacy concerns raised over surveillance and stigmatization; ongoing debate on safeguards.	Ethical risks identified (bias, opacity, overreliance); lack of clear national guidelines.
Cybersecurity Measures	Integrated into national strategy; proactive monitoring of risks.	Secure access via digital ID; strong interoperability but vulnerable to surveillance risks.	General cybersecurity laws apply; insufficient coverage of educational AI systems.
Stage of Development	Mature, strategic, and holistic integration of AI in education.	Advanced infrastructure with strong digital identity systems; AI widely adopted.	Early stage; fragmented initiatives, strong interest but limited regulation and coherence.

Comparative analysis suggests that Romania could benefit significantly from adapting international best practices to its local context. This would involve not only importing technological solutions but also embedding them within a framework of ethical reflection, teacher empowerment, and policy coherence.

Building a Balanced Ecosystem

The overarching lesson from these case studies is that AI in education must be integrated within a balanced ecosystem. This ecosystem should promote innovation while safeguarding fundamental rights. Key components include:

- Coherent policies that regulate high-risk applications and ensure accountability.
- Adequate teacher training that combines technical skills with ethical literacy.
- Active involvement of all educational actors, including students, parents, policymakers, and technologists.
- Continuous evaluation of AI's impact on equity, inclusion, and academic integrity.

Such an ecosystem would not only enhance learning outcomes but also strengthen trust in digital education, ensuring that AI contributes to sustainable development rather than exacerbating existing inequalities.

For Romania, building a balanced AI-supported educational ecosystem requires coordinated efforts across multiple levels: national policy, institutional governance, teacher training, and community engagement. AI tools must be evaluated not only for their technical performance but also for their pedagogical relevance and ethical implications. Ensuring equitable access to digital resources, strengthening cybersecurity, and promoting a culture of critical engagement with technology are essential steps toward a sustainable digital transformation. When these conditions are met, AI can contribute to a more inclusive and learner-centered Romanian education system.

Conclusion of the Discussion

The discussion highlights that the integration of AI in education involves interconnected challenges related to ethics, governance, teacher preparedness, and infrastructural disparities. International models demonstrate that coherent strategies, transparent regulation, and sustained professional development can support responsible AI adoption. In Romania, the

growing interest in AI-assisted learning is accompanied by significant vulnerabilities, including legislative gaps, uneven digital infrastructure, and limited training opportunities. Addressing these issues requires coordinated efforts that balance innovation with ethical safeguards and institutional accountability. Overall, the discussion shows that AI can contribute meaningfully to educational transformation only when embedded within a transparent, equitable, and pedagogically grounded framework.

Conclusion

The analysis conducted in this study demonstrates that artificial intelligence has the potential to reshape educational ecosystems by enhancing personalization, supporting data-informed teaching, and improving administrative efficiency. However, the transformative promise of AI can only be realized when technological innovation is embedded within a coherent framework of ethical safeguards, transparent governance, and pedagogical purpose. The comparative examination of Finland, Estonia, and Romania reveals that the maturity of national digital strategies, the robustness of regulatory frameworks, and the quality of teacher training significantly influence the extent to which AI contributes to equitable and sustainable educational development.

Finland and Estonia illustrate how long-term investments in digital infrastructure, continuous professional development, and clear national strategies can create favorable conditions for responsible AI integration. Their experiences show that AI adoption is most effective when aligned with democratic values, human rights, and participatory governance. These models also highlight the importance of transparency, explainability, and systematic evaluation of AI tools—elements that help maintain trust and ensure that automated systems support, rather than constrain, pedagogical autonomy.

In contrast, Romania's emerging initiatives reveal both opportunities and vulnerabilities. Teachers and institutions display increasing interest in AI-assisted learning, yet the absence of a dedicated national strategy, limited ethical guidelines, and uneven digital infrastructure create significant barriers to responsible implementation. The Romanian context underscores the need for coordinated national policies that address algorithmic transparency, data protection, and the governance of high-risk educational systems. Strengthening teacher training in AI literacy and digital ethics is equally essential, as educators play a central role in mediating the pedagogical and ethical implications of AI tools.

The findings of this study suggest that the successful integration of AI in education requires a holistic, human-centered approach. This includes:

- **robust regulatory frameworks** that ensure accountability and protect fundamental rights,
- **continuous professional development** that empowers teachers to critically evaluate and adapt AI tools,
- **ethical governance mechanisms** that promote fairness, transparency, and explainability,
- **equitable access to digital infrastructure**, ensuring that technological innovation does not exacerbate existing inequalities.

Ultimately, AI should not be viewed as a substitute for human judgment but as a tool that can enhance the quality of teaching and learning when used responsibly. By aligning technological capabilities with pedagogical objectives and ethical principles, educational systems can harness the potential of AI to support inclusive, transparent, and future-oriented learning environments. For Romania, this alignment represents both a challenge and an opportunity: a chance to build a coherent, equitable, and sustainable digital education ecosystem that reflects European standards while responding to local needs.

Limitations of the Study

Despite its comprehensive documentary and comparative approach, the study presents several limitations that should be acknowledged. First, the analysis relies exclusively on secondary sources, without incorporating primary empirical data such as interviews, classroom observations, or surveys with educators and students. This limits the ability to capture lived experiences and contextual nuances regarding AI adoption in educational settings. Second, the focus is predominantly on European policy frameworks and case studies, which may reduce the generalizability of the findings to non-European contexts with different regulatory, cultural, or infrastructural conditions. Third, the rapidly evolving nature of AI technologies and regulations means that some policy references may become outdated, requiring continuous updates. Future research could address these limitations by integrating empirical fieldwork, expanding the geographical scope, and conducting longitudinal analyses of AI implementation in education.

Policy Recommendations

Governance and Regulation

- Develop a *coherent national strategy* for AI in education, aligned with European frameworks (AI Act, Council of Europe Convention).
- Classify educational AI systems as *"high risk"*, requiring audits, transparency reports, and accountability mechanisms.
- Establish *interdisciplinary regulatory bodies* including educators, policymakers, technologists, and civil society.
- Mandate *continuous evaluation* of AI's impact on equity, inclusion, and academic integrity, with public reporting.

Teacher Training and Professional Development

- Integrate *AI literacy and digital ethics* into teacher education programs.
- Provide *continuous professional development* through workshops, online modules, and immersive labs.
- Empower teachers as *co-designers of AI tools*, involving them in testing and adaptation.
- Develop *ethical classroom guidelines* to help teachers decide when and how AI should be applied.

Data Protection and Cybersecurity

- Implement *strict data protection protocols* tailored to educational contexts, beyond general GDPR compliance.
- Require *explainable AI systems* to ensure transparency in algorithmic decision-making.
- Introduce safeguards against *surveillance and stigmatization*, limiting predictive analytics to supportive interventions.
- Invest in *secure digital infrastructure*, including encryption and robust identity verification systems.

References

Adăscăliței, A. (2025), *Introducere în utilizarea inteligenței artificiale în educație*, Polirom, Iasi, Romania.

Bostan, C.G. (2024), *Digital Pedagogy, a New Trend in Education*, CNPEE-UCE, Bucharest, https://ibn.idsi.md/sites/default/files/imag_file/9-17_19.pdf

Holmes, W., Tuomi, I., (2022), *State of the art and practice in AI in education*, European Journal of Education, 00, 1–29. https://doi.org/10.1111/ejed.12533, State of the art and practice in AI in education; European Commission Joint Research Centre, https://publications.jrc.ec.europa.eu/repository/handle/JRC128299

European Commission, (2022), *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for Educators*, Publications Office of the EU, https://op.europa.eu/en/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1

Council of Europe, (2024), *Council of Europe Framework Convention on artificial intelligence and human rights, democracy, and the rule of law* (CETS No. 225), https://www.coe.int/ro/web/chisinau/-/council-of-europe-making-history-with-a-first-ever-legally-binding-global-treaty-on-ai-and-human-rights-with-republic-of-moldova-amid-signers;

European Commission, (2023), *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Istrate, O., Velea, S., Ștefănescu, D., (2022), *Ariadna Experiment. The Role of Artificial Intelligence in Education Sciences*, Journal of Digital Pedagogy, https://digital-pedagogy.eu/ariadna-experiment-the-role-of-artificial-intelligence-in-education-sciences/

Negrea-Busuioc, E., Zbuchea, A., Costea, A.-M., (2024), *Ghid de utilizare a instrumentelor de inteligență artificială în procesele academice*, AI ETHICS Project, https://snspa.ro/wp-content/uploads/2024/12/Ghid-de-utilizare-a-instrumentelor-AI-in-procesele-academice.pdf

Sămărescu, N., Bumbac, R., Zamfiroiu, A., Iorgulescu, C.M., (2024), *Artificial Intelligence in Education: Next-Gen Teacher Perspectives*, Amfiteatru Economic, 26(65), 145–161, https://doi.org/10.24818/EA/2024/65/145

UNESCO, (2019), *Artificial intelligence in education: challenges and opportunities for sustainable development*, https://unesdoc.unesco.org/ark:/48223/pf0000366994.locale=en

UNESCO, (2025), *AI and the future of education. Disruptions, dilemmas and directions*, https://www.unesco.org/en/articles/ai-and-future-education-disruptions-dilemmas-and-directions

AI-DRIVEN DISINFORMATION AS A GLOBAL CYBERSECURITY THREAT TO DEMOCRATIC SYSTEMS

Murat EMEÇ[*]
ORCID ID: 0000-0002-9407-1728

Declaration[*]

Abstract

The proliferation of generative artificial intelligence has fundamentally transformed our current information landscape, facilitating the widespread production and consumption of photorealistic yet fictitious media. So, misinformation is no longer just a communication issue; it is an absolute cybersecurity threat, especially for democracies that depend on trust, transparency, and an informed public. This paper views AI-fueled disinformation as a cognitive cyber-attack aimed at modifying perceptions, shaping belief formation, and undermining the legitimacy of institutions, rather than solely targeting technical systems. Based on literature in cybersecurity, political communication, and AI governance, the research investigates how generative AI augments disinformation, which systemic failures in democracies it exploits, and why current interventions fall short. The results, he says, are that AI-supported disinformation damages public trust, increases social and political fragmentation, and distorts electoral and governmental processes in ways that are often hard to detect and even harder to put right. The research suggests expanding existing cybersecurity strategies to protect democracies in the age of generative AI by considering not only information integrity and cognitive security but also societal resilience to disinformation and propaganda, as well as technical protections.

Keywords: AI-driven disinformation, Cybersecurity, Democratic Systems, Cognitive Security, Generative Artificial Intelligence

[*] Computer Science Application and Research Centre, Istanbul University, murat.emec@istanbul.edu.tr
[*] AI tools were utilized to enhance linguistic clarity, refine structure, and support the overall organization of the manuscript. All conceptual insights, analytical interpretations, and final conclusions were independently developed and verified by the author. All analysis and conclusions are solely the author's responsibility.

Introduction

The accelerating advance of artificial intelligence has decidedly altered the world's information landscape, redefining how it is produced, distributed, and understood. Whilst AI-based technology offers great promise in automation, communication, and decision-making, it also introduces new and pervasive risks to the integrity of the informational fabric of digital spaces. Chief among these risks is AI-fueled disinformation, one of the most disruptive and potentially damaging threats, especially to democratic systems that require informed citizens, transparent institutions, and trust-based governance.

The difference between AI-fueled disinformation and previous varieties of misinformation is the former's scale and its ability to adapt in real time — like Russia's so-called firehose of falsehood and, eventually, its personalised pitch. It is not just you or me, however: Big language models are now able to spew forth coherent, contextually relevant narratives in the blink of an eye. Meanwhile, deepfake technology means even experts can fall for unnatural-looking, sounding clips. These technologies, when used alongside automated bot networks and algorithmic pushes on social media, enable propagandists to influence operations orders of magnitude faster, more effectively, and more widely than in the past. As a result, disinformation has evolved from an opportunistic informational weapon to a tool in an organised, technology-enabled attack strategy. From a cybersecurity perspective, AI-generated disinformation represents a new level of attack that democratic societies must oppose. Conventional cybersecurity principles focus on protecting networks, including data and critical infrastructure. By contrast, AI disinformation attacks these layers at the cognitive level: it hones perceptions, alters belief systems, and plays with emotions. Such campaigns can undermine trust, skew public discourse, and sway electoral results without ever hacking any technical system. This change requires a new definition of cybersecurity that includes cognitive security alongside traditional technical defences. The cognitive threats are specifically pernicious to democratic systems, given their open and pluralistic nature – largely dependent on free information flow. Elections, media ecosystems and public deliberation are at risk when disinformation can organise across social divisions, identity-based tensions and political polarisation. These threats are further exacerbated by AI, which can facilitate customised messaging to specific groups based on their views, demographic profiles, or psychological triggers. It is therefore not just a matter of pockets of untruths, but of the erosion over time of democratic legitimacy and trust.

The world-spanning, border-blurring design of digital platforms only exacerbates the geopolitical dimensions of AI-generated disinformation. Influence operations can occur beyond national borders, making the problem of attribution and responsibility even more complicated. AI can be exploited by state and non-state actors to meddle in other nations' politics, engage in information warfare, or cause societies to unravel without ever pushing a key on their keyboards through traditional cyberattacks. This intersection of cybersecurity, information warfare, and hybrid threats has significant implications for traditional governance and defence arrangements. While policymakers and academics are increasingly assuming AI-based disinformation as a serious social threat, they also find that this phenomenon is often mishandled in cybersecurity discussions. There is a place for technical tools — or detection algorithms and content moderation systems, but they cannot solve the problem on their own. Effective responses require a combination of technological protections, regulatory efforts, institutional coordination, and societal resilience. Considering AI-based disinformation a cybersecurity problem provides a consistent framework for compiling these components. Against this background, the current work identifies AI-generated disinformation as a key security threat to democracy. It synthesises the best available knowledge, explains how AI-enabled disinformation attacks on democracy work, and identifies traditional cybersecurity strategies. By focusing on cognitive security and democratic resilience, the study seeks to contribute to the ongoing debate about how democracies can defend themselves in an age of generative AI.

Literature Review

With the advent of AI and content-generation tools, there is a new form of disinformation. Previous studies on digital media and democratic life highlighted that algorithmic systems curate what people see online, shape their political behaviour at scale, and, in turn, influence citizen participation in democratic processes (Lorenz-Spreen, 2022; Vaccari, 2020). AI-powered disinformation is distinctive because, unlike traditional misinformation, this technology is characterised by automation that seamlessly adapts to craft personalised messages that can reach millions. Recent research has characterised AI-facilitated disinformation as a hybrid socio-technical phenomenon that emerges from the interplay between algorithmic behaviours, platform infrastructure and human biases. Computational studies confirm that generative models have the potential to generate continuous narratives that are coherent over time and aligned with a specific ideology, and to adapt to an audience's attention rather than relying on occasional falsehoods (Romănishyn, 2025; Saeidnia, 2025).

This understanding has led scholars to regard disinformation not as an isolated anomaly but as a structural, ongoing risk. Generative AI techniques like large language models and synthetic media systems have made it much less resource-intensive to create realistic-looking fake content. A study shows that AI-generated text can closely replicate human writing style, rhetorical devices, and emotional tone (Drolsbach, 2025; Olanipekun, 2025). This has enabled bad actors to astroturf convincing but false narratives on a scale never before seen on digital platforms. The emergence of deepfake audio and video technology heightens the risks. Research on deepfakes has also shown that these recordings can even provoke disbelief among citizens who regard themselves as digitally literate, particularly in imitations of political figures, journalists, or public institutions (Ratnawita, 2015; Kaczmarek, 2015). As a result, the distinction between real and fake facts becomes fuzzier, eroding the credibility of visual and audio evidence in democratic discourse. An increasing number of authors now cast AI-powered disinformation as a cybersecurity problem, even though it doesn't break into technical systems. Instead of relying on hardware or software weaknesses, AI-powered disinformation attacks the human/cognitive layer: by influencing perceptions, trust, and decision-making (Mazurczyk, 2023; Mirzoyan, 2023). This reconceptualisation puts disinformation in line with the broader cyber threat model, which acknowledges that psychological and social manipulation is part and parcel of these types of threats. According to these scholars, such cognitive attacks may have strategic effects similar to those of conventional cyber operations, undermining political institutions or destroying social unity (Kreps, 2023; Singh, 2025). It is hard to attribute their mysterious origins, reflecting legal and jurisdictional grey areas in which cyber-extortion artists can operate. This imbalance, therefore, contributes to the appeal of AI-generated disinformation in hybrid and information warfare campaigns. Numerous empirical studies have demonstrated the negative impacts of AI-powered disinformation on democratic processes. Elections are especially vulnerable: there is evidence that AI narratives can alter voters' perceptions, suppress turnout, and undermine electoral legitimacy (Bennett, 2025; Yilmaz, 2024). More generally, continued exposure to disinformation increases polarisation and undermines the coherence of the public sphere. At the institutional level, AI-generated misinformation erodes confidence in journalism, scientific authority, and democratic institutions. When individuals can no longer trust the veracity of information, public deliberation diminishes, and trust in public institutions wanes (Marsden, 2021; Pawelec, 2022). Longitudinal studies indicate that such loss of trust is cumulative, eroding democratic resilience (Schipper, 2025; Wong, 2025). Increased attention to regulatory and governance responses has been driven by growing

awareness of these risks. Comparative studies reveal diverse efforts – from platform-led campaigns to state-driven approaches to cybersecurity strategies that incorporate information integrity into the national security agenda (Shoaib, 2023; Zhou, 2023). However, academics say such interventions also have to strike a careful balance between protecting democracy and fundamental rights, including the freedom of expression. Policy analysis is also illuminating the limitations of purely technical remedies. Content discovery tools may lag what the most sophisticated generative AI technologies can do (MacDonald, 2025; Schroeder, 2025). Therefore, scholars are increasingly promoting a multi-level counter-strategy that introduces technical measures combined with institutional collaboration and social resilience. Education, digital skills and cross-border cooperation are often mentioned as key elements of long-term defence mechanisms (Bontridder, 2021; Corsi, 2024; Imam, 2025).

Combined, this body of work illustrates that disinformation powered by AI poses a serious challenge to democratic societies, encompassing not only the provision of misinformation but also new cognitive and cybersecurity threats. Despite important strides in charting the contours of its mechanisms and effects, significant shortfalls persist — most notably in consolidating insights from the cybersecurity literature, governance science, and democratic theory into a comprehensive framework (Aditya, 2025; Sophia, 2025). These gaps need to be addressed by integrating interdisciplinary efforts to adapt to an evolving, increasingly intricate threat (Sambur, 2025; Wahab, 2025).

Methodology

This research is based on a qualitative, conceptual, analytical, and empirical design to investigate AI-generated disinformation as a global cyber threat to democracies. Given that no single empirical dataset captures all the dimensions of extent, pace, and complexity of AI-facilitated disinformation, a conceptual approach is warranted and legitimate. It allows insights from cybersecurity, political science, media studies, and AI governance to be systematically incorporated.

The methodological approach is based on the premise that AI-facilitated disinformation is hybrid, spanning technological, cognitive, and institutional dimensions. Rather than testing a particular hypothesis, the study's objective is to identify standard mechanisms, structural weaknesses, and cross-contextual structures that recur across different democratic contexts (Lorenz-Spreen & Endres, 2022; Mazurczyk et al., 2023). This design facilitates theory-building and policy-relevant analysis, which are both central to the study's objectives.

The insights are based solely on peer-reviewed academic and policy papers and interdisciplinary studies, referenced in the endnotes. The study uses a controlled corpus to maintain concept consistency and reduce the risk of arbitrary selection of data sources.

The literature conveys various disciplinary viewpoints, including:

- Artificial intelligence and generative models
- Cybersecurity and information warfare
- Democratic procedures and political communication
- Regulation of platforms and digital policy

It specifically prioritised references to AI-generated content, automation, synthetic media, and algorithmic amplification in democratic environments (Vaccari, 2020; Shoaib, 2023; Bennett et al., 2025).

The research uses thematic content analysis and cybersecurity-focused threat-mapping. Each publication was read and coded using predetermined analytical categories, developed in the literature and adjusted throughout the coding process.

The analysis was conducted in 3 stages:

1. First Coding: Repeated themes on AI capabilities, disinformation methods and democratic effects.
2. Thematic Clustering: We grouped codes that seemed similar to one another into overarching thematic categories resembling cybersecurity threat models.
3. Integrated Interpretation: Combining themes to analyse how AI-based disinformation operates as a systemic cybersecurity risk.

This allows for an overview of studies in a systematic comparison and of context-specific variation (Mirzoyan, 2023; Singh, 2025). To bridge disinformation research with cybersecurity theory, the work maps classical threat-modelling constructs, attack vectors, targets, vulnerabilities and impacts to AI-powered disinformation. Democracies are seen as complex systems in which human cognition is a key security layer.

Table 1. Cybersecurity-Oriented Threat Mapping Framework

Threat Component	Description	Application to AI-Driven Disinformation
Threat Actors	State and non-state entities	Political groups, foreign actors, coordinated networks
Attack Vectors	Means of attack	AI-generated text, deepfake audio/video, automated bots
Targets	Assets under threat	Voters, public trust, and democratic institutions
Vulnerabilities	System weaknesses	Cognitive biases, polarisation, platform algorithms
Impacts	Consequences	Electoral interference, trust erosion, and democratic instability

This mapping illustrates the similarities between AI-enabled disinformation and established cyber threats, strengthening the claim that information integrity must be recognised as a core cybersecurity asset (Kreps, 2023; Zhou, 2023).

Based on the thematic analysis, three analytical dimensions were consistently applied across all sources.

Table 2. Core Analytical Dimensions Used in the Study

Dimension	Focus	Analytical Purpose
Technological	Generative AI, automation, synthetic media	Identify enabling capabilities
Cognitive–Cyber	Perception, trust, and belief manipulation	Assess cognitive attack mechanisms
Democratic–Institutional	Elections, governance, public discourse	Evaluate systemic impacts on democratic systems

These dimensions facilitated a systematic review of each publication, which, in turn, enabled the examination to capture micro-level levers and consider how they contributed to broader systemic impact.

To increase analytical validity, this study adopts a triangulation approach across various academic disciplines and publication types. A result is only highlighted when it is consistently observed across studies. Explicit coding criteria and analytical dimensions provide reliability through clear definitions, allowing for the replication of this study in future research.

The qualitative nature of the study means that transferability, but not statistical generalisation, is a feature, yet conceptually straightforward and explanatorily deep analyses are necessary to engage with fast-moving threats (Schipper, 2025; Wong, 2025).

The study acknowledges several limitations. It is based on secondary work and does not include experimental or large-scale empirical evidence. Moreover, secondly, progress in generative AI technologies often moves so rapidly that some of the particular tools mentioned may themselves be outdated. However, by examining underlying mechanisms and structural patterns, the results remain relevant in the long term (MacDonald, 2025; Schroeder, 2025).

Combining cybersecurity threat modelling with democratic theory and AI governance research, this methodological approach provides a new lens for studying AI-mediated disinformation. It broadens a traditional cybersecurity architecture to include the realms of cognition and institution, supporting both further empirical work and future policy development and analysis.

Discussion

This article contributes to and extends this work by highlighting that AI-enabled disinformation should not be conceptualised merely as a communication challenge, but rather as a core cybersecurity threat to democratic societies. Results postulate that generative AI-powered disinformation has become a scalable, adaptable, and asymmetric attack surface at the cognitive layer of socio-technical systems, rather than at their technical infrastructure (Mazurczyk, 2023; Mirzoyan, 2023).

This reframing has profound implications for cybersecurity theory. The traditional cybersecurity model is focused on protecting the confidentiality, integrity, and availability of digital assets, but has proven insufficient at preventing threats that undermine cognitive trust or information integrity. AI-generated disinformation plays on just these kinds of gaps. The debate thus complements recent calls elsewhere to broaden cybersecurity regimes to encompass cognitive and information security, especially in democratic settings that require public trust (Kreps, 2023; Singh, 2025).

One of the most important observations from the analysis is a structural asymmetry between attackers and defenders. Generative AI dramatically lowers the threshold for influence operations by enabling even small groups and non-state actors to wield political influence on a level hitherto reachable only by states (Drolsbach, 2025; Olanipekun, 2025). Democratic institutions, on the other hand, are bound by law and ethics to react at the same tempo and with the same immediacy.

This imbalance is similar to that in other cybersecurity areas, but is amplified by the difficulty of attribution and jurisdictional complications. “The whole idea of disinformation is it is like a virus,” he said. These perceptions alone, in the absence of any technical violation, can undermine democratic legitimacy (Bennett, 2025; Yilmaz, 2024).

The conversation also underscores that, though machine-generated disinformation accumulates significant resistance over the long term, it undermines democratic resilience. Perpetual exposure to synthetic content not only produces misinformed individuals but also undermines trust in all sources of information, in general, real and authoritative (Vaccari, 2020; Marsden, 2021). That slow erosion of epistemic trust is especially hard to recover once it is established.

Table 3. Key Discussion Themes and Cybersecurity Implications

Theme	Key Insight	Cybersecurity Implication
Cognitive attacks	Disinformation targets perception and trust	Extend threat models beyond technical infrastructure
Asymmetry	AI enables low-cost, high-impact operations	Attackers gain a strategic advantage
Trust erosion	Long-term democratic harm	Trust becomes a central security asset
Platform amplification	Algorithms intensify the spread	Requires shared governance responsibility

From an ecosystem perspective, trust is a fundamental system precondition for democratic governance. Further, it undermines accountability and democracy, reduces citizen participation, and deepens political polarisation. The results are consistent with the literature, which argues that disinformation is an emerging cause of democratic fatigue (Schipper, 2025) and reduced political participation (Wong, 2025).

Another layer of the debate has revolved around the relationship between AI-generated disinformation and platforms' algorithms. Engagement-driven recommendation systems tend to spread emotionally heightened or divisive synthetic narratives (Lorenz-Spreen, 2022; Jaidka, 2024). Integrated into generative AI, these mechanisms generate self-reinforcing feedback loops that amplify the visibility and spread of harmful content. These issues bring to the fore acute questions of platform accountability and government regulation. While platforms have the technical power to reduce amplification, commercial motivations and the international scope of their business make enforcement challenging. The conversation supports a collaborative (platforms-regulators-civil society) shared-risk approach to information integrity (Zhou, 2023; MacDonald, 2025).

It also highlights the folly of relying solely on technical rectification. Detection mechanisms, watermarking systems, and automated moderation are necessary but can be easily bypassed and do not address the underlying social structures that sustain disinformation (Kaczmarek, 2025; Schroeder, 2025). Heavy reliance on automated moderation can also result in (over)blocking incidents and a risk of suppressing freedom of expression (Bontridder, 2021; Marsden, 2021). Given these constraints, the multi-layered nature of defence strategies that fuse technological detractors with institutional coordination and social policies is a necessity. Education, public enlightenment and media literacy stand out as crucial long-term factors contributing to the resilience of democracy (Corsi, 2024; Aditya, 2025).

The exchange also highlights important implications for cybersecurity policy and governance. Integrating AI-fueled disinformation into national cybersecurity strategy can improve readiness to address hybrid threats and foster coherence among government institutions (Shoaib, 2023; Imam, 2025). That kind of integration would better link election security, platform governance and information integrity efforts. Crucially, the report also argues that democratic resilience should be seen as a security objective in its own terms. This entails a shift from reactive tactics to proactive initiatives, such as capacity-building, international cooperation, and norm-building, aimed at maintaining trust and transparency in digital public spheres.

Table 4. Comparative Assessment of Countermeasures

Countermeasure Type	Strengths	Limitations
Technical detection	Scalable and automated	Easily bypassed; risk of errors
Platform moderation	Allows rapid intervention	Incentive misalignment; limited transparency
Regulation	Provides accountability and deterrence	Jurisdictional constraints
Societal resilience	Builds long-term capacity	Slow to develop; resource-intensive

The overall debate highlights that AI-enabled disinformation is a multidimensional cybersecurity threat, with effects that extend well beyond proximate political events and contribute to the erosion of the structural foundations of liberal democracy. Dealing with this threat will require rethinking cybersecurity in ways that include cognitive and institutional dimensions, informed by coherent policy action supported by technological innovation and societal-wide engagement.

Conclusion

AI-fueled disinformation has become a worldwide cybersecurity threat that strikes at the heart of democratic systems: public trust and decision-making. In contrast to traditional cyber-attacks, which focus on digital infrastructure, AI-driven influence operations target the cognitive and social foundations of democracy by amplifying persuasive falsehoods, generating counterfeit credibility, and hastening the proliferation of manipulative narratives across the net. This pivot widens the democratic attack surface, from networks and systems to perceptions, legitimacy and institutional credibility.

The argument developed here is that the strategic threat posed by AI-generated disinformation is much deeper than merely convincing citizens of particular falsehoods. Its damage is greater in eroding the foundations of democratic governance itself — common understandings, accountability, and informed participation. This legitimacy crisis — which does not disappear when false frames are eventually debunked — trains us to internalise artificial and conflicting information, normalise uncertainty, increase division amongst us, and wane confidence in the

media, elections, and public institutions. In that sense, the greatest danger isn't any one fake news story but the slow disintegration of epistemic stability itself.

The dynamics of these realities have a significant bearing on their cybersecurity strategy. Ensuring democratic integrity, in sum, means going beyond the narrow life-and-death technicality of security toward a broader approach that also encompasses information integrity, institutional coordination, and social robustness. This means that mitigation needs to work across several layers: improving transparency and accountability by online platforms, enhancing the rapid-response and strategic communication capabilities of public institutions, and investing in sustainable civic resilience through education and media literacy. These measures should, of course, be conceived as safeguards for democracy, not as undermining factors that control too much, take decisions behind closed doors, or impose unjustified limitations on freedom of expression.

In sum, the challenge to democracy posed by generative AI is not just a matter of protecting technical infrastructure but of defending democratic reality. Policymakers, cybersecurity experts, digital platforms, and civil society need to consider AI-powered disinformation as a significant security threat that requires coordinated, flexible, rights-respecting responses that adapt to the changing threat environment.

References

Aditya, S. (2025). The misinformation epidemic: combating AI-generated fake content and deepfakes. World Journal of Advanced Research and Reviews. https://doi.org/10.30574/wjarr.2025.26.2.1752

Bennett, W. L. & Livingston, S. (2025). Platforms, Politics, and the Crisis of Democracy: Connective Action and the Rise of Illiberalism. Perspectives on Politics. https://doi.org/10.1017/s1537592724002123

Bienvenue, E. (2020). Computational propaganda: political parties, politicians, and political manipulation on social media. International Affairs. https://doi.org/10.1093/ia/iiaa018

Bontridder, N. & Poullet, Y. (2021). The role of artificial intelligence in disinformation. Data & Policy. https://doi.org/10.1017/dap.2021.20

Corsi, G., Marino, B., & Wong, W. (2024). The spread of synthetic media on X. Harvard Kennedy School Misinformation Review. https://doi.org/10.37016/mr-2020-140

Demartini, G., Mizzaro, S., & Spina, D. (2020). Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. IEEE Data Eng. https://consensus.app/papers/humanintheloop-artificial-intelligence-for-fighting-demartini-mizzaro/6fe75dfbff

Drolsbach, C. & Pröllochs, N. (2025). Characterising AI-Generated Misinformation on Social Media. ArXiv. https://doi.org/10.48550/arxiv.2505.10266

Imam, M. (2025). The Threats of AI and Disinformation in Times of Global Crises. Bulletin of Islamic Research. https://doi.org/10.69526/bir.v3i4.394

Jaidka, K., Chen, T., Chesterman, S., Hsu, W., Kan, M., Kankanhalli, M., Lee, M., Seres, G., Sim, T., Taeihagh, A., Tung, A., Xiao, X., & Yue, A. (2024). Misinformation, Disinformation, and Generative AI: Implications for Perception and Policy. Digital Government: Research and Practice.

Kaczmarek, K., Karpiuk, M., & Melchior, C. (2025). Disinformation as a Threat to State Security. Przegląd Nauk o Obronności. https://doi.org/10.37055/pno/205780

Kreps, S. E. & Kriner, D. (2023). How AI Threatens Democracy. Journal of Democracy. https://doi.org/10.1353/jod.2023.a907693

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. Nature Human Behaviour. https://doi.org/10.1038/s41562-022-01460-1

MacDonald, E. (2025). Digital Authoritarianism and the Erosion of Democratic Norms: A Comparative Study of State Surveillance in Hybrid Regimes. OTS Canadian Journal. https://doi.org/10.58840/2an15325

Marsden, C., Brown, I., & Veale, M. (2021). Responding to Disinformation. Regulating Big Tech. https://doi.org/10.1093/oso/9780197616093.003.0012

Mazurczyk, W., Lee, D., & Vlachos, A. (2023). Disinformation 2.0 in the Age of AI: A Cybersecurity Perspective. Communications of the ACM. https://doi.org/10.1145/3624721

Miller, M. L. & Vaccari, C. (2020). Digital Threats to Democracy: Comparative Lessons and Possible Remedies. The International Journal of Press/Politics. https://doi.org/10.1177/1940161220922323

Mirzoyan, A. (2023). Digital Authoritarianism as a Modern Threat to Democratic Stability: Restriction of Freedom or Network Politicisation?. Journal of Political Science: Bulletin of Yerevan University. https://doi.org/10.46991/jops/2023.2.6.062

Olanipekun, S. O. (2025). Computational propaganda and misinformation: AI technologies as tools of media manipulation. World Journal of Advanced Research and Reviews. https://doi.org/10.30574/wjarr.2025.25.1.0131

Ratnawita, R. (2025). Cybersecurity in the AI Era Measures Deepfake Threats and Artificial Intelligence-Based Attacks. Journal of the American Institute. https://doi.org/10.71364/s3emxx77

Românishyn, A., Malytska, O., & Goncharuk, V. (2025). AI-driven disinformation: policy recommendations for democratic resilience. Frontiers in Artificial Intelligence. https://doi.org/10.3389/frai.2025.1569115

Saeidnia, H. R., Hosseini, E., Lund, B., Tehrani, M. A., Zaker, S., & Molaei, S. (2025). Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches. https://doi.org/10.1007/s10115-024-02337-7

Sambur, B. (2025). Artificial intelligence (AI) and institutional religion. Cyberpolitik Journal, 10(19), 94–97.

Schipper, T. (2025). Disinformation by design: leveraging solutions to combat misinformation in the Philippines' 2025 election. Data & Policy. https://doi.org/10.1017/dap.2025.18

Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N., Garcia, D., Goldenberg, A., Kyrychenko, Y., Leyton-Brown, K., Lutz, N., Marcus, G., Menczer, F., Pennycook, G., Rand, D. G., Schweitzer, F., Summerfield, C., Tang, A., Bavel, J. V., Linden, S. V. D., Song, D., & Kunst

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. 2023 International Conference on Computer and Applications (ICCA). https://doi.org/10.1109/icca59364.2023.10401723

Singh, K. & Kumar, H. (2025). Futuristic Media Information Literacy to Counter AI Generative Deepfake Media Content and Its Implications. Jharkhand Journal of Development and Management Studies. https://doi.org/10.70994/jjdms.10689.10703

Sophia, L. (2025). The Social Harms of AI-Generated Fake News: Addressing Deepfake and AI Political Manipulation. Digital Society &amp";" Virtual Governance. https://doi.org/10.6914/dsvg.010105

Vaccari, C. & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social media + Society. https://doi.org/10.1177/2056305120903408

Wong, W. (2025). Trends in Political Science Research: Artificial Intelligence and Voter Disinformation. International Political Science Abstracts. https://doi.org/10.1177/00208345251339324

Yılmaz, I., Akbarzadeh, S., Abbasov, N., & Bashirov, G. (2024). The Double-Edged Sword: Political Engagement on Social Media and Its Impact on Democracy Support in Authoritarian Regimes. Political Research Quarterly. https://doi.org/10.1177/10659129241305035

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & Choudhury, M. D. (2023). Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3544548.

Wahab, A. A. (2025). Artificial intelligence (AI) and cybersecurity. Cyberpolitik journal, 10(19), 85–93.

DİJİTAL KAPİTALİZM ÇAĞINDA YAPAY ZEKÂ, GÖZETİM VE İNSAN HAKLARI: MAHREMİYETİN VE ÖZGÜRLÜĞÜN GELECEĞİ

Demet Şefika MANGIR[*]
ORCID ID: 0000-0002-2542-8551

Declaration[*]

Özet

Dijital kapitalizm çağında insan hakları ve özgürlüklerin korunması hem uluslararası hukuk hem de küresel politika açısından giderek daha kritik bir mesele haline gelmektedir. Yapay zekâ temelli gözetim sistemlerinin yaygınlaşmasıyla birlikte, bireylerin mahremiyeti, ifade özgürlüğü ve veri güvenliği gibi temel haklar, dijital altyapıların ekonomik ve siyasal işleyişi içinde yeniden tanımlanmaktadır. Bu çalışma, dijital kapitalizmin ekonomik mantığı ile gözetim teknolojilerinin toplumsal etkilerini kesiştiren kuramsal bir çerçeve sunarak, insan haklarının dijital çağda nasıl dönüşüme uğradığını incelemektedir. Sonuç olarak, gözetim kapitalizminin yıkıcı etkilerini önlemede mahremiyet yasalarının yeterli olmadığı, aynı zamanda kurumsal yapılar ve uluslararası hukuk düzenlemelerinde köklü reformlarla mümkün olabileceğini göstermektedir. Bu kapsamda çalışma, dijital haklar, algoritmik hesap verebilirlik ve veri egemenliği konularındaki küresel tartışmalara katkı sunmayı amaçlamaktadır. Ayrıca politika yapıcılar ve hukuk reformcuları için dijital gözetim karşısında bireysel özgürlükleri koruyacak normatif mekanizmaların geliştirilmesine teorik ve pratik bir temel oluşturmayı da amaçlamaktadır. Sonuç olarak makale, dijital kapitalizm çağında mahremiyetin, özgürlüğün ve insan onurunun geleceğine ilişkin bütüncül bir değerlendirme ortaya koymaktadır.

Anahtar Kelimeler: Dijital kapitalizm, İnsan hakları, Dijital mahremiyet, Uluslararası hukuk, Özgürlük.

[*] Doç. Dr., SÜ İİBF Uluslararası İlişkiler Bölümü, demetacar@selcuk.edu.tr.
[*] Bu çalışmada yapay zekâ araçlarından yalnızca dilsel düzenleme ve yazım kurallarının iyileştirilmesi ve yabancı dildeki metinlerin çevrilmesi amacıyla yararlanılmış olup, çalışmanın bilimsel içeriği, kuramsal çerçevesi ve sonuçları tamamen yazara aittir.

Abstract

In the age of digital capitalism, the protection of human rights and freedoms is becoming an increasingly critical issue in terms of both international law and global politics. With the proliferation of AI-based surveillance systems, fundamental rights such as individual privacy, freedom of expression, and data security are being redefined within the economic and political functioning of digital infrastructures. This study examines how human rights are transformed in the digital age by presenting a theoretical framework that intersects the economic logic of digital capitalism with the social impacts of surveillance technologies. Ultimately, it demonstrates that privacy laws alone are insufficient to prevent the destructive effects of surveillance capitalism and that this can only be achieved through fundamental reforms in institutional structures and international legal regulations. In this context, the study aims to contribute to global debates on digital rights, algorithmic accountability, and data sovereignty. It also aims to provide a theoretical and practical basis for policymakers and legal reformers to develop normative mechanisms that protect individual freedoms in the face of digital surveillance. Ultimately, the article presents a comprehensive assessment of the future of privacy, freedom, and human dignity in the age of digital capitalism.

Keywords: Digital capitalism, Human rights, Digital privacy, International law, Freedom.

Giriş

Dijital kapitalizm, verinin birincil bir sermaye biçimi olarak işlev kazandığı, bilgi ve iletişim teknolojilerinin yanı sıra ağ temelli platformların ekonomik faaliyetlerin merkezi hâline geldiği çağdaş bir kapitalizm türünü ifade etmektedir. Bu çerçevede Zuboff (2019), gözetim kapitalizmi "insan davranışlarının izlenip kaydedilmesinden, bu davranışların tahmin edilmesine ve nihayetinde kâr elde edilmesine yönelik yeni bir iktisadi mantık" olarak tanımlamaktadır. Söz konusu sistem yalnızca özel sektörle sınırlı kalmamış, devletler ve uluslararası kurumlar da bu yapının içine dahil olmuştur. Özellikle gözetim kapitalizmi artık sadece bir iş modeli olmanın ötesinde, düzenlemeler, ticaret kuralları ve güçlü devletler tarafından desteklenen kurumsal bir çerçeveye sahip küresel bir güç düzenidir. Bu nedenle, dijital ortamlarda toplanan veriler ve bu veriler üzerine geliştirilen algoritmalar, küresel güç ilişkilerinin yeniden yapılandırılmasında belirleyici bir rol oynamaktadır (Kilic & Carr-Ryan, 2025). Dolayısıyla, az sayıda teknoloji devinin -Amazon, Google, Meta, Alibaba, Tencent,

Huawei gibi- elinde toplanmış olan küresel teknoloji hegemonisi bu bağlamda, hükümetlerin egemenliğine rakip ya da bazen ondan daha güçlü bir etki alanına sahip olmuştur.

Bu bağlamda, dijital kapitalizmin kurumsal ve ekonomik boyutlarının ötesinde, toplumsal denetim ve bireysel özgürlük alanlarını da yeniden şekillendiren gözetim pratikleri ortaya çıkmıştır. Yapay zekâ destekli gözetim teknolojileri -yüz tanıma sistemlerinden öngörücü polislik algoritmalarına ve sosyal medya filtreleme mekanizmalarına kadar- devletlerin ve özel sektör aktörlerinin bireyleri izleme, sınıflandırma ve yönlendirme biçimlerini köklü bir şekilde dönüştürmektedir. Günümüzde gözetim, bir polislik aracından daha fazlasıdır, dijital dünyamızın altyapısal bir koşulu haline gelmiştir. Lyon'un (2017) da belirttiği üzere, gözetim kavramı tarihsel olarak şüphelilerin, mahkûmların veya potansiyel suçluların yakından gözlemlenmesi anlamında kullanılırken, dijital çağda bu anlam köklü bir dönüşüm geçirmiştir. Günümüzde gözetim, bireylerin izlenmesi, dijital platformlar aracılığıyla kişisel verilerin toplanması, işlenmesi ve ekonomik değer üretiminde kullanılması sürecini kapsamaktadır. Bu durum gözetimi artık kültürel ve ekonomik bir olguya dönüştürmüştür. Benzer biçimde Zuboff (2019), bu yeni aşamayı *gözetim kapitalizmi* olarak tanımlayarak, gözetimin iktidar pratiği olmasının ötesinde, bütün bir siyasi-ekonomik düzenin temeli hâline geldiğini vurgulamaktadır. Böylece gözetim, klasik anlamdaki kontrol mekanizmalarından çıkıp dijital kapitalizmin temel üretim mantığına dönüşmüştür.

Bu bağlamda, dijital kapitalizm iş modeli düzeyinden, ulusal güvenlik, ticaret politikaları ve platform düzenlemelerinin iç içe geçtiği stratejik bir alan haline gelmiştir. Birkaç büyük teknoloji şirketi veri, bulut altyapısı ve algoritmaları üzerinden küresel ölçekte standart belirleme gücünü elde ederken, başka ülkeler veya aktörler bu standartlara uyum sağlama ya da bağımsız alternatif geliştirme noktasında dezavantajlı konuma düşmektedir. Bu durum, dijital ekonomide eşitsizliklerin derinleşmesine ve hegemonya mücadelelerinin sanal ortama taşınmasına olanak tanımaktadır. Dolayısıyla, dijital kapitalizm çerçevesinde az sayıda büyük aktör tarafından kontrol edilen veri ve altyapılar hem ekonomik hem siyasal düzeyde küresel adaletsizlikleri artırma potansiyeli taşımakta, uluslararası sistemde yeni hegemonik ilişkilerin temelini atmaktadır. Böylece küresel kapitalizmin dijitalleşme süreciyle birlikte üretim ilişkilerinin temelinde yaşanan dönüşüm, günümüz kapitalizmini artık klasik üretim ve emek sömürüsü biçimlerinden ziyade veri, algoritma ve dijital altyapı sahipliği üzerinden değer üreten bir sistemin parçası haline getirmiştir. Bu yeni evre, *dijital platform kapitalizmi* ya da *rant kapitalizmi* olarak adlandırılmaktadır (Özlük, 2023). Dijital kapitalizmin temelinde, üretim veri akışının ve algoritmik yönlendirme kapasitesinin kontrolü yer almakta, dolayısıyla

ekonomik güç, üretimden ziyade bilgi tekellerinden doğmaktadır. Bu yeni ekonomik düzende, bireylerin davranışları, tercihleri ve iletişim biçimleri sürekli olarak izlenmekte, kaydedilmekte ve ekonomik değere dönüştürülmektedir.

Günümüzde kişisel veriler her yerde toplanmakta, analiz edilmekte ve giderek daha fazla biçimde davranış tahmini amacıyla kullanılmaktadır. Bu durum, bireylerin mahremiyetinin ötesinde, ifade, düşünce ve bilgi özgürlüğü gibi diğer temel hakların korunması açısından da ciddi sorular doğurmaktadır. Avrupa hukuku bu gelişmeye güçlü bir yanıt vermiştir: Avrupa Birliği Genel Veri Koruma Tüzüğü (GDPR), kişisel verilerin korunmasını bir temel hak olarak tanımakta ve bu hakkın diğer özgürlüklerle dengelenmesi gerektiğini açıkça belirtmektedir. GDPR'nin 1. gerekçesinde (Recital 1) veri koruma hakkının temel bir hak olduğu vurgulanırken, 4. gerekçede (Recital 4) bu hakkın mutlak olmadığı, düşünce, ifade ve bilgi özgürlüğü gibi diğer haklarla uyum içinde uygulanması gerektiği ifade edilmektedir. Bu nedenle GDPR, veri gizliliğini koruyarak ifade ve bilgi özgürlüğü hakkı da dahil olmak üzere diğer AB temel haklarını destekleyen bütüncül bir yaklaşım benimsemektedir (European Union, 2016, Recital 1, Recital 4, Article 85). Benzer şekilde, Avrupa İnsan Hakları Sözleşmesi (AİHS/ECHR, 1950) bireylerin özel hayatına saygı hakkını (Madde 8) ve ifade özgürlüğünü (Madde 10) güvence altına almaktadır. Ancak gelişmiş yapay zekâ temelli gözetim sistemleri, bu güvenceleri yeni sınamalarla karşı karşıya bırakmaktadır. AİHM içtihadına göre (örneğin Big Brother Watch and Others v. the United Kingdom, ECHR 2021), kitle gözetimi uygulamaları kişisel verilerin orantısız biçimde toplanması yoluyla Madde 8 kapsamındaki özel hayat hakkını ihlal edebilmektedir. Benzer biçimde, algoritmik profilleme ve otomatik sınıflandırma süreçleri bireyleri şeffaf olmayan kriterlerle kategorize ederek ifade ve bilgi edinme özgürlükleri üzerinde caydırıcı bir etki yaratabilmektedir. Avrupa Konseyi'nin Mass Surveillance Factsheet (ECHR, 2023) belgesinde de belirtildiği gibi, bu tür uygulamalar AİHS'nin 8. ve 10. maddeleri arasında hassas bir denge kurulmasını zorunlu kılmaktadır. Dolayısıyla, yapay zekâ destekli gözetimin gelişimi, mahremiyetin ve ifade özgürlüğünün korunması açısından Avrupa insan hakları sisteminin temel ilkelerini yeniden değerlendirmeyi gerektirmektedir.

Yapay zeka destekli gözetim teknolojileri, mahremiyet, ifade özgürlüğü ve eşit muamele ilkelerini ciddi biçimde tehdit edebilmektedir. Özellikle otomatik içerik filtreleme ve sosyal kredi sistemleri aracılığıyla, demokratik toplumlarda ifade özgürlüğü ve eşit muamele ilkeleri üzerinde karmaşık sonuçlar doğurmaktadır. Otomatik içerik filtreleri, dijital platformlarda yasa dışı veya zararlı içerikleri önleme amacı taşımakla birlikte, bağlamı tam olarak

değerlendiremedikleri için sıklıkla eleştirel veya muhalif söylemleri yanlışlıkla bastırabilmektedir. Bu durum, algoritmik sansür olarak adlandırılan ve ifade özgürlüğünü dolaylı biçimde kısıtlayan yeni bir olguyu ortaya çıkarmaktadır (Article 19, 2018, Yeung, 2018). Benzer biçimde, Çin'de uygulanan sosyal kredi sistemleri, vatandaşların çevrimiçi ve çevrimdışı davranışlarını değerlendirerek bir güven puanı üretmekte, bu puan bireylerin kamu hizmetlerine erişimini, istihdam fırsatlarını veya sosyal statüsünü etkileyebilmektedir. Ancak bu sistemler, hangi davranışların *iyi* veya *kötü* olarak tanımlandığına dair şeffaflık eksikliği nedeniyle politik uyum ve öz-sansür eğilimini teşvik edebilmekte ve azınlık gruplarını orantısız biçimde etkileyebilmektedir (Drinhausen & Brussee, 2021, Citron & Pasquale, 2014). Her iki örnek de, yapay zekâ sistemlerinin teknik önyargı ve veri temelli karar verme süreçleri aracılığıyla, demokratik katılımı ve temel hakları istemeden zayıflatma potansiyeline sahip olduğunu göstermektedir. Yeni dijital altyapıların uluslararası siyaseti dönüştürmesi bağlamında, veri ve ağ platformlarına yönelik hakimiyet yalnızca ekonomik düzeyle sınırlı kalmayıp, jeopolitik güç dengelerini de derinden etkilemektedir. Örneğin UNCTAD'ın Digital Economy Report 2021 adlı çalışmasında, veri-paylaşımı olmayan ya da çok sınırlı olan ülkelerin küresel dijital kazançlardan oldukça düşük pay aldığı belirtilmiş ve dijital ekonomide hâkimiyetin Kuzey Amerika ve Çin merkezli büyük platformlara kaydığına dikkat çekilmiştir.

Öte yandan yapay zeka destekli gözetim uygulamaları, kamu güvenliği ya da dolandırıcılıkla mücadele gibi meşru amaçlarla savunulabilir. Avrupa Birliği de bu tür ödünleşmelerle yakından ilgilenmektedir: önerilen AI Yasası (AI Act), yüksek riskli yapay zeka sistemlerini (örneğin işe alımda ya da kolluk uygulamalarında kullanılanları) özel denetim ve şeffaflık yükümlülüklerine tabi tutmayı hedeflerken, otomatik GDPR'nin Madde 22 kapsamında karar süreçlerini sınırlandırmakta ve insan müdahalesi, itiraz hakkı gibi koruyucu mekanizmalar içermektedir. Bu bağlamda, modern gözetimin yeniden şekillenmesiyle temel hakların korunması arasındaki gerilim, hukuk ve haklar açısından ele alınması gereken hayati bir meseledir. Dolayısıyla makale, ampirik bir araştırma sunmaktan ziyade kuramsal ve normatif analiz yoluyla dijital gözetimin insan hakları üzerindeki etkilerini tartışmaktadır. Bu nedenle çalışma üç kuramsal geleneğin kesişiminde konumlanmaktadır. Gözetim çalışmaları geleneği, yapay zekâ temelli izleme pratiklerinin nasıl işlediğini, yani gözetimin arkasındaki mekanizmaları, bilgi akışlarını ve güç dinamiklerini çözümlememize imkân tanımaktadır. Haklar teorisi geleneği, bu gözetim biçimlerinin ne zaman meşruluğunu yitirdiğine ilişkin normatif ölçütleri -gereklilik, orantılılık ve insan onuru ilkeleri gibi- sağlamaktadır. Hukuk ve

teknoloji geleneği ise teknik kapasite ile hukuki denetim arasındaki yapısal boşlukları görünür kılarak, dijital çağda hesap verebilirliğin ve şeffaflığın nasıl sağlanabileceği sorusuna odaklanmaktadır. Bu üç gelenek birlikte, analizimizin kuramsal çerçevesini oluşturmakta ve bu çerçeve doğrultusunda, Avrupa hukuk sisteminin -özellikle AİHS, GDPR ve ilgili Avrupa Birliği Adalet Divanı (ABAD/CJEU) ile AİHM içtihatlarının- bu çok katmanlı zorluğa nasıl yanıt verdiği incelenecektir.

Dijital Kapitalizm, Gözetim ve Mahremiyetin Teorik Haritası

Makalenin kuramsal zemini, dijital kapitalizm, gözetim ve mahremiyet ilişkilerini açıklayan üç temel düşünsel geleneğin kesişiminde şekillenmektedir. Dijital kapitalizm, verinin birincil sermaye biçimi haline geldiği, bilgi teknolojileri ve algoritmik sistemlerin ekonomik üretim süreçlerinin merkezine yerleştiği yeni bir toplumsal-ekonomik düzendir. Bu yapı, piyasa dinamiklerini ve bilgi, iktidar ve gözetim ilişkilerini yeniden biçimlendirerek bireylerin mahremiyet alanlarını ekonomik değer üretiminin bir parçası hâline getirmektedir. Gözetim kuramı, yapay zekâ temelli izleme biçimlerinin ardındaki bilgi akışlarını, iktidar ilişkilerini ve davranışsal yönlendirme mekanizmalarını çözümleyerek dijital kapitalizmin gözetim mantığını açığa çıkarır. Haklar kuramı, bu gözetim yapılarının meşruiyet sınırlarını belirleyen normatif ilkeleri -özellikle gereklilik, orantılılık ve insan onuru ölçütlerini- tanımlayarak dijital mahremiyetin korunmasına teorik bir dayanak oluşturur. Hukuk ve teknoloji yaklaşımı ise dijital çağda kod, algoritma ve hukuk arasındaki karşılıklı etkileşimi inceleyerek teknik altyapıların normatif düzen üretme kapasitesini ve hukukun bu alanı düzenleme gücünün sınırlarını tartışır. Bu üç kuramsal çizgi birlikte ele alındığında, dijital kapitalizmin ekonomik mantığı, gözetimin iktidar biçimleri ve mahremiyetin normatif sınırları arasındaki ilişkileri bütüncül biçimde çözümleyen bir teorik çerçeve ortaya çıkmaktadır.

Yapay zekâ, dijital kapitalizm bağlamında yeni bir üretim faktörü olarak çapraz sektörsel yayılım kabiliyeti olan genel amaçlı bir teknoloji olarak kavramsallaştırılmaktadır. Yapay zekanın üretim süreçlerine entegrasyonu arttıkça iş modellerinin, gelir dağılımının ve sermaye formasyonunun köklü biçimde değişeceği öngörülmektedir. Gerçekten de, sanayi ve hizmet alanlarında otomasyon, makine öğrenimi ve veri-yoğun algoritmalar metalaşmış emek biçimleriyle birlikte ekonomik büyümenin yeni lokomotifleri hâline gelmektedir. Bu bağlama uygun olarak, yapay zeka patent sayıları ile uzun dönem ekonomik büyüme arasında pozitif ve istatistiki olarak anlamlı bir ilişki tespit edilmiştir, özellikle gelişmiş ekonomilerde bu etki daha güçlü gözlemlenmektedir. Bu dönüşümün ardında yatan dinamikler arasında veri ve

hesaplama gücü yoğunlaşması, yani büyük teknoloji firmalarının devasa veri setlerine erişimi ve bulut altyapısı kontrolü ile üst düzey yapay zeka personeline sahip oluşu yer almaktadır. Dolayısıyla eleştirel siyasal ekonomi literatüründe, bu yapı yapay zeka kapitalizmi olarak tanımlanmakta ve veri sömürüsü, ticarileştirme ve kazanan her şeyi alır (winner-takes-all) mantığı çerçevesinde değerlendirilmektedir (Gonzales, 2023).

Günümüzde yapay zeka destekli gözetim sorununa yaklaşımda, Foucault'nun disipliner toplum kavramı, Zuboff'un teknolojik gözetim kapitalizmi analizi ve Lyon'un gözetimin toplumsal boyutlarına vurgu yapan perspektifleri önemli kavramsal çerçeveler sunmaktadır. Bu nedenle bu kuramsal perspektiflerden yola çıkılarak yapay zeka çağında gözetimin nasıl yeniden üretildiği ve klasik teorilerle hangi noktalarda örtüştüğünün analiz edilmesi gerekmektedir. Bu bağlamda klasik gözetim teorisi yapay zeka çağında da önemini korumaktadır. Michel Foucault, Discipline and Punish: The Birth of the Prison adlı eserinde Jeremy Bentham (1995)'ın panoptikon modelini bir metafor olarak ele almakta ve modern disiplin toplumunun iktidar yapısını bu çerçeveden okumaktadır. Foucault (1977), gözetim mimarisinin işleyişini "…merkezdeki gözetleme kulesi tüm hücreleri görebilir, fakat gözetlenenler kimin onları izlediğini bilemez" şeklinde betimlemektedir. Dolayısıyla bu yapı, bilinçli ve kalıcı bir görünürlük durumu yaratırken, bireyler her an izlenme ihtimalini dikkate alarak davranışlarını düzenleyebilmektedir. Aslında Foucault'nun Bentham'ın Panopticon'u üzerine yaptığı analiz, asimetrik görünürlüğün nasıl öz disiplini teşvik ettiğini vurguladığını göstermektedir. Burada, gözetimin fiziksel bir sürekliliğe sahip olmasa bile algılanan olasılıkla bireylerin içsel denetim geliştirmesi, ceza mimarisinin, norm üretimi, disiplin ve iktidar-bilgi ilişkisinin kodlandığı bir sistemdir. Günümüzün dijital panoptikonu ise yaygın ve her yerde mevcuttur. Bireyleri hiçbir insan operatör aktif olarak izlemiyor olsa bile kameralar, dronlar, sosyal medya izleme ve biyometrik sensörler tarafından izlenmektedir.

Shoshana Zuboff'un *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* adlı eseri de, dijital platformların kullanıcı davranışlarını veri olarak sermayeleştirme süreçlerini incelemektedir. Zuboff'un temel iddialarından biri, kullanıcı davranışlarının *gözetim varlıkları* hâline gelmesidir, bu varlıklar algoritmalar aracılığıyla analiz edilir ve geleceğe dair tahmin modellerine dönüştürülür. Bu bağlamda, klasik iktidar modellerinden farklı olarak, davranışsal müdahale odaklı bir *enstrümantarya iktidar* kavramını geliştirmiştir. Bu iktidar biçimi, baskı ya da açık zor kullanımı yerine detaylı bir şekilde yapılandırılmış tekniklerle bireylerin seçim mekanizmalarını yönlendirmeye odaklanmaktadır. Dolayısıyla, Zuboff'un yaklaşımı, Foucault'nun norm

üretimi ve iktidar-bilgi ilişkisinin dijital bağlamda yeniden kodlandığını göstermektedir. Bu durumu *epistemik darbe* olarak adlandırmakta, çünkü bilme yetisi -yani hangi davranışların bekleneceği, hangi müdahalelerin yapılacağı- artık teknolojik aktörlerin eline geçmiştir (Zuboff, 2019). Bu bağlamda bireyler verinin üreticisi ve aynı zamanda deneği konumuna gelmişlerdir.

Gözetim çalışmalarına katkıda bulunan önemli bir kuramcı olan Lyon, gözetimi toplumsal süreçlerle iç içe geçen bir fenomen olarak ele almaktadır. Lyon'a göre gözetim, bireysel verilerin toplanmasından öte, bu verilerin sınıflandırılması, sıralanması ve bu sıralamaya göre müdahale edilmesi süreçlerini içermektedir. Bu süreç Lyon tarafından *sosyal sıralama* olarak kavramsallaştırmıştır. Lyon, gözetim süreçlerinde devletin yanı sıra özel aktörlerin, kurumsal yapıların ve bireylerin de önemli roller oynadığını vurgulayarak, gözetim yapılarının çok katmanlı ağlar şeklinde işlediğini ve mekânsal sınırların giderek silikleştiğini elektronik panoptikon ya da akışkan gözetim kavramlarıyla tartışır. Lyon'un analizinde, gözetim toplumsal düzeni yeniden üretme ve ayrımcı sınıflandırmalara aracılık etme kapasitesine sahiptir. Algoritmik sistemlerdeki yanlılık ve adalet sorunları, Lyon'un sosyal sıralama analizinin güncel teknolojilerle kesiştiği önemli alanlardır (Lyon, 2017). Dolayısıyla, gözetim sistemleri insanları kredi puanı, risk profili, davranış kalıpları gibi veriye dayalı kategorilere ayırarak farklı muameleye olanak tanımaktadır.

Günümüzde yapay zeka sistemleri ile gözetim pratikleri yeni biçimler kazanmıştır. Özellikle, bireylerin davranışlarını gerçek zamanlı olarak analiz edip sınıflandırarak panoptik algıyı algoritmik düzeyde yeniden üretebilmektedir. Bu bağlamda izlenen birey gözetim mekanizmasının nasıl işlediğini tam olarak bilmemekte, bu durum klasik panoptikon metaforundaki *görülür ama görünmez* dinamiğini daha karmaşıklaştırmaktadır. Dahası Zuboff'un enstrümantarya iktidar anlayışı ile norm üretim süreci teknolojik düzeyde yeniden ele alınmakta ve normlar artık algoritmalar aracılığıyla uygulanabilmekte ve müdahaleler otomatik biçimde yapılabilmektedir. Yapay zeka destekli sistemler, gözetimi tek merkezden çok aktörlü ağlar biçiminde ele alan Lyon'un yaklaşımını da derinleştirmektedir. Görüldüğü üzere, Foucault'nun panoptikon kuramı, Zuboff'un gözetim kapitalizmi ve Lyon'un sosyal sıralama analizleri birlikte incelendiğinde, gözetimin tarihsel sürekliliği ve teknolojik dönüşümü arasındaki bağ daha görünür hale gelmektedir. Bu bağlamda, yapay zeka çağında gözetim, norm üretme, müdahale, sınıflama ve öngörü pratikleri üzerinden yeniden yapılandırılmıştır (Foucault, 1977, Lyon, 2018, Zuboff, 2019) Bu yeni ortamda klasik

kuramsal araçlar hala anlamlıdır, ancak teknolojik özgünlükleri hesaba katacak şekilde eleştirel olarak uyarlanmalıdır.

Bu noktada, gözetimin meşruiyeti meselesi doğrudan insan hakları teorisinin normatif temelleriyle kesişmektedir. Gözetim teknolojilerinin meşru sınırları, teknik güvenlik, toplumsal fayda kriterleri ve hakların temel statüsü ile belirlenmelidir. Çağdaş hukuk teorisi, belirli özgürlüklerin -ifade, vicdan ve mahremiyet gibi- dokunulmaz bir konumda olduğuna dikkat çekmektedir. Bu çerçevede, John Rawls'un (A Theory of Justice, 1971/1999) adalet kuramında belirttiği üzere, temel hak ve özgürlükler *toplumsal fayda* adına feda edilemez, çünkü bu haklar, yalnızca özgürlüğün kendisiyle çatıştığında ve eşit özgürlük ilkesinin korunması gerektiğinde sınırlanabilmektedir. Benzer biçimde, Ronald Dworkin (1977) de hakların genel refah ya da fayda hesaplamalarına karşı *koz* işlevi gördüğünü savunmaktadır. Bu bağlamda, mahremiyet veya ifade özgürlüğü gibi hakların yalnızca diğer temel haklar ya da olağanüstü derecede önemli toplumsal çıkarlar söz konusu olduğunda sınırlandırılabileceğini, ancak sıradan verimlilik ya da fayda artışları için sınırlandırılamayacağını ileri sürmektedir. Ayrıca, Jeremy Waldron (2010) hakların onur boyutunu ön plana çıkararak, hakların uygulanmasındaki zorlukları kabul etmekle birlikte korunmalarının dikkatli bir dengeleme gerektirdiğini -ancak hiçbir zaman terk edilmemesi gerektiğini- vurgulamaktadır. Özellikle mahkemeler ifade özgürlüğü veya onur gibi temel değerleri korurken genellikle farklı değerleri dengelemek ve zorlu değer yargılarında bulunmak zorunda kalmaktadır. Kısacası, haklar teorisi mahremiyet, ifade özgürlüğü ve eşitliği adil bir toplumun önkoşulları olarak görür, dolayısıyla getirilecek sınırlamalar insan onurunu hukuka aykırı biçimde tehlikeye atmayacak şekilde -yani yasaya uygunluk, gereklilik ve orantılılık ölçütlerine uyarak- düzenlenmelidir. Bu yaklaşım, çıkış noktasını doğrudan AİHS ve AB Temel Haklar Şartı'ndan almaktadır: her türlü dijital gözetim rejimi, keyfî ihlallere karşı korunan bu haklarla uyumlu olmalıdır.

Mahremiyetin yalnızca gizlilik değil, kişinin kendi kimliğini temsil etme, sınırlarını çizme ve özerkliğini sürdürme hakkı olduğunu vurgulayan Allen'in (2008) etik zemini Moore'un (2010) mahremiyeti bireyin ahlaki özerkliğini ve bilişsel alanını koruyan temel hak olarak görmesiyle derinleşmektedir. Nissenbaum (2004) ise dijital çağın veri akışlarında mahremiyetin, mutlak bir gizlilikten ziyade *bağlamsal bütünlük* ilkesiyle korunabileceğini savunmaktadır. Buna göre, bilginin akışı ancak uygun toplumsal norm ve bağlamlar içinde meşru sayılabilir. Bu normatif yaklaşımlar, AİHS ve AB Temel Haklar Şartı'nın insan onurunu ve özel yaşamın gizliliğini ayrılmaz haklar olarak kabul eden düzenlemelerinde

yansımasını bulmuştur. Dolayısıyla dijital gözetim, algoritmik sınıflandırma ve yapay zekâ temelli karar süreçleri, bireyin mahremiyetini, özerkliğini ve eşitlik ilkesini tehdit ettiğinde, insan onuruna dayalı hak kuramı, hukukun müdahale sınırlarını belirleyen temel ölçütleri ortaya koyar. Bu yaklaşım, her türlü dijital denetim veya veri işleme faaliyetinin yalnızca gereklilik, orantılılık ve insan denetimi ilkelerine uygun olduğu sürece meşru sayılabileceğini savunur. Böylelikle insan onuru, klasik liberal hak anlayışından dijital çağın etik yönetişim paradigmasına uzanan bir süreklilik içinde hem hakların nihai dayanağı hem de onların sınırlandırılmasına yön veren normatif ilke olarak yeniden tanımlanmaktadır. Böylece, dijital çağda mahremiyet hakkı, insan onurunun, demokratik katılımın ve teknolojik yönetişimin meşruiyetinin de normatif dayanağı haline gelmiştir.

Bu çerçevede, hukuk ve teknoloji alanında yürütülen çalışmalar, gözetim aygıtlarının biçimlenmesinde yazılım altyapılarının *kod* ve düzenleyici normların birbiriyle etkileşime girdiğini ve sıklıkla çatıştığını ortaya koymaktadır. Örneğin, Lawrence Lessig (2000) *kod kanundur* ifadesiyle siber uzayın teknik mimarisinin toplumsal normları ve hukuki düzenlemeleri biçimlendiren bir düzenleyici etkinliğe sahip olduğunu vurgulamıştır. Bu bağlamda, bir sosyal medya platformunun gizlilik ayarları, hangi bilgilerin görünür ya da görünmez olacağını belirleyen mimari kodlar olarak işlev görmektedir. Dolayısıyla, bir yapay zekânın veya gözetim sisteminin tasarımı -tıpkı bir yasa gibi-gözetlemeyi ya kolaylaştırabilecek ya da sınırlandırabilecek bir mimari düzen anlamına gelmektedir. Bu noktayı daha ileri taşıyan Hildebrandt ve Tielemans (2013), tasarım yoluyla yasal koruma kavramını geliştirmiştir. Buradaki yaklaşıma göre, ağ bağlantılı çağımızda temel haklar -mahremiyet, adil yargılama hakkı, ayrımcılığa uğramama hakkı gibi- yazılı hukuk metinleri ve teknolojik altyapıların kendi içinde kodlanmış normlarla korunmalıdır. Dolayısıyla, pratikte hakların sistematik olarak güvence altına alınması -örneğin: varsayılan gizlilik ayarları, algoritmik etki değerlendirmeleri ve kullanıcıların veri akışı üzerindeki kontrolünü sağlayan mekanizmalar- için teknolojik çözümlerin uygulanması anlamına gelmektedir.

Teorik açıdan bakıldığında, dijital gözetim ve yapay zekâ temelli karar sistemleri, hukukun ve teknolojinin kesişiminde yer alan kritik bir normatif uyarıya işaret etmektedir. Bu bağlamda, temel hakların korunması için yasal düzenlemelerin yanı sıra bu düzenlemeleri somutlaştıran teknoloji mimarisine de hak koruma ilkelerinin entegre edilmesi gerekmektedir. Dolayısıyla, dijital gözetim rejimlerinin normların uygulanma biçimini belirleyen teknolojik tasarımlarla da şekillendiği görülmektedir. Bu nedenle, algoritmik süreçlerin şeffaflığı ve hesap verebilirliği sağlanmadıkça, hakların korunması yönündeki yasal çabalar teknik mimarinin

arkasında etkisiz kalma riski taşımaktadır. Bu noktada, Pasquale'nin algoritmik şeffaflık ve hesap verebilirliğe ilişkin eleştirileri, hak koruma ilkelerinin teknoloji tasarımının yapısal unsurlarında da içselleştirilmesi gerektiğini ortaya koymaktadır. Pasquale (2015), teknoloji destekli gözetimin *karanlık yüzünü* vurgulayarak, veri işleme süreçlerinin yoğunlaştığı dijital ekonomide güçlü aktörlerin ticari sırların ardına gizlendiğini ve algoritmaların sözde tarafsızlık iddialarının bu nedenle doğrulanamaz hâle geldiğini belirtir. Dahası, mevcut hukuk düzenleri çoğu zaman şirketlerin iş modellerini ve gözetim sistemlerinin gizli formüllerini de koruyarak hem bireyler hem de denetleyici kurumlar açısından ciddi bir şeffaflık sorunu yaratmaktadır. Bu durum, dijital düzenin giderek hukuki şeffaflığın yerini teknik gizliliğin aldığı bir yapıya dönüşmesine ve böylece denetim mekanizmalarıyla hukukun geleneksel işlevi arasında derin bir uyumsuzluk doğmasına yol açmaktadır.

Görüldüğü üzere, dijital çağda ortaya çıkan temel sorunlardan biri hesap verebilirlik boşluğudur. Tarihsel olarak insan karar vericilere yönelik geliştirilen hukuk sistemi, artık otomatikleştirilmiş, veri odaklı ve kodla işleyen güç biçimlerine uyum sağlamakta zorlanmaktadır. Bu yapısal boşluğu gidermeyi amaçlayan AB düzenleyicileri, önemli adımlar atmıştır: GDPR, bireyler üzerinde önemli etkiler doğuran tamamen otomatik karar alma süreçlerini yasaklamakta (Madde 22) ve veri sahiplerine açıklama talep etme, karar sürecine itiraz etme gibi haklar tanımaktadır. Benzer biçimde, Yapay Zekâ Yasası (AI Act) ve Dijital Hizmetler Yasaları, algoritmik sistemlerin tasarımı ile kullanımını düzenleyerek, insan denetimi, şeffaflık ve orantılılık ilkeleri çerçevesinde yeni bir hesap verebilirlik rejimi inşa etmeyi hedeflemektedir.

Sonuç olarak, bu kuramsal yapı dijital gözetimin teknik bir olgunun ötesinde derin bir normatif ve hukuki mesele olduğunu göstermektedir. Gözetim, haklar ve teknoloji arasındaki etkileşim, insan onurunu ve demokratik değerlere dayalı hukuk düzenini korumak için yeniden düşünülmesi gereken bir alan yaratmaktadır. Bu nedenle, dijital çağın gözetim pratikleri, hukukun meşruiyetini sınayan bir etik sorumluluk alanı haline gelmiştir.

Dijital Avrupa'da Gözetim ve Mahremiyet: Normatif Sınırlar ve Meşruiyet İlkeleri

Gözetim uygulamaları devlet çıkarları ile bireysel haklar arasında temel bir gerilim doğurmaktadır, liberal demokrasiler bağlamında mahremiyete yönelik her türlü müdahale, ancak güçlü ve meşru bir gerekçeyle savunulabilir. Normatif olarak sorgulanması gereken husus, gözetimin hangi koşullar altında meşru sayılabileceği ve ne zaman temel hakların yasadışı biçimde ihlal edildiği sorusudur. Bu soru özellikle, insan onuru ve mahremiyetin

temel değerler olarak ön plana çıktığı Avrupa İnsan Hakları Hukuku çerçevesinde büyük önem taşımaktadır. Nitekim, mahremiyet ile insan onuru arasındaki denge, gereklilik ve orantılılık ilkelerinin uygulanması ve gözetim faaliyetlerinin insan denetimine tabi olması gibi boyutlar, gözetimin meşruiyetini değerlendirmede kilit roller üstlenmektedir (Lissens, 2024, Fikfak & Izvorova, 2022).

Normatif hak teorileri, gözetimin ne zaman ve hangi koşullar altında bireysel hakları haklı olarak geçersiz kılabileceğini analiz etmek için önemli bir çerçeve sunmaktadır. Bu bağlamda Dworkin, bireysel hakları (örneğin gizlilik hakkını) kolektif hedeflerin üstünde konumlandırmakta ve ancak gerçek bir acil durum dışında ihlal edilemezliğini öne sürmektedir (Dworkin, 1977). Buna göre, gözetim uygulamaları ancak başka bir temel hakkı korumak amacıyla ya da gerçekten olağanüstü bir durumda meşru olabilmekte, rutin bir kamu refahı ya da güvenlik gerekçesi gizlilik hakkını geçersiz kılmak için yeterli olmamaktadır. Öte yandan Rawls'ın teorisi de benzer biçimde bireysel temel haklara özel bir öncelik tanımakta, temel eşit özgürlüklerin sosyal ve ekonomik hedeflerin önünde yer aldığını savunmaktadır (Rawls, 1971/1999). Bu bağlamda hakların üstünlüğünü esas alan anlayış, mahremiyet ve özgürlüklerin, kamu güvenliği gibi kolektif çıkarlar uğruna kolaylıkla sınırlandırılamayacağını öne sürmektedir.

Normatif hak kuramları perspektifinden bakıldığında, gözetimin hangi durumlarda bireysel hakların sınırlanmasını meşru kılabileceği tartışmalıdır. Bu konuda, özellikle 11 Eylül sonrasında benimsenen güvenlik-merkezli politikaların, hakları yalnızca güvenlik uğruna değiştirilebilir değişkenler olarak görme eğilimi eleştirilmektedir. Bu eleştirinin özünde, hakların maliyet-fayda analizine indirgenmesi ve özgürlük-güvenlik arasında nicel bir denge kurulabileceği varsayımının hem kavramsal hem de ahlaki açıdan yanıltıcı olduğu, dolayısıyla hakların kendi başına taşıdığı normatif önceliğin, araçsal gerekçelere indirgenmemesi gerektiği vurgulanmaktadır (Waldron, 2010). Dolayısıyla böyle bir yaklaşım, hakların doğasında var olan normatif üstünlüğü zayıflatmakta ve onları pragmatik bir çıkar analizinin unsuru hâline getirmektedir. Bu bağlamda özellikle temel medeni ve siyasal haklar, yalnızca değişen faydacı hesaplamalara dayanarak ödün verilemeyecek taahhütlerdir. Buna göre özgürlük ile güvenlik arasında kayan bir ölçek tahayyülü, hangi hakların kimin güvenliği için feda edildiği sorusunu ya da ölçütünü belirsizleştirmektedir (Waldron, 2010). Kısacası, özgürlüklerin sosyal ve ekonomik hedeflerin önünde konumlandırılması, bu özgürlüklerin kamu yararı gerekçesiyle sıradan biçimde feda edilemeyeceğini, gözetimin meşrulaştırılabilmesi için hak temelli ve ilkeli bir gerekçenin zorunlu olduğunu ortaya

koymaktadır. Böylelikle, güvenlik tehditleri hakların otomatik biçimde aşınmasına neden olmamalıdır.

Anita L. Allen (2008)'ın gizlilik etiğine ilişkin çalışması, gözetimin *prima facie* yani ilk bakışta ahlaki olarak yanlış olduğunu ortaya koyarak, hangi istisnai koşullar altında meşru görülebileceğine dair daha incelikli bir normatif çerçeve sunmaktadır. Allen'ın ortaya koyduğu *casusluk karşıtı ilke*, bireylerin kasıtlı biçimde gözetlenmesinin -gizlilik normlarını ihlal ettiği ve çoğunlukla aldatma veya gizliliğin ihlali üzerine kurulu olduğu için- doğası gereği etik dışı olduğunu öne sürülmektedir. Bu öncül, gözetlemenin ahlaki bakımdan tarafsız bir yönetim aracı değil, bireyin haysiyetine ve özerkliğine içkin bir tehdit oluşturduğunu vurgulamaktadır. Bununla birlikte Allen, belirli istisnai durumlarda gözetlemenin etik açıdan kabul edilebilir, hatta zorunlu olabileceğini de kabul etmektedir. Bu durumlar, inkâr edilemeyecek ölçüde iyi amaçlar -örneğin ciddi bir zararın önlenmesi- güdüldüğünde ve daha az müdahaleci bir alternatifin mevcut olmadığı hallerde görülmektedir. Bu yönüyle Allen'ın yaklaşımı, gözetimi hak temelli ve ilke odaklı bir etik çerçeveye yerleştirerek, onun meşruiyetini orantılılık, gereklilik ve insan onuruna saygı ilkeleriyle koşullandırmaktadır. Dolayısıyla, istisnai durumlarda dahi gözetleme pratiklerine katı ahlaki sınırlamaların getirilmesi gerektiğini ısrarla savunmaktadır. Bu bağlamda tanımladığı *erdemli casus* figürü, mahremiyete ancak ciddi bir isteksizlikle ve gözetlenen kişinin ahlaki iradesine saygı göstererek müdahale eden, bu müdahaleyi ise zararı en aza indiren, ölçülü ve özenli yöntemlerle gerçekleştiren bir aktörü temsil etmektedir (Allen, 2008). Bu anlayışa göre, etik olarak haklı gösterilebilecek her türlü gözetleme, en az müdahaleci araçları kullanmalı, yalnızca zorunlu olan hedefleri gözetmeli ve bireyin insan onuru ile özerkliğine koşulsuz bir saygı göstermelidir. Bu özellikle aşağılayıcı, keyfi ya da aşırı müdahaleci yöntemlerden sistematik biçimde kaçınmayı gerektirmektedir. Dolayısıyla Allen'ın normatif yaklaşımı, gözetimin meşruiyet sınırını belirleyen temel ölçütün insan onuru ve özerkliğin korunması olduğunu vurgulamaktadır. Zira gözetim, bireyleri yalnızca birer nesne ya da veri noktasına indirgediği anda -amacı ne kadar iyi olursa olsun- etik sınırın ötesine geçmiş olacaktır.

Görüldüğü üzere bu teorisyenlerin ortak yaklaşımı, gözetim olgusuna ilişkin hak temelli bir normatif varsayım ortaya koymaktadır. Gözetim, doğası gereği etik açıdan şüpheli bir eylem olarak değerlendirilmektedir. Bu şüpheyi ortadan kaldırmak için kamu güvenliği ya da genel çıkar gibi soyut gerekçeler yeterli görülmemektedir. Bunun yerine, yalnızca diğer temel hakların korunmasına veya acil durumların yönetimine yönelik, açıkça tanımlanmış ve ikna edici etik gerekçeler sunulmalıdır. Bu nedenle, her türlü gizlilik ihlali dar kapsamlı bir

biçimde tanımlanmalı ve sıkı ahlaki sınırlamalara tabi tutulmalıdır. Bu normatif duruş, gözetimin meşruiyetinin belirlenmesinde gizlilik, özgürlük ve insan onuru gibi temel hakların belirleyici olduğu anlayışıyla örtüşmektedir. Dolayısıyla, bu kriterleri karşılamayan her türlü gözetim uygulaması, meşru bir güvenlik önlemi değil, açık biçimde hak ihlali olarak değerlendirilmelidir.

Dijital Kapitalizm Çağında Mahremiyet ve İnsan Onuru

İnsan onuru, özerk, eşit ve ahlaki değere sahip bir özne olarak, kurucu bir insan hakları ilkesi olarak çalışmamızda ele alınmaktadır. Bu bağlamda, Avrupa insan hakları söylemi, mahremiyetin insan onuruyla ayrılmaz bir biçimde bağlantılı olduğunu vurgulamaktadır. Özellikle 20. yüzyılın totaliter rejimlerinin deneyimleri, kişisel verilerin kötüye kullanılmasının insan onurunu doğrudan zedeleyebileceğini açık biçimde göstermiştir. Nazi Almanyası'nın azınlıkları izlemek, sınıflandırmak ve yok etmek amacıyla veri sistemlerini kullanması, gözetimin etik sınırlarının tarihsel olarak nasıl aşıldığını somut biçimde ortaya koymuştur (Bygrave, 2014, Whitman, 2004). Bu tarihsel travmaların ardından oluşturulan Avrupa hukuk araçları -başta Alman Temel Yasası (Grundgesetz, 1949) ve Avrupa İnsan Hakları Sözleşmesi (AİHS) olmak üzere- mahremiyetin korunmasını kasıtlı olarak insan onuru kavramına dayalı bir hak olarak yeniden tanımlamıştır (ECHR, 1950, Article 8, Grundgesetz, 1949, Article 1). Dolayısıyla bu yaklaşım, bireylerin yalnızca devletin kontrol nesnesi veya veri kaynağı olarak görülmeksizin, kişiliklerini ve kimliklerini bağımsız biçimde geliştirme hakkına sahip olmaları gerektiği anlamına gelmektedir. Böylelikle mahremiyet, salt bir bilgi gizliliği meselesi olmaktan çıkarak, insan onurunu koruyan özgür ve özerk varoluşun etik temeli hâline gelmiştir.

Avrupa'daki mahremiyet-onur ilişkisi, Anglo-Amerikan yaklaşımından belirgin biçimde ayrılmaktadır. Whitman'ın (2004) ileri sürdüğü gibi, kıta Avrupası'nda mahremiyet hukukunun kökeni esasen onur ve itibarın korunmasına dayanan bir tarihsel gelenekten türemiştir. Buna karşın Amerikan yaklaşımı, bireyin devlet müdahalesinden özgürlüğü ilkesine dayanmaktadır. Bu ayrım, mahremiyetin Avrupa düşüncesinde özel alanın gizliliği dışında kişiliğin, özsaygının ve sosyal tanınmanın korunması olarak kavramsallaştırıldığını göstermektedir. Bu bağlamda Avrupa normatif perspektifinde gizlilik ihlalleri, doğrudan kişiliğin bütünlüğüne ve insan onuruna aykırı eylemler olarak değerlendirilmektedir. Sürekli gözetim pratikleri -anlık zarar doğurmasa bile- bireyleri kalıcı bir maruz kalma, güçsüzlük ve aşağılanma konumuna yerleştirerek onurla bağdaşmayan bir varoluş biçimine zorlayacaktır

(Bygrave, 2014, De Hert & Gutwirth, 2006). Bu durum özellikle, bireylerin özel iletişimlerinin, davranışlarının veya hareketlerinin izlenmesi gibi uygulamalarda daha belirgindir. Zira bu tür eylemler bireyi varsayılan olarak şüpheli konumuna indirgemekte ve özgürce kendini ifade etme kapasitesini zayıflatmakta ve onurun gerektirdiği kişisel saygıyı aşındırmaktadır. Sonuç olarak, Avrupa insan hakları söylemi, mahremiyet hakkını negatif özgürlük biçimi olmanın ötesinde insan onurunun kurucu unsuru olarak yeniden tanımlamaktadır (ECHR, S. and Marper v. United Kingdom, 2008, para. 66).

Mahremiyet ile insan onuru arasındaki ilişki, sıklıkla gerilimli bir değerler alanı olarak tartışılsa da Avrupa insan hakları hukukunda bu iki kavram birbirini karşılıklı olarak tamamlayan ve meşrulaştıran hak ilkeleri biçiminde kavramsallaştırılmaktadır. Mahremiyetin korunması, bireyin onurunu teyit etmenin ve kişisel bütünlüğünü sürdürmenin asli bir yoludur. Bu çerçevede etik zorluk, mahremiyet ile onuru birbirine karşı dengelemekten ziyade, her ikisini de güvenlik ve kamu yararı gibi diğer toplumsal hedeflerle uzlaştırmaktır. Bu bağlamda, gözetim pratiklerinde *onur* kavramının öne çıkarılması, sınırlama argümanını güçlendirir, çünkü bu vurgu, bireylerin her şeyi gören bir devletin şeffaf öznelerine indirgenmemesi gerektiğini hatırlatır (Bygrave, 2014). Dolayısıyla, gözetim uygulamaları meşru sayılabilecek olsa dahi, bireylerin onurunu ortadan kaldıracak biçimde yürütülmemelidir. Özellikle zalimane, aşağılayıcı veya son derece kişisel müdahale içeren yöntemlerden kaçınılması gerekmektedir. Nitekim AİHM, aşırı müdahaleci gözetleme biçimlerini, özel hayata saygı hakkı (ECHR, 1950, Article 8) kapsamında kişisel bütünlük ve onur temelli ihlallerle ilişkilendirmiştir. Daha aşırı vakalarda ise, son derece aşağılayıcı gözetim uygulamaları, bireyin insani özsaygısını zedeleyerek AİHS Madde 3 (insanlık dışı veya aşağılayıcı muamele yasağı) kapsamında değerlendirilmiştir. Böylece AİHM, mahremiyet ve onurun bir arada, insan kişiliğinin özüne ilişkin normatif bir bütünlük oluşturduğunu teyit etmiştir.

Görüldüğü üzere, onur merkezli bir etik anlayış, gözetim düzenlemelerinin kamu yararının yanı sıra bireyin özerklik ve eşit özne statüsüne saygı ilkesine de uygun biçimde sınırlandırılmasını gerektirir. Böyle bir yaklaşım, gözetimin meşruiyetinin amaçlarının ve kullandığı araçların bireysel onur üzerindeki etkilerinin değerlendirilmesine bağlı olduğunu varsaymaktadır. Dolayısıyla herhangi bir meşru gözetim politikası, bireysel onurun korunması için gerekli olan mahrem alanı koruduğunu, örneğin, mahrem kişisel verilerin toplanmasını sınırlayarak, bireyin şeref ve itibarına saygı göstererek ve kişilerin utanç verici ya da aşağılayıcı koşullarda izlenmesini önleyerek gösterebilmelidir. Bu çerçevede, mahremiyetin

insan onurunun korunmasıyla yakından ilişkili olduğu kabul edilmelidir. Güvenlik ile mahremiyet arasında sıfır toplamlı bir karşıtlık kurmak, etik açıdan yanıltıcıdır, zira her iki değer de özgür ve haysiyetli bir yaşam biçiminin önkoşullarını oluşturur (Solove, 2007). Sonuç olarak, insan onurunu zedeleyen gözetim, ahlaki meşruiyetini kaybedecek: güvenliği sağlama iddiası taşısa bile, böyle bir sistem yaşamaya değer bir toplumsal düzenin korunmasına hizmet etmekten uzaklaşacaktır.

Hukuki ve Normatif Perspektiften Gözetim: Gereklilik, Orantılılık ve İnsan Denetimi

Avrupa insan hakları hukukunda, özellikle AİHS 8. maddesi uyarınca, bireylerin özel hayatına yönelik her türlü müdahale -gözetim uygulamaları dâhil- yalnızca yasallık, meşru amaç, gereklilik ve orantılılık kriterlerinin eşzamanlı biçimde karşılanması hâlinde meşru kabul edilmektedir (Harris vd., 2018). Bu iki temel ilke, demokratik bir toplumda hakların sınırlanmasının keyfilikten korunması ve devlet müdahalesinin ölçülülüğünün sağlanması için normatif bir çerçeve oluşturmaktadır (Letwin, 2023). Bu bağlamda gereklilik ilkesi, devletin gözetim önlemini yalnızca acil bir toplumsal ihtiyaç durumunda uygulayabileceğini, yani söz konusu amaca ulaşmak için daha az müdahaleci bir alternatif bulunmadığını göstermesi gerektiğini öngörmektedir (Guida & Tozzi, 2020). Buna karşılık orantılılık ilkesi, gözetimin bireyin mahremiyet hakkına müdahalesinin, elde edilmek istenen kamu yararıyla makul bir denge içinde olmasını zorunlu kılmalı ve müdahale, ulaşılmak istenen amaçla karşılaştırıldığında aşırı bir yük getirmemelidir.

AİHM içtihadında bu ilkeler, gözetim önlemlerinin meşruiyetini belirleyen başlıca değerlendirme araçları olarak yerleşmiştir. Örneğin Roman Zakharov v. Russia kararında Mahkeme, kitlesel telefon dinleme rejimlerinin demokratik bir toplumda gerekli olabilmesi için yeterli denetim mekanizmalarına ve sınırlı hedefleme ölçütlerine sahip olması gerektiğini belirtmiştir (ECHR, 2015, para. 232). Benzer biçimde, Szabó and Vissy v. Hungary kararında, ulusal güvenlik gerekçesiyle yürütülen geniş kapsamlı izlemelerin *orantılılık* koşulunu karşılamadığı sonucuna varılmıştır (ECHR, 2016, para. 89). Dolayısıyla, Avrupa insan hakları hukukunda gereklilik ve orantılılık, gözetim politikalarının yasallığını ve ahlaki meşruiyetini de sınırlandıran normatif ilkeler olarak işlev görmektedir.

Bu ilkelerin uygulama düzeyinde yorumlanması, gözetimin ultima ratio -yani son çare- niteliğinde olması gerektiği anlamına gelmektedir. Buna göre, gözetim yalnızca daha az müdahaleci önlemlerin yetersiz kaldığı, belirli ve ölçülebilir bir tehdidin mevcut olduğu durumlarda kullanılmalı ve bu tehdidin yoğunluğu ve niteliğiyle orantılı biçimde

sınırlandırılmalıdır. Benzer şekilde, AB Adalet Divanı (ABAD), AB Temel Haklar Şartı'nın 7. ve 8. maddelerini yorumlarken, gözetimin meşruiyetini gereklilik ve orantılılık kriterleriyle sıkı biçimde ilişkilendirmiştir. Digital Rights Ireland v Ireland ve Tele2 Sverige AB v Post- och telestyrelsen kararlarında Divan, ayrım gözetmeyen, genel veri saklama rejimlerinin bireylerin mahremiyetine ve veri koruma hakkına ölçüsüz biçimde müdahale ettiğine hükmetmiş ve bu tür uygulamaları *gerçek gereklilik* ve *orantılılık* ilkelerinden yoksun oldukları gerekçesiyle geçersiz kılmıştır (CJEU, 2014, paras. 54–69, CJEU, 2016, paras. 99–107). Bu içtihatlar, gözetimin meşru sayılabilmesi için yasal temele ve sıkı normatif gerekçelendirmeye ve birey-merkezli bir sınırlandırma rejimine dayanması gerektiğini teyit etmektedir. Görüldüğü üzere ABAD içtihadı, geniş kapsamlı ve ayrım gözetmeyen veri toplama uygulamalarının, bireylerin özel yaşamına yönelik ağır bir müdahale teşkil ettiğini açık biçimde ortaya koymaktadır. Dolayısıyla, bu tür kapsamlı veri saklama önlemlerinin yalnızca eşit derecede ciddi tehditlerle -örneğin organize suç veya terörizmle mücadele gibi- haklı gösterilebileceğini ancak, soyut, genel veya varsayımsal risklerin bu ölçüde bir müdahaleyi meşrulaştıramayacağını vurgulamıştır.

Gereklilik ve orantılılık ilkelerinin uygulanması, devletlerin gözetim faaliyetlerini her bir vaka için ayrı ayrı sürekli biçimde gerekçelendirmesini zorunlu kılmaktadır. Geniş kapsamlı genel yetkilendirmeler veya süresiz kitlesel izleme uygulamaları bu ilkelerle bağdaşmamaktadır. Bunun yerine gözetim, belirli bir amaca yönelik, ölçülü ve zamanla sınırlı şekilde kullanılmalıdır. AİHM'nin birçok kararında vurgulandığı gibi, demokratik bir toplumda alınan önlemler yararlı olmasının yanı sıra kesinlikle gerekli olması bağlamında anlamlı olmaktadır. Dolayısıyla gereklilik ve orantılılık ilkesi, aynı amaca daha az müdahaleci araçlarla ulaşılıyorsa daha kapsamlı önlemlerin ölçüsüz sayılacağı anlamına gelir. Örneğin, tüm vatandaşların iletişim verilerini süresiz biçimde saklamak yerine yalnızca makul şüphe altındaki kişilerin iletişimlerinin izlenmesi hem daha dar kapsamlı hem de temel haklara daha saygılı bir yaklaşım benimsenmelidir. Bu yaklaşım, devletleri ayrım gözetmeyen kitlesel gözetimden uzaklaştırarak hedefli ve bağlama duyarlı gözetim yöntemlerini benimsemeye yönelten normatif bir teşvik yaratacaktır. Dahası gereklilik ve orantılılık ilkeleri düzenli denetimi de zorunlu kılmakta, ani bir tehdit veya olağanüstü durum sırasında gerekli görülen bir gözetim tedbiri, tehdit ortadan kalktığında artık gerekçesini yitirebilmektedir. Bu nedenle gözetim yetkilendirmeleri belirli sürelerle sınırlanmalı, devam edecekse yeni bir gerekçelendirme sürecinden geçmelidir. Avrupa hukuk literatüründe bu anlayış, hukukun

teknolojik dönüşüme uyum sağlayarak insan onuru, hesap verebilirlik ve demokratik denetim ilkelerini koruma çabasının bir parçası olarak değerlendirilmektedir.

Kısacası gereklilik ve orantılılık ilkeleri, temel hakların en önemli güvenceleri arasında yer alarak, demokratik bir toplumda gözetim faaliyetlerinin korumayı amaçladığı özgürlükleri ortadan kaldırmamasını sağlamaktadır. Araçlarla amaçlar arasında makul bir orantı kurulmasını ve haklara yalnızca zorunlu olduğu ölçüde müdahale edilmesini şart koşarak, bu ilkeler güvenlik politikalarının özgür, adil ve onurlu bir toplum düzeniyle bağdaşabilir kalmasını teminat altına almaktadır. Nitekim akademik literatürde de vurgulandığı üzere, güvenlik ile mahremiyeti karşıt kutuplar şeklinde konumlandırmak yanlış bir ikilemdir, çünkü güvenliğin meşru bir biçimde sağlanabilmesi, mahremiyetin özünü oluşturan iki temel değere -bireysel özerklik ve keyfi müdahalelere karşı korunma hakkına- saygı gösterilmesine bağlıdır. Dolayısıyla yalnızca gerekli ve orantılı nitelik taşıyan gözetim uygulamaları hukuken ve etik olarak meşru sayılmalı ve bu ölçütleri karşılamayan her türlü gözetim faaliyeti ise hak ihlali olarak değerlendirilmelidir.

Normatif kuram ve hukuk doktrininde sürekli vurgulanan bir tema, gözetim mekanizmalarında insan gözetiminin normatif zorunluluğudur. İnsan gözetimi, özellikle dijital çağın otomatik, özerk ve çoğu zaman şeffaf olmayan izleme teknolojilerinin yaygınlaştığı bir bağlamda, hukukun üstünlüğü, hesap verebilirlik ve temel hakların korunması açısından merkezi bir konuma sahiptir. Bu kavram, gözetim faaliyetlerinin yalnızca teknik süreçlere veya algoritmik yapılara bırakılmaması, aksine demokratik meşruiyet ve etik sorumluluk ilkeleri doğrultusunda, yetkilendirilmiş bireyler veya kurumsal aktörler tarafından sürekli denetlenmesi gerekliliğini ifade etmektedir (Green, 2022). Dolayısıyla, yüz tanıma gibi algoritmik gözetim teknolojilerinin uygulanmasından önce yetkili kurumlarca hazırlanacak etki değerlendirme raporlarının kamu denetimine açılması, insan karar vericilerin sürece dahil edilmesi gerekmektedir.

Bu çerçevede insan gözetimi, düzenleyici kurumlar, bağımsız denetim organları veya yargısal otoriteler gibi hesap verebilir mercilerin, gözetim uygulamalarının amaç, kapsam ve araçları üzerinde süreklilik arz eden bir kontrol mekanizması kurmasını gerektirmektedir. Böyle bir denetim, hem bireylerin özel hayatının gizliliğini ve ifade özgürlüğünü korurken hem de devletin gözetim yetkisinin keyfileşmesini önleyerek hukuk devletinin normatif sınırlarını güçlendirmektedir (Sterz et al., 2024). Dolayısıyla, insan gözetimi teknik bir güvenlik önlemi olmanın ötesinde demokratik yönetişimin ve insan onuruna dayalı hukuki düzenin

vazgeçilmez bir unsuru olmaktadır. Bu yaklaşım hem hukuki meşruiyet hem de demokratik denetim açısından kritik bir önem taşımakta ve insan gözetimi ve özgürlük boyutu bağlamında denetimsiz gözetim yetkilerinin bireysel özgürlük için ciddi bir tehdit oluşturduğunu kabul etmektedir. Bu nedenle özgürlüğün korunması, gözetim süreçlerinin her aşamasında bilinçli insan yargısına, şeffaf işleyişe ve bağımsız dış denetim mekanizmalarına tabi kılınmasını gerektirmektedir.

Normatif bir perspektiften bakıldığında, insan denetimi bireysel özgürlüklerin korunması ve devletin gözetim yetkisinin sınırlandırılması açısından temel bir güvence işlevi görür. Bu çerçevede, Dworkin'in hakların yalnızca diğer haklara veya olağanüstü durumlara tabi olabileceği yönündeki yaklaşımı, özgürlüklerin sınırlanmasının ancak istisnai ve rasyonel gerekçelere dayanabileceğini vurgular (Dworkin, 1977). Dolayısıyla, bir gözetim önleminin meşru kabul edilebilmesi için, bireysel haklara yapılan müdahalenin kamu yararıyla kurduğu denge, yalnızca teknik kurumlar tarafından değil, aynı zamanda bağımsız ve tarafsız insan aktörleri tarafından da normatif bir değerlendirmeye tabi tutulmalıdır.

Sonuç ve Değerlendirmeler

Bu çalışma, dijital kapitalizm çağında yapay zekâ destekli gözetim teknolojilerinin insan hakları üzerindeki etkilerini kuramsal ve hukuki bir çerçevede değerlendirmiştir. Özellikle mahremiyet, ifade özgürlüğü ve insan onuru gibi temel hakların, yapay zekâ teknolojilerinin yaygınlaştığı bir dijital altyapı içinde nasıl yeniden tanımlandığını ve kimi zaman tehdit altına girdiğini ortaya koymuştur. Dijital gözetim sistemlerinin teknik bir meselenin ötesinde etik, hukuki ve siyasal bir sorun olduğu sonucuna varılmıştır. Dolayısıyla, gereklilik, orantılılık ve insan denetimi ilkeleri, dijital gözetimin meşruiyetinin sağlanmasında temel normatif dayanaklar olarak tanımlanmıştır. Mahremiyet hakkı, bireysel bir özgürlük olmanın dışında demokratik toplumun temel taşı olan insan onurunun korunması bakımından da vazgeçilmez bir ilkedir. Bununla birlikte çalışmanın ortaya koyduğu en önemli bulgulardan biri, mevcut hukuk düzenlerinin hızlı dijital dönüşüm karşısında yetersiz kalmasıdır. Ayrıca hesap verebilirlik boşluklarının giderek artması ve algoritmik karar sistemlerinin şeffaflık ilkesinin zayıflamasıdır. Dijital gözetim, yalnızca yasal normlarla değil, aynı zamanda teknolojik tasarımlarla da şekillenmektedir. Dolayısıyla mahremiyetin ve özgürlüğün korunması, hukuk ve teknolojinin birlikte yeniden inşasını gerektirmektedir.

Bu bağlamda, algoritmaların ve veri işleme sistemlerinin, tasarım aşamasından itibaren temel haklara duyarlı bir şekilde yapılandırılması, varsayılan gizlilik ve kullanıcı özerkliği

ilkelerinin kod mimarisine entegre edilmesi gerekmektedir. Özellikle yüksek risk taşıyan yapay zekâ uygulamalarında, algoritmaların işleyişine dair şeffaflık sağlanmalı, değerlendirmeler kamuya açık hale getirilmelidir. Gözetim faaliyetleri teknik sistemlerin yanı sıra bağımsız kurumlar ve insan aktörler tarafından sürekli denetime tabi tutulmalıdır. Bireylerin veriler üzerindeki kontrolünü artıracak şekilde dijital egemenlik ilkesi geliştirilmelidir. Yapay zekâ ve gözetim teknolojilerinin sınır aşan niteliği göz önünde bulundurulduğunda, uluslararası hukukta koordineli düzenlemelerin yapılması, dijital hakların evrensel insan hakları çerçevesinde yeniden tanımlanması ve toplumun dijital haklar, algoritmik karar alma ve veri güvenliği konusunda bilinçlendirilmesi, özellikle genç kuşakların dijital yurttaşlık bilinciyle donatılması gerekmektedir.

Sonuç olarak, dijital kapitalizmin gözetim mantığına karşı insan haklarını korumak, teknik düzenlemelerle tek başına yeterli görülmemektedir. Daha derin etik, normatif ve yapısal müdahaleler gerekmektedir. Bu müdahaleler, mahremiyet ve ifade özgürlüğü gibi temel hakların demokratik toplum düzeninin devamlılığının teminatı olduğu düşüncesiyle yapılmalıdır. Bu bağlamda, insan onurunu merkeze alan ve teknolojik tasarımları hukuki normlarla bütünleştiren yeni bir dijital haklar rejimi inşa edilmelidir.

Kaynakça

Allen, A. L. (2008). The virtuous spy: Privacy as an ethical limit. University of Pennsylvania Law School, 1–17. https://scholarship.law.upenn.edu/faculty_scholarship/168/.

ARTICLE 19. (2018). *Side-stepping rights: Regulating speech by algorithm*. https://www.article19.org.

Bentham, J. (1995). Panopticon (Preface). In Miran Bozovic (ed.), The Panopticon Writings, London: Verso, 29-95.

Bygrave, L. A. (2014). Data privacy law: An international perspective. Oxford University Press.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1–33.

Court of Justice of the European Union (CJEU). (2014). *Digital Rights Ireland Ltd v Minister for Communications and Others*, Joined Cases C-293/12 and C-594/12, Judgment of 8 April 2014. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0293.

Court of Justice of the European Union (CJEU). (2016). *Tele2 Sverige AB v Post-och telestyrelsen and Secretary of State for the Home Department*, Joined Cases C-203/15 and

C-698/15, Judgment of 21 December 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62015CJ0203.

De Hert, P., & Gutwirth, S. (2006). Privacy, data protection and law enforcement: Opacity of the individual and transparency of power. In E. Claes, A. Duff, & S. Gutwirth (Eds.), *Privacy and the criminal law, Intersentia*, 61–104.

Drinhausen, K., & Brussee, V. (2021). *China's social credit system in 2021: From fragmentation towards integration*. Mercator Institute for China Studies (MERICS). https://merics.org/en/report/chinas-social-credit-system-2021-fragmentation-towards-integration.

Dworkin, R. (1977). Taking rights seriously. Harvard University Press. https://www.law.nyu.edu/sites/default/files/Ronald%20Dworkin%20-%20Hard%20Cases.pdf.

European Court of Human Rights. (1950). *European Convention on Human Rights*. Official Journal of the Council of Europe. https://www.echr.coe.int/documents/d/echr/convention_ENG.

European Court of Human Rights. (2008). *S. and Marper v. The United Kingdom*, Applications nos. 30562/04 and 30566/04, Judgment of 4 December 2008.

European Court of Human Rights. (2015). *Roman Zakharov v. Russia*, Application no. 47143/06, Judgment of 4 December 2015.

European Court of Human Rights. (2016). *Szabó and Vissy v. Hungary*, Application no. 37138/14, Judgment of 12 January 2016.

European Court of Human Rights. (2023). *Mass surveillance – Factsheet*. Press Unit. https://www.echr.coe.int/documents/d/echr/fs_mass_surveillance_eng.

Fikfak, V., & Izvorova, L. (2022). Language and persuasion: Human dignity at the European Court of Human Rights. *Human Rights Law Review, 22*(3), 1–30.

Foucault, M. (1977). *Discipline and punish: The birth of the prison* (A. Sheridan, Trans.). Vintage Books.

Gonzales, J. T. (2023). Implications of AI innovation on economic growth: A panel dataset analysis. *Journal of Economic Structures, 12*(1). https://doi.org/10.1186/s40008-023-00307-w

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review, 45*, 1–22. https://ssrn.com/abstract=3921216 or https://doi.org/10.2139/ssrn.3921216.

Grundgesetz, Article 1 of the Basic Law for the Federal Republic of Germany, (1949). Federal Law Gazette I p. 1.

Guida, S., & Tozzi, D. (2020). The assessment of the proportionality of the measures that limit the privacy fundamental rights in the new guidelines of the European Data Protection Supervisor. *European Journal of Privacy Law & Technologies, 2020*(1), 195–207.

Harris, D., O'Boyle, M., Bates, E., & Buckley, C. (2018). *Law of the European Convention on Human Rights* (4th ed.). Oxford University Press.

Hildebrandt, M., & Tielemans, L. (2013). Data protection by design and technology neutral law. *Computer Law & Security Review, 29*(5), 509–521. https://doi.org/10.1016/j.clsr.2013.07.005.

Kilic, B., & Carr-Ryan, D. (2025). The geopolitics of surveillance capitalism [Working paper]. *Harvard Kennedy School*. https://www.hks.harvard.edu/sites/default/files/2025-10/25_Kilic_Tech_Paper_0.pdf?utm.

Lessig, L. (2000). Code is law: On liberty in cyberspace. *Harvard Magazine*. https://www.harvardmagazine.com/2000/01/code-is-law-html.

Letwin, J. (2023). Proportionality, stringency and utility in the jurisprudence of the European Court of Human Rights. *Human Rights Law Review, 23*(3), 478–503. https://doi.org/10.1093/hrlr/ngad014.

Lissens, S. (2024). The foundations of EU personal data protection law: Privacy and human dignity. *EU-Renew*. https://eu-renew.eu/the-foundations-of-eu-personal-data-protection-law-privacy-and-human-dignity/.

Lyon, D. (2017). Surveillance culture: Engagement, exposure, and ethics in digital modernity. *International Journal of Communication, 11*, 824–842. https://ijoc.org/index.php/ijoc/article/view/5527.

Moore, A. D. (2010). *Privacy rights: Moral and legal foundations*. Penn State University Press.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*(1), 119–157. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10.

Özlük, D. (2023). Next stage of global capitalism: Digital platforms and rentier capitalism. *Ekonomi, Politika ve Finans Araştırmaları Dergisi (EPFAD), 8*(4), 681–695. https://doi.org/10.30784/epfad.1350739.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Rawls, J. (1999). *A theory of justice* (Rev. ed.). Harvard University Press. (Original work published 1971)

https://www.cita.or.id/wp-content/uploads/2016/06/John-Rawls-A-Theory-of-Justice-Belknap-Press-1999.pdf.

Solove, D. J. (2007). *The digital person: Technology and privacy in the information age*. New York University Press.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. *arXiv preprint*, arXiv:2404.04059. https://arxiv.org/abs/2404.04059.

Waldron, J. (2010). Dignity, rights, and responsibilities (Public Law Research Paper No. 10-83). *New York University School of Law*.

Whitman, J. Q. (2004). The two Western cultures of privacy: Dignity versus liberty. *Yale Law Journal, 113*(6), 1151–1221. https://doi.org/10.2307/4135739.

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance, 12*(4), 505–523. https://doi.org/10.1111/rego.12160

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

YAPAY ZEKA VE KÜRESEL GÜVENLİK MİMARİSİ: GÜÇ DAĞILIMININ MEKANİZMALARI VE YAPISAL DÖNÜŞÜM

Kürşat KAN[*]
ORCID ID: 0000-0003-2195-255

Declaration[**]

Özet

Bu çalışma, yapay zekanın küresel güvenlik mimarisi üzerindeki yapısal etkilerini ve güç dağılımını dönüştürme mekanizmalarını incelemektedir. Mevcut literatürde YZ'nin ürettiği nedensel zincirlerin ve kurumsal basıncın yeterince açıklanamaması temel araştırma problemini oluşturur. Makale, YZ'yi tekil bir araç yerine; hesaplama gücü (compute) yoğunlaşması, operasyonel hız ve modüler yönetişim üzerinden işleyen bir "çarpan teknoloji" olarak tanımlar. Araştırmada kuram güdümlü nitel bir tasarım benimsenmiş; yapılandırılmış belge analizi ve süreç izleme mantığına dayalı mekanizma temelli akıl yürütme yoluyla, güvenlik mimarisinin normatif, kurumsal ve operasyonel katmanlarında ortaya çıkan basınçlar ile uyum/gerilim dinamikleri izlenmiştir.

Bulgular, YZ'nin karar alma döngülerini sıkıştırdığını ve komuta-kontrol süreçlerini dönüştürdüğünü göstermektedir. Ancak bu teknolojik çarpanın etkisi homojen değildir. Hesaplama gücü, ileri çip üretimi ve veri merkezleri üzerindeki kontrol, devletler arasında sert bir hiyerarşi üretmektedir. Güç rekabeti artık model performansından ziyade altyapısal üstünlük ve kritik girdiler üzerinde denetim mücadelesine dönüşmüştür. Bu dönüşüm; BM, NATO ve silah kontrol rejimleri gibi çok katmanlı yapılar üzerinde yapısal bir baskı oluşturmaktadır. Bulgular; evrensel rejimlerin tıkandığı mevcut koşullarda, teknik standartlar ve ittifak içi düzenlemeleri kapsayan modüler araçların risk azaltımı için daha uygulanabilir bir yönetişim hattı sunduğuna işaret etmektedir.

Anahtar Kelimeler: Yapay zeka ve küresel güç dağılımı, Küresel güvenlik mimarisi, Stratejik istikrar ve tırmanma riski, Hesaplama gücü yoğunlaşması

[*] Dr. Öğr. Üyesi, Selçuk Üniversitesi, Uluslararası İlişkiler Bölümü
[*] AI language model (Claude) was minimally employed to help refine the writing style and clarity of certain sections in this paper.
[*] Bu çalışma Selçuk Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü tarafından 23401184'nolu proje kapsamında desteklenmiştir.

ARTIFICIAL INTELLIGENCE AND GLOBAL SECURITY ARCHITECTURE: MECHANISMS OF POWER DISTRIBUTION AND STRUCTURAL TRANSFORMATION

Abstract

This study examines the structural impact of artificial intelligence (AI) on the global security architecture and the mechanisms transforming the distribution of power. The core research problem is the lack of a clear causal chain explaining how AI-driven shifts exert pressure on norms, institutions, and operational tools. This article conceptualizes AI not as a singular capability but as a "multiplier technology" that redefines power competition through compute concentration, operational speed, and modular governance. Adopting a theory-driven qualitative design, the research utilizes structured document analysis and mechanism-based reasoning to trace structural pressures across the normative, institutional, and operational layers of the security architecture.

The findings indicate that AI compresses decision cycles and reshapes command-and-control and intelligence processes. However, the impact of this multiplier is not distributed uniformly. Concentration in computing power, advanced chips, and data centers produces a rigid strategic hierarchy. Consequently, power competition has evolved into an infrastructural struggle over critical inputs and the control of enabling infrastructures. This shift exerts structural pressure on the multi-layered security architecture, including the UN-centered collective security system, NATO, and arms control regimes. Given the deadlock in universal regimes caused by great power rivalry, the findings suggest that modular regulatory tools—such as technical standards and alliance-based arrangements—offer a more viable pathway for risk mitigation and governance.

Keywords: Artificial Intelligence and Global Power Distribution, Global Security Architecture, Strategic Stability and Escalation Risk, Compute Concentration

Giriş

Yapay zeka (YZ) bilimkurgu filmlerinin neredeyse yarım asırdır kullandığı objelerden biriydi. Sinemanın iyi veya kötü sanal karakteri, son on yıldır devletlerin ve insanların hayatına giren çok önemli bir karaktere evrildi. Bugün YZ sadece sinemanın bir karakterinden ötede, günlük yaşamın önemli bir parçası haline geldi. Bireysel hayatta bu kadar alan kaplayan bir unsurun devletlere etkisi de kaçınılmazdı. Ancak bu etki, on yıl önce konuşulan ve tartışılanlardan çok daha büyük ve asimetrikti. Bu etki kuramsal açıdan da pratikler açısından da yıkıcı ve yeniden inşa edici boyutlar taşımaktadır.

YZ, devletlerin göreli güç kazanımı arayışını hızlandırarak veri, hesaplama kapasitesi (compute) ve kritik altyapı üzerindeki yoğunlaşmayı stratejik üstünlüğün yeni ögelerine dönüştürmekte ve uluslararası eşitsizliği sert bir hiyerarşi şeklinde yeniden üretmektedir. Ayrıca karar süreçlerine asimetrik katılım ve düzenleme kapasitesindeki uçurumlar, teknolojik yönetişimi bazı aktörlerin lehine kurumsal bir eşitsizlik mekanizmasına çevirmektedir. YZ ekosisteminin küresel sermaye, platform gücü ve bilgi tekelleri etrafında örgütlenmesi, uluslararası ilişkilerde eşitsizliği yalnızca bir sonuç olarak değil; üretim ve denetim ilişkileri içinde süreklileşen yapısal bir tahakküm biçimi olarak da görünür kılmıştır.

Uluslararası sistem, tarih boyunca teknolojik dönüşümlerin yarattığı yapısal etkilerle defalarca yeniden şekillenmiştir. Sanayi Devrimi, buhar gücü, elektrik ve nükleer teknoloji; güç dengelerini dönüştürerek devletler arasındaki hiyerarşiyi derinden etkilemiştir. YZ günümüzde benzer bir dönüşüme öncülük eden teknolojidir.

YZ, devletlerin ekonomik kapasitesinden askeri yeteneklerine, toplumsal yapısından bilgi işleme kabiliyetine kadar uzanan geniş bir alanda yeni güç parametreleri üretmektedir. Akademik literatür, YZ'nin uluslararası rekabeti yeniden tanımladığını vurgulamaktadır (Horowitz, 2018; Johnson, 2019a; Sastry vd., 2024). Bu dönüşüm süreci, yalnızca askeri kapasite artışıyla sınırlı değildir. Rekabet; karar döngülerinin hızlanması ve komuta-kontrol sistemlerinin dönüşümü üzerinden şekillenmektedir. Ayrıca altyapısal yoğunlaşma, bu yeni güç mücadelesinin temel belirleyicilerinden biri olarak öne çıkmaktadır (Horowitz, 2018; Johnson, 2019a; Sastry vd., 2024).

Bu çalışmada YZ'nin küresel güç dağılımını nasıl dönüştürdüğü ve küresel güç ve güvenlik mimarisine etkisi ele alınacaktır. Bu çalışma, YZ'nin küresel güç dağılımını dönüştüren

mekanizmalarını kuramsal bir çerçevede belirlemeyi amaçlamaktadır. Bu mekanizmaların küresel güvenlik mimarisinin normatif, kurumsal ve operasyonel katmanlarındaki etkileri, süreç izleme ve yapılandırılmış belge analizi yöntemleriyle analiz edilmektedir.

Bu çalışma, YZ'nin küresel güç dağılımını hangi yapısal kanallar üzerinden dönüştürdüğünü sorunsallaştırmaktadır. Araştırmanın temel odağı, bu dönüşümün küresel güvenlik mimarisi üzerinde ürettiği dinamikleri analiz etmektir. Bu bağlamda, YZ'nin mimarideki farklı katmanlarda ne tür uyum ve gerilimlere yol açtığı incelenmektedir. Analitik çerçeve; BM merkezli kolektif güvenlik sistemi, bölgesel düzenlemeler ve silahların kontrolü rejimlerini kapsamaktadır. Ayrıca uluslararası insancıl hukuk ve ceza adaleti mekanizmaları da çalışmanın kapsamına dahil edilmiştir. Çalışma, YZ'yi tekil bir kapasite artışından ziyade; karar döngülerini sıkıştıran, bilgi işleme maliyetlerini düşüren ve komuta-kontrol/istihbarat/hedefleme süreçlerini dönüştüren bir "çarpan teknoloji" olarak ele alır.

Çalışmanın araştırma sorusu, birbirini tamamlayan üç temel hipotez üzerinden sınanmaktadır.

İlk hipotez, YZ kapasitesinin fiziksel altyapıdaki yoğunlaşmasına odaklanmaktadır. İleri düzey çipler, bulut sistemleri ve veri merkezleri bu kapasitenin temel bileşenleridir. Günümüzde güç rekabeti, model performansından ziyade bu kritik girdilerin ve hesaplama gücü akışlarının kontrolüne kaymaktadır. Söz konusu eğilim, uluslararası sistemdeki hiyerarşik yoğunlaşmayı derinleştiren temel bir unsurdur.

İkinci hipotez, YZ'nin karar alma süreçlerindeki hız ve otomasyon etkisini ele almaktadır. Daralan karar süreleri, kriz yönetiminde yanlış anlama ve hatalı hesaplama riskini belirgin şekilde artırmaktadır. Özellikle YZ-siber etkileşimi, komuta-kontrol süreçlerinde tırmanma riskini tetikleyerek stratejik istikrar üzerinde ciddi bir baskı oluşturmaktadır.

Üçüncü hipotez, küresel yönetişim mekanizmalarındaki yapısal dönüşümü incelemektedir. Büyük güç rekabeti ve mevcut kurumsal tıkanıklıklar, bağlayıcı evrensel rejimlerin inşasını giderek zorlaştırmaktadır. Bu süreçte güvenlik yönetişimi, daha kademeli ve modüler araçlara evrilmektedir. Teknik standartlar, ittifak içi düzenlemeler ve tedarik zinciri üzerinden kurulan fiili uyum baskıları, bu yeni dönemin temel yönetişim enstrümanlarıdır.

Bu çalışma, metodolojik olarak kuram güdümlü nitel bir araştırma tasarımı üzerine inşa edilmiştir. Araştırmada birbirini tamamlayan iki temel teknik kullanılmaktadır. Bunlar; yapılandırılmış belge analizi ve süreç-izleme (process tracing) mantığına dayalı nedensel akıl yürütmedir (Bowen, 2009; Beach & Pedersen, 2019; Bennett & Checkel, 2015).

Yapılandırılmış belge analizi, küresel güvenlik mimarisinin normatif ve kurumsal metinlerine odaklanır. Bu aşamada; YZ'ye ilişkin risk tanımları, düzenleme tercihleri ve önerilen yönetişim araçları sistematik olarak ayrıştırılmaktadır (Bowen, 2009).

Süreç-izleme mantığı ise teorik mekanizmalar ile ampirik çıktılar arasındaki nedensel bağları değerlendirir. Örneğin; hesaplama gücü yoğunlaşması veya otomasyon hızı gibi mekanizmaların, stratejik istikrar ve yönetişimde parçalanma gibi sonuçlarla tutarlılığı analiz edilmektedir (Beach & Pedersen, 2019; Bennett & Checkel, 2015).

Araştırmanın veri seti geniş bir yelpazeye dayanmaktadır. Resmi kurum metinleri ve politika belgeleri birincil kaynakları oluştururken; hakemli literatür ve araştırma raporları ikincil kaynaklar olarak kullanılmaktadır.

Araştırmanın yöntemsel geçerliliği üç temel dayanak üzerine inşa edilmiştir. Bu süreç; kavramların operasyonel olarak tanımlanmasını, veri setinin şeffaf sınırlandırılmasını ve bulguların literatürle üçgenlenmesini (triangulation) kapsamaktadır (Bowen, 2009; Beach & Pedersen, 2019).

Makalenin temel argümanı belirli bir analitik zincir doğrultusunda ilerlemektedir. İlk aşamada, YZ'nin güç dağılımını dönüştüren temel mekanizmaları kavramsal düzeyde belirlenmektedir. Bu kapsamda hesaplama gücü yoğunlaşması, otomasyon hızı ve ağ merkeziliği gibi unsurlar analiz edilmektedir. İkinci aşamada, bu mekanizmaların küresel güvenlik mimarisi üzerinde yarattığı baskılar sistematik olarak izlenmektedir. Analiz; mimarinin normatif, kurumsal ve operasyonel katmanlarında yoğunlaşmaktadır. Bu süreçte stratejik istikrar, hukuki uyum ve yönetişimde parçalanma gibi temel sorunlar ele alınmaktadır. Son aşamada ise bağlayıcı rejimlerin sınırları ve modüler düzenleme seçeneklerinin uygulanabilirliği tartışılmaktadır. Söz konusu tartışma, büyük güç rekabeti ve mevcut kurumsal tıkanıklıklar ekseninde yürütülmektedir. Bu kurgu, çalışmadaki tüm kavram ve bulguları hipotezlerin öngördüğü mekanizma-sonuç ilişkisi içinde konumlandırmayı amaçlamaktadır.

Bu makalenin özgün katkısı, YZ'yi yalnızca yeni bir askeri kapasite artışı olarak değil; bulut altyapısı, ileri çipler ve yüksek ölçekli veri-merkezleri üzerinden şekillenen bir "altyapısal çarpan" olarak kavramsallaştırmasında yatmaktadır. Bu çerçevede "hesaplama kapasitesi egemenliği", bir aktörün yüksek performanslı hesaplama kaynaklarına (ileri yarı iletkenler, bulut ve veri merkezi kapasitesi, kritik tedarik zincirleri) erişimi kontrol etme ve bu

kaynakları stratejik amaçlar doğrultusunda yönlendirme yeteneğini ifade eder. Dolayısıyla YZ rekabeti, model yeteneği kadar, bu yeteneği mümkün kılan altyapının yoğunlaşması ve boğaz noktalarının yönetimi üzerinden de güç dağılımını yeniden üretmektedir (Farrell & Newman, 2019; Hawkins vd., 2025; U.S. Department of Commerce, Bureau of Industry and Security, 2023).

Literatüre katkı üç düzlemde yapılandırılmıştır. Birincisi, güç tartışmasını salt "platform/algoritma" rekabeti düzeyinde bırakmak yerine, YZ'nin üretim ve kullanım koşullarını belirleyen altyapı-asimetriyi (hesaplama kapasitesi yoğunlaşması) gücün yapısal bir bileşeni olarak görünür kılar (Horowitz, 2018; Johnson, 2019a). İkincisi, bu altyapısal dönüşümün küresel güvenlik mimarisinin normatif, kurumsal ve operasyonel katmanlarında nasıl bir basınç ürettiğini mekanizma-temelli bir argüman zinciri içinde eşler. Üçüncüsü, bağlayıcı rejimlerin tıkanıklıklarını dikkate alarak, risk azaltımı ve uyum için "modüler düzenleme" yaklaşımını pratik bir yönetişim alternatifi olarak tartışır.

Bu çalışma, kuram güdümlü nitel bir araştırma tasarımı çerçevesinde, yapılandırılmış belge analizi ve süreç-izleme mantığıyla mekanizma-temelli bir değerlendirme yürütmektedir (Bowen, 2009; Beach & Pedersen, 2019; Bennett & Checkel, 2015). Dolayısıyla amaç, YZ'nin etkilerini tekil bir vaka üzerinden nedensel olarak "ölçmekten" ziyade; literatürde dağınık halde bulunan nedensel kanalları tutarlı bir analitik zincire bağlamak ve bu zincirin güvenlik mimarisinin katmanlarına nasıl yansıdığını göstermekti.

Ampirik kapsam, başlıca uluslararası örgüt belgeleri, strateji dokümanları, düzenleyici metinler ve alan raporları gibi açık kaynak metinlere dayanmaktadır. Bu tercih, bulguların genellenebilirliğini istatistiksel bir temsilden değil; kaynak çeşitliliği, kavramsal tutarlılık ve mekanizma-çıktı eşlemesinin izlenebilirliğinden türetir. Bu nedenle çalışma, firma düzeyinde performans kıyasları, model değerlendirme metrikleri veya ayrıntılı teknik doğrulama analizleri sunmamaktadır.

Siber alan bu makalede bağımsız bir alt disiplin veya "siber güç/siber düzen" literatürünün başlı başına bir inceleme alanı olarak ele alınmamaktadır. Bunun yerine siber alan, YZ'nin hız, otomasyon ve bilgi bütünlüğü üzerindeki etkilerinin somutlaştığı bir uygulama düzlemi olarak konumlandırılır ve YZ-siber etkileşimi stratejik istikrar ve bilgi güvenliği bağlamında tartışılır (Johnson, 2019b). Bu sınırlama, çalışmanın analitik odağını YZ'nin küresel güvenlik mimarisinin katmanlarında ürettiği yapısal basınçlara yoğunlaştırmak ve siber alanı bu basınçların görünürleştiği mekanizmalardan biri olarak ele almak amacıyla tercih edilmiştir.

Son olarak, çalışma öngörü (forecasting) iddiası taşımaz; belirli bir ülke veya bölgeye ilişkin ayrıntılı alan incelemesi yerine, büyük güç rekabeti bağlamında ortaya çıkan genel mekanizmaları ve yönetişim seçeneklerini tartışır. Bu nedenle vaka dışlamaları (ör. tekil savaş alanı uygulamaları veya kapalı veri setlerine dayalı değerlendirmeler) bilinçli bir tercih olup, makalenin kapsamını kavramsal bütünlük içinde sınırlamayı hedefler.

Gücün Doğası, Değişim ve Dönüştürücü Mekanizmalar

Bu bölüm, araştırmanın bağımsız değişkenini inşa etmektedir. YZ'nin güç üretimini dönüştürme süreçleri bu kısımda tanımlanmaktadır. Özellikle birinci ve ikinci hipotezlerde (H1 ve H2) öngörülen mekanizmalar kavramsal olarak temellendirilmektedir. Bu sayede, küresel güvenlik mimarisi üzerindeki baskı noktaları analitik olarak netleştirilmektedir. Analiz, teknolojik dönüşümün kurumsal yapılar üzerindeki etkisini anlamayı kolaylaştıran bir çerçeve sunmaktadır.

Uluslararası ilişkiler literatüründe güç, uzun süre askeri kapasite, ekonomik büyüklük ve teknolojik üretim gücüyle ölçülmüştür. YZ bu ölçütleri ortadan kaldırmaktan ziyade, onları yeniden işleyen bir ara katman üretir ve bunların etkinliğini artıran/azaltan bir çarpan işlevi görür. Karar alma döngülerini sıkıştırır, bilgi üretim maliyetlerini düşürür, üretimde otomasyon imkanlarını genişletir ve güvenlik alanında hedefleme-istihbarat-komuta-kontrol süreçlerini dönüştürür (Horowitz, 2018).

YZ'nin küresel güç dağılımı üzerindeki dönüştürücü etkisini belirgin kılan temel dinamik, bu teknolojinin homojen bir yayılım sergilememesidir. Aksine YZ ekosistemi; ileri nesil çipler, bulut altyapısı, yüksek kapasiteli veri merkezleri, araştırma ekosistemleri ve norm belirleyici kurumlar gibi stratejik odak noktalarında yoğunlaşma eğilimi göstermektedir. Bu yapısal yoğunlaşma merkezinde üzerine odaklanılması gereken soru şudur: YZ, küresel güç dağılımını hangi yapısal kanallar üzerinden dönüştürmekte ve bu süreç hangi aktörler lehine asimetrik bir güç yoğunlaşması üretmektedir?

Bu sorunsalın aciliyeti birbiriyle kesişen iki eğilimin yarattığı gerilimden doğmaktadır. İlk eğilim, hesaplama gücü talebindeki artışın endüstriyel yoğunlaşmayı hızlandırmasıdır. İleri seviye YZ modellerinde hesaplama gücü ihtiyacı hızla artarken, model geliştirme kapasitesinin büyük ölçüde özel sektör odaklı hale gelmesi dikkat çekicidir. Stanford AI Index 2025 raporu, 2024 yılı itibarıyla en öne çıkan modellerin ezici çoğunluğunun endüstri kaynaklı olduğuna işaret etmektedir (Maslej vd., 2025). İkinci eğilim ise jeopolitik rekabet ve

dışlayıcı erişim rejimleridir. YZ üretim zincirinin temel girdilerine (ileri seviye yarı iletkenler ve üretim ekipmanları) erişimin bir ulusal güvenlik ve stratejik rekabet unsuruna dönüşmesi; ihracat kontrolleri vasıtasıyla hesaplama ekosisteminin bir kısıtlama ve dışlama alanına evrilmesidir (U.S. Department of Commerce, Bureau of Industry and Security, 2023).

Bu iki eğilim birlikte okunduğunda, YZ'de rekabet avantajının salt yenilikçilikten ibaret olmadığı; sermaye yoğun altyapı, tedarik zinciri boğazları ve erişim rejimleriyle şekillenen daha sert bir hiyerarşi ürettiği görülür. Dolayısıyla güç mücadelesi, model performansının ötesinde, hesaplama gücü akışlarının kontrolü ve kritik girdilere erişim üzerinden kurumsallaşmaktadır.

YZ'nin ileri seviye modellerinde rekabet avantajı; yüksek ölçekli veri merkezleri, hızlandırıcı çipler ve bulut altyapısı gibi sermaye yoğun girdilere dayanır. Hesaplama gücü yönetişimi çerçevesi, bu girdinin yoğunlaşmış tedarik zinciri nedeniyle merkezi aktörlere yapısal avantaj sağladığını ve devletlerin hesaplama gücü akışlarını politika araçlarıyla etkileyebileceğini belirtir (Sastry vd., 2024). Bu tartışmayı somutlayan hesaplama gücü egemenliği literatürü, hesaplama gücünün yalnızca miktar değil, mülkiyet ve denetim boyutlarını da öne çıkarır. Bazı akademisyenler hesaplama gücü egemenliğini; (i) ülke sınırları içindeki hesaplama gücü kapasitesi, (ii) bu kapasiteye sahip bulut şirketlerinin milliyeti/mülkiyeti, (iii) hızlandırıcı çip ekosisteminin menşei gibi üç düzeyde ayrıştırır ve küresel bulut pazarının büyük kısmını temsil eden baskın sağlayıcılar üzerinden ampirik bir okuma önerir (Hawkins vd., 2025). Hesaplama gücü yoğunlaşması, YZ kabiliyet üretimini belli merkezlerde toplarken; çevrede kalan aktörleri maliyet, tedarik ve erişim kısıtları üzerinden bağımlı hale getirebilir.

Ayrıca Farrell ve Newman'ın (2019) ortaya koyduğu gibi, küresel ağlarda merkezi düğümlerin kontrolü devletlere zorlayıcı güç sağlayabilir. YZ bağlamında bu merkezi düğümler; bulut altyapısı, model dağıtım kanalları, veri akışları ve yarı iletken tedarik zincirleri olarak düşünülebilir. Buradaki kritik nokta, güç projeksiyonunun yalnızca askeri değil, erişimi sınırlamak, hizmeti kesmek, uyum maliyeti yaratmak veya izleme kapasitesi elde etmek gibi altyapısal olmasıdır. Özel sektörün notable model üretimindeki ağırlığına işaret eden bulgular, bu altyapı/kapasite birikiminin şirketleşmiş doğasını da görünür kılar (Maslej vd., 2025). Böylece güç dağılımı, yalnızca devletler arası rekabet olmaktan çıkarak devlet–şirket eklemlenmesi üzerinden hibrit bir form kazanmaktadır.

YZ askeri güç dengesini bir etkinleştirici teknoloji olarak etkileme potansiyeline sahiptir. Nitekim YZ istihbarat, hedefleme, lojistik ve komuta-kontrol alanlarında dönüşüm

yaratacaktır (Horowitz, 2018). Johnson (2019a), YZ'nin hızlı yayılımının belirsizlik ve kırılganlıklar üreterek stratejik istikrarsızlık ve büyük güç rekabetini keskinleştirebileceğini ileri sürer. Bu yaklaşımın ima ettiği temel sonuç, askeri gücün yalnızca daha fazla platform üretimiyle değil; daha hızlı karar, daha yoğun veri işleme ve operasyonel koordinasyonla yeniden tanımlanmasıdır. Bu dönüşüm, kriz anlarında tırmanma risklerini artırabilir.

YZ'nin ekonomik kazançlarının girdi yoğunlaşması ve ağ merkeziliği nedeniyle eşitsiz dağılmaya meyilli olduğu da görülmektedir. Emek piyasası düzeyinde otomasyon ağırlığına dönük tartışmalar (Acemoglu & Restrepo, 2020) ile kalkınma düzeyindeki uyarılar (Korinek & Stiglitz, 2021) birlikte okunduğunda, YZ'nin küresel güç farklarını artırma riski belirginleşmektedir (Acemoglu & Restrepo, 2020). Bu durum, özellikle orta ve gelişmekte olan ülkeler açısından iki stratejik zorunluluk üretir: hesaplama gücü erişimi ve bulut/altyapı seçeneklerini çeşitlendirmek; bununla eş zamanlı olarak veri yönetişimi ve yerli yetenek ekosistemini güçlendirmek.

Genel sonuç şudur: YZ, uygun politikalarla yaygın refah artışı potansiyeli taşısa da mevcut ekosistem yapısı (hesaplama gücü yoğunlaşması, şirketleşmiş kapasite ve tedarik zinciri boğazları) nedeniyle küresel güç birikimini hızlandırma ve eşitsizliği derinleştirme eğilimi taşır. Bu nedenle YZ çağında güç; askeri ve ekonomik stokların yanı sıra hesaplama gücü egemenliği, ağ merkeziliği, veri yönetişimi ve standart belirleme kapasitesinin bileşiminden oluşan yeni bir mimariye dayanmaktadır.

Bu noktada güç parametrelerindeki dönüşümün yalnızca rekabet ve kalkınma sonuçları üretmediği; bununla beraber uluslararası düzenin güvenlik üretme biçimini de yeniden şekillendirdiği görülür. Çünkü güçteki yeniden ölçeklenme, kriz yönetimi, caydırıcılık, silahların kontrolü ve uluslararası hukukun uygulanabilirliği gibi alanlarda kurumsal mimariyi doğrudan baskılayan bir etki yaratır. Dolayısıyla YZ'nin etkisini tam anlamak için, küresel güvenlik mimarisinin dayandığı normlar ve kurumlar ile bu mimarinin sahadaki işleyişine bakmak gerekmektedir.

Takip eden bölüm, araştırmanın bağımlı değişkeni olan güvenlik mimarisini tüm katmanlarıyla tanımlamaktadır. Bu tanımlama süreci, YZ'nin söz konusu katmanlarda yarattığı etkilerin incelenebileceği analitik bir zemin hazırlamaktadır. Böylece çalışmanın ilerleyen aşamalarında yapılacak değerlendirmeler için gerekli kuramsal çerçeve oluşturulmaktadır.

Küresel Güvenlik Mimarisi

Bu bölüm, küresel güvenlik mimarisini normatif, kurumsal ve operasyonel katmanlar şeklinde tasnif etmektedir. Bu yaklaşım, analizin ortak referans çerçevesini oluşturmaktadır. Temel amaç, YZ'nin etkilerini tekil örnekler üzerinden sunmak değildir. Bunun yerine, her bir etkinin ilgili katmanlarda ne tür bir uyum veya gerilim ürettiğini metodolojik olarak izlenebilir kılmaktır.

Küresel güvenlik mimarisi, devletlerin ve uluslararası örgütlerin barışı korumak, çatışmaları önlemek/sonlandırmak ve savaşın insani etkilerini sınırlamak amacıyla geliştirdiği normlar, kurumlar ve operasyonel araçlar bütününü ifade eder. Bu mimari tekil bir dünya hükümeti değil; BM merkezli kolektif güvenlik sistemi, bölgesel düzenlemeler, silahların kontrolü rejimleri, insancıl hukuk ve uluslararası ceza adaletinin iç içe geçtiği çok katmanlı bir yönetişim alanıdır.

Mimari, normatif çekirdeğini 1945 BM Şartı'ndan alır: devletlerin egemen eşitliği ve uluslararası ilişkilerde kuvvet kullanma/tehdit etme yasağı, savaşın meşru bir dış politika aracı olmasına sınır çizer (United Nations, 1945).

Kolektif güvenlik mantığı, saldırganlığın tüm uluslararası topluma karşı ihlal sayılması ve BM Güvenlik Konseyi'nin barışa tehdit/ihlal veya saldırı fiilini tespit ederek bağlayıcı tedbirlere karar verebilmesiyle işler. Şart'ın VII. Bölümü ekonomik yaptırımlar, iletişim/diplomatik ilişkilerin kesilmesi gibi zorlayıcı araçları; yetersiz kaldığında askeri tedbirleri öngörerek mimarinin hukuki omurgasını oluşturur (United Nations, 1945). Bu klasik normatif omurga, YZ'nin barış ve güvenlik boyutunu da kurumsal gündeme taşımaktadır. BM Genel Kurulu'nun güvenli, emniyetli ve güvenilir YZ odağındaki kararı (United Nations, 2024a), YZ'nin insan hakları, kapasite geliştirme ve yönetişim boyutlarını küresel düzeyde ilkesel bir çerçeveye bağlayarak güvenlik mimarisinin norm üretim katmanına somut bir ek yapmıştır.

Ayrıca BM sisteminin kendi içinde, YZ'nin kurumsal kullanımı için etik ilke setleri benimsemesi (United Nations System Chief Executives Board for Coordination, 2022), güvenlik mimarisinin yalnızca dışarıya dönük norm üretimiyle değil kurumsal örneklik üzerinden de yeniden güncellenmeye çalışıldığını göstermektedir.

Küresel güvenlik mimarisinin kurumsal çekirdeğinde BM bulunsa da güvenliğin fiili üretimi çoğu zaman bölgesel örgütler ve alt bölgesel mekanizmalar üzerinden gerçekleşir. Bu iş

bölümünün Avrupa-Atlantik ayağında NATO, kolektif savunma ve caydırıcılığı kurumsallaştırır. NATO'nun 2022 Stratejik Konsepti, İttifak'ın temel amacını kolektif savunma olarak tanımlarken caydırma ve savunma; kriz önleme ve yönetimi, işbirliğine dayalı güvenlik şeklinde üç temel görevi açık biçimde sıralar (NATO, 2022). NATO, bu görev setini teknoloji yönetişimi ile tamamlayarak savunmada YZ'nin sorumlu kullanımına ilişkin ilkeleştirme ve uygulama hedefleri geliştirmiş; 2024'te yayımlanan revize YZ stratejisinde Principles of Responsible Use ile birlikte birlikte-çalışabilirlik, YZ ekosisteminin genişletilmesi ve uygulama kapasitesinin artırılması gibi somut öncelikleri vurgulamıştır (NATO, 2024). Bu, küresel mimarinin bir parçası olarak yalnızca askeri savunma değil, kriz yönetimi ve ortak güvenlik ağlarıyla daha geniş bir güvenlik portföyü sunmaktadır.

Avrupa güvenlik düzeninin bir diğer normatif-kurumsal sütunu, 1975 Helsinki Nihai Senedi ile şekillenen ve bugün Avrupa Güvenlik ve İşbirliği Teşkilatı (AGİT) çerçevesinde sürdürülen kapsamlı güvenlik yaklaşımıdır. Helsinki metni, katılımcı devletlerin eşit haklar ve halkların kendi kaderini tayin hakkına saygı ilkesini vurgular ve devletler arası işbirliğini uluslararası hukuka uyumla ilişkilendirir (Conference on Security and Co-operation in Europe, 1975). Bu kapsamlı güvenlik hattı, YZ kaynaklı güvenlik etkilerini de gündemine almaya başlamıştır: AGİT Parlamenter Asamblesi'nin 2024 Bükreş Deklarasyonu, YZ'nin güvenlik etkilerine ilişkin farkındalık/izleme vurgusu yapmış ve teknoloji kaynaklı risklerin demokratik denetim ve şeffaflık ilkeleriyle ele alınması gerektiğini belirtmiştir (Organization for Security and Co-operation in Europe Parliamentary Assembly, 2024).

Küresel mimarinin çok-merkezli niteliği Avrupa dışındaki bölgesel güvenlik düzeneklerinde de görülür. Örneğin Afrika Birliği bünyesindeki Barış ve Güvenlik Konseyi (PSC), çatışmaların önlenmesi-yönetimi-çözümü için kalıcı karar organı ve kolektif güvenlik/erken uyarı düzenlemesi olarak kurulmuştur (The African Union, 2002). Aynı protokol, PSC'yi kıtasal erken uyarı sistemi ve Afrika Daimi Hazır Gücü gibi araçlarla desteklenmiş bir mimari olarak tasarlar (The African Union., 2002). Bu bize küresel güvenliğin yalnızca BM merkezli değil, bölgesel kapasite inşası ile de şekillendiğini gösterir. Afrika Birliği bu kapasite inşasını teknoloji yönetişimine de genişletmiş; 2024'te Kıtasal YZ Stratejisini kabul ederek etik, kapsayıcılık, kapasite geliştirme ve yönetişim başlıklarında çerçeve oluşturmuştur (African Union, 2024). PSC ise 20 Mart 2025 tarihli toplantısında YZ'nin barış, güvenlik ve yönetişime etkilerini doğrudan gündem başlığı yaparak danışma/uzman desteği, kapasite geliştirme ve düzenleyici çerçevelerin güçlendirilmesi gibi

somut adımlar çağrısında bulunmuştur (African Union Peace and Security Council [AU PSC], 2025).

Küresel güvenlik mimarisinin sahadaki en görünür enstrümanlarından biri olan BM barış operasyonları, meşruiyetini 2008 tarihli Capstone Doktrininde tanımlanan üç temel ilke üzerine inşa etmiştir. Bu ilkeler; tarafların rızası, tarafsızlık ve meşru müdafaa ya da manda savunması haricinde kuvvet kullanımından kaçınılmasıdır (United Nations, 2008). Söz konusu normatif çerçeve, barış operasyonlarını klasik askeri çatışma modellerinden ayrıştırarak operasyonun başarısını saf askeri üstünlükten ziyade siyasi uzlaşı ve güven inşasına bağlar. Bununla eşzamanlı olarak BM Barış Operasyonları Dijital Dönüşüm Stratejisi ile (veri/analitik dahil) yeni teknolojilerin sahada kullanımını kapasite ve risk yönetimi mantığıyla ele alarak misyonların daha etkin ve güvenli biçimde yürütülmesine dönük somut bir kurumsal modernizasyon hattı açmıştır (United Nations, 2021).

Küresel güvenlik mimarisinin bir diğer temel katmanı, silahların kontrolü ve yayılmanın önlenmesi rejimleridir. Nükleer Silahların Yayılmasının Önlenmesi Antlaşması (NPT), nükleer silah sahibi devletlerin bu silahları devretmemesi ve nükleer silah sahibi olmayan devletlerin de nükleer silah edinmemesi yönünde çekirdek yükümlülükler getirir (United Nations, 1970). YZ'nin askeri alandaki yayılımı arttıkça, silah kontrolü katmanı da teknoloji yönetişimiyle kesişmektedir. BM Genel Kurulu'nun askeri alanda YZ ve uluslararası barış ve güvenlik gündemli kararı (United Nations, 2024b), uluslararası hukukun uygulanabilirliği, risklerin değerlendirilmesi ve kapsayıcı diyalog çağrısıyla bu alanı kurumsal bir müzakere hattına bağlamıştır.

Güvenlik mimarisi yalnızca çatışmayı önlemekle değil, çatışma çıktığında insani tahribatı sınırlamak ve ağır ihlaller için hesap verebilirliği güçlendirmekle de ilgilidir. Uluslararası insancıl hukuk (UİH), silahlı çatışmanın etkilerini sınırlamayı; çatışmaya katılmayan ya da artık katılmayan kişileri korumayı hedefler. Hesap verebilirlik boyutunda ise Uluslararası Ceza Mahkemesi'ni kuran Roma Statüsü önemlidir. Statü, Mahkemenin yargı yetkisini ve yargılamanın koşullarını düzenlerken, BMGK'nin VII. Bölüm kapsamında bir kararla soruşturma/kovuşturmayı belirli süreyle ertelemesine ilişkin mekanizmayı da içerir (United Nations, 1998). Bu, küresel güvenlik mimarisindeki siyasal güvenlik yönetimi ile ceza adaleti arasındaki gerilimli ama kurumsallaşmış ilişkiye işaret etmektedir.

Özetle küresel güvenlik mimarisi; Birleşmiş Milletler Şartı çerçevesindeki kuvvet kullanma yasağına dayalı kolektif güvenlik sistemi, NATO ve Afrika Birliği gibi bölgesel

düzenlemeler, barış operasyonları, NPT odaklı silahların kontrolü rejimleri ile uluslararası insancıl hukuk ve ceza adaleti mekanizmalarının iç içe geçtiği çok katmanlı bir yapıdır. Bu mimari, uluslararası barışın tesisi için müşterek bir hukuki-siyasi zemin sunsa da büyük güçler arası rekabet, veto mekanizmaları, asimetrik tehditler ve bölgesel kapasite farkları nedeniyle uygulama aşamasında sıklıkla yapısal tıkanıklıklar yaşamaktadır.

YZ'nin katmanlı mimari üzerindeki etkisi, araştırmanın hipotezlerinde tanımlanan üç mekanizma aracılığıyla incelenmektedir. İlk olarak, altyapı ve hesaplama gücü yoğunlaşması, mimarinin kurumsal kapasitesini ve asimetrik bağımlılık ilişkilerini derinleştirmektedir. İkinci olarak, hız ve otomasyon faktörleri, krize müdahale süreçleri ile stratejik istikrar pratikleri üzerinde baskı oluşturmaktadır. Büyük güç rekabeti koşullarında yaşanan yönetişim tıkanıklıkları ise bağlayıcı rejimler yerine modüler araçlara yönelimi teşvik etmektedir. Bu eğilim, güvenlik mimarisinin norm üretim kapasitesini parçalı bir biçimde dönüştürmektedir. Bu gerekçelerle izleyen bölüm, söz konusu etkileri üç temel analitik izlek üzerinden değerlendirmektedir. Bu izlekler; stratejik istikrar, hukuki uyum ve hesap verebilirlik ile YZ-siber güvenlik başlıklarından oluşmaktadır.

Yapay Zekanın Küresel Güvenlik Mimarisine Etkileri

Küresel güvenlik mimarisi; devletlerin askeri kapasiteleri kadar, NATO benzeri ittifak sistemleri, caydırıcılık ve stratejik istikrar pratikleri, silahların kontrolü/insancıl hukuk rejimleri, kriz yönetimi mekanizmaları ve giderek artan biçimde dijital altyapı güvenliği üzerinden işleyen bir kurumlar, normlar ve kapasiteler bütünüdür. YZ, bu mimarinin her katmanına aynı anda temas eden genel amaçlı bir teknoloji olduğu için, etkisi tekil bir silah sisteminin ötesine geçerek karar alma hızını, bilgi üstünlüğü üretimini, savaşın yürütülme biçimini ve silahların kontrolünün doğrulanabilirliğini dönüştürmektedir (Horowitz, 2018).

Bu dönüşüm iki nedenle mimariyi yeniden düzenleyici bir basınç üretmektedir. Birincisi, YZ'nin askeri alandaki yayılımı hız/otomasyon/entegre sensör-ağları üzerinden kriz zamanında yanlış hesap ve tırmanma risklerini büyütebilir (Johnson, 2019b; Zala, 2024). İkincisi, YZ'nin çift kullanımlı doğası, özel sektörün ve sivil ekosistemlerin savaş kapasitesi üretimindeki ağırlığını artırarak güvenlik mimarisinin geleneksel devlet merkezli araçlarını zorlar; bu da norm koyma ve denetimi kurumsal olarak güçleştirmektedir (Horowitz vd., 2020).

Analiz, söz konusu basıncı üç farklı düzlemde incelemektedir. İlk aşamada stratejik istikrar ve tırmanma dinamikleri ele alınmaktadır. Bu kapsamda karar sürelerinin daralması ve nükleer-konvansiyonel eklemlenme süreçleri tartışılmaktadır. İkinci düzlemde otonom sistemler aracılığıyla uluslararası insancıl hukuk ve hesap verebilirlik konuları değerlendirilmektedir. Özellikle "anlamlı insan kontrolü" kavramı üzerine yürütülen tartışmalara odaklanılmaktadır. Üçüncü düzlemde YZ ile siber alan arasındaki etkileşim analiz edilmektedir. Dezenformasyon ve deepfake kaynaklı bilgi güvenliği risklerinin güvenlik mimarisinin meşruiyeti üzerindeki etkileri gösterilmektedir. Son olarak, bu risklerin kriz yönetimi kapasitesini nasıl dönüştürdüğü ortaya koyulmaktadır.

Nükleer çağın güvenlik mimarisi; erken uyarı, ikinci vuruş kapasitesi, yanlış alarm yönetimi ve kriz iletişimi gibi stratejik istikrar varsayımları üzerine inşa edilmiştir. YZ bu istikrar ortamını iki temel eksende etkilemektedir: bir yandan erken uyarı ve istihbarat verilerini işleme kapasitesini artırarak karar vericilere daha kapsamlı bir farkındalık sunabilmekte; diğer yandan karar sürelerini daraltıp belirsizliği artırarak kriz anlarında yanlış hesaplama riskini yükseltmektedir (Johnson, 2020; Stockholm International Peace Research Institute [SIPRI], 2019). Johnson (2020), YZ'nin geliştirilmiş konvansiyonel kabiliyetlerinin nükleer kuvvetlere dolaylı baskı yaratabileceğini; nükleer ve konvansiyonel alanların iç içe geçmesinin istem dışı tırmanma olasılığını artırdığını savunur. Zala (2024) ise YZ'nin bazı senaryolarda ilk vuruş teşviklerini artırıp artırmayacağını, özellikle erken uyarı ve hedefleme alanındaki riskler üzerinden tartışmaktadır.

YZ'nin nükleer alanda kullanımı, insan hatasını azaltma ve bilgi kalitesini artırma gibi faydalarla gerekçelendirilebilse de güvenlik mimarisi açısından kritik soru şudur: karar otoritesi ve hesap verebilirlik kimde kalacaktır? Bu noktada anlamlı insan kontrolü ilkesi, nükleer komuta-kontrol için özellikle hassas hale gelir; çünkü otomasyonun hata payı kriz koşullarında geri döndürülemez sonuçlar doğurabilir (Johnson, 2020; Horowitz vd., 2020). Yakın dönem nükleer risk literatürü ayrıca görünürlük/görünmezlik yarışına işaret eder: YZ hem tespit etme kapasitesini artırabilir hem de aldatma/maskeleme tekniklerini güçlendirebilir. Bu, karşılıklı kırılganlık ve belirsizlik üzerinden güvenlik mimarisinin istikrar ayarlarını zorlayabilir (Allison & Herzog, 2025).

Michael Horowitz'in 2018 yılında Texas National Security Review dergisinde yayımlanan çalışması, YZ'nin modern askeri inovasyon literatüründeki konumunu tanımlamak açısından temel bir referans noktasıdır. Horowitz'e göre YZ, tek başına savaşın sonucunu tayin eden

müstakil bir silah sisteminden ziyade, mevcut askeri kabiliyetlerin etkinliğini artıran ve savaşın doğasından ziyade icra ediliş biçimini kökten değiştiren bir etkinleştirici veya genel amaçlı teknoloji olarak değerlendirilmelidir (Horowitz, 2018). Bu çerçeve, YZ'nin uluslararası güvenlik mimarisi üzerindeki etkilerini daha sistematik analiz edebilmek için taktik/operasyonel ve stratejik boyutları birlikte okumayı mümkün kılmaktadır.

Otonom silah sistemlerinin teknolojik bir gerçeklik olarak ortaya çıkışı, küresel güvenlik mimarisinin temel taşlarını oluşturan uluslararası insancıl hukuk ve silahların kontrolü rejimlerini ciddi bir sınamayla karşı karşıya bırakmaktadır. Akademik literatürdeki tartışmalar, bu sistemlerin hukuki statüsünü ve kullanım sınırlarını belirlemek amacıyla temelde iki eksen etrafında yoğunlaşmaktadır. Birinci eksen, geleneksel savaş hukukunun temel ilkeleri olan askeri hedefler ile sivilleri ayırma, orantılılık ve önleyicilik gibi kuralların, karar alma süreçlerinde insan müdahalesinin bulunmadığı otonom sistemlere nasıl uyarlanabileceğini sorgulamaktadır. İkinci eksen ise, makinelerin öldürücü güç kullanma yetkisine sahip olması durumunda ortaya çıkacak anlamlı insan kontrolü eksikliği ve buna bağlı gelişen hukuki ve cezai sorumluluk boşluklarını ele almaktadır.

Bu tartışma zeminini güçlendiren çalışmalar, otonom sistemlerin uluslararası insancıl hukuk ilkeleriyle yapısal uyumunu analiz etmektedir. Bu teknolojilerin savaş alanındaki karmaşık dinamikleri kavrayarak sivil–muharip ayrımını yapabilme kapasitesi ve saldırının beklenen askeri avantajla orantılı olup olmadığını değerlendirme yeteneği kritik sorun alanları olarak işaret edilmektedir (Güneysu, 2024). Öte yandan, öldürücü otonom silah sistemleri için küresel ölçekte bağlayıcı ve dirençli bir kontrol rejiminin tesis edilmesinin önündeki engeller de öne çıkmaktadır. Devletlerin stratejik çıkarlarından kaynaklanan siyasal dirençler ve yazılımsal tabanlı bu silahların doğrulama süreçlerindeki teknik zorluklar, uluslararası bir denetim mekanizmasının kurulmasını güçleştirmektedir (Qerimi, 2023).

Uluslararası hukuk perspektifinden güncel yaklaşımlar ise, otonom silah sistemlerinin doğası gereği hukuka aykırı olup olmadığı sorusuna odaklanmaktadır. Casey-Maslen'e göre bu sistemlerin varlığı kendi başına uluslararası hukukun ihlali sayılmasa da kullanılmaları durumunda her zaman uluslararası insancıl hukuk ve uluslararası insan hakları hukuku ile tam uyum içinde olmak zorundadır (Casey-Maslen, 2025). Bu durum, teknolojinin kendisinden ziyade bu teknolojinin savaş alanındaki uygulama biçimlerinin hukuki meşruiyeti belirleyeceğini göstermektedir. Sonuç olarak, YZ tabanlı silah sistemlerinin askeri doktrinlere entegrasyonu; mevcut hukuk normlarının yeniden yorumlanmasını ve sorumluluk

hiyerarşisinin bu yeni teknolojik gerçekliğe göre yeniden tanımlanmasını kaçınılmaz kılmaktadır.

YZ teknolojilerinin küresel güvenlik mimarisi üzerindeki derin etkilerinden biri de devletlerin savaşa girme kararı alma süreçlerindeki siyasal maliyet analizlerini ve stratejik enformasyon akışını dönüştürmesidir. Bu bağlamda askeri alanda sorumlu YZ kullanımına ilişkin tartışmalar ve beraberindeki riskler, savaş kararının geleceğine dair normatif ve siyasal sonuçlar çerçevesinde ele alınmaktadır (Erskine & Miller, 2024). YZ destekli sistemlerin sağladığı hız ve otomasyon, bir yandan karar vericilere operasyonel avantaj sunarken, diğer yandan savaşın insani ve siyasi maliyetini görünmez kılma eğilimi üzerinden demokratik hesap verebilirlik mekanizmalarını zayıflatma potansiyeli taşımaktadır. Dolayısıyla YZ, modern güvenlik ekosisteminde yalnızca bir kuvvet çarpanı değil; bunun yanı sıra devletin savaş ve barış arasındaki kritik kararı verme biçimini ve bu kararın meşruiyet zeminini etkileyen kurumsal bir dinamik olarak konumlanmaktadır.

AI-Siber Etkileşimi: Stratejik İstikrar ve Bilgi Güvenliği

YZ ve siber kapasitelerin birleşimi, modern güvenlik mimarisinin en belirsiz alanlarından biri olan gri bölge faaliyetlerini daha karmaşık ve öngörülemez bir boyuta taşımaktadır. Soğuk Savaş sonrası dönemde tırmanma yönetimini zorlaştıran siber tehditler, YZ algoritmalarının entegrasyonuyla birlikte stratejik istikrar üzerinde doğrudan bir risk unsuru haline gelebilir. Bu bağlamda YZ destekli siber kabiliyetlerin özellikle nükleer kuvvetlerin güvenilirliği, komuta-kontrol sistemleri ve operasyonel hazır bulunuşluk üzerinde dolaylı fakat derin etkiler yarattığını savunan görüşler ortaya çıkmıştır (Johnson, 2019b). Bu birleşme, sadece teknik sabotaj ihtimalini değil; yanlış anlama veya hatalı sinyal okuma nedeniyle ortaya çıkabilecek istem dışı tırmanma riskini de büyütebilir.

Güvenlik mimarisinin bir diğer kritik sütunu, askeri kapasitelerin ötesinde yer alan toplumsal meşruiyet ve algı yönetimidir. YZ destekli dezenformasyon faaliyetleri ve deepfake teknolojileri, kriz anlarında kamuoyu algısını manipüle ederek liderlerin karar alma mekanizmalarını ve müttefikler arası dayanışmayı hedef alabilmektedir. YZ'nin dezenformasyon üretimini devasa ölçeklere ulaştırma potansiyeli, güvenlik risklerini ulusal sınırların ötesine taşımakta ve koordineli bir uluslararası yanıtı zorunlu kılmaktadır (Kertysova, 2018). Hynek vd. (2025) ise deepfake uygulamalarının kurumsal güven erozyonu ve kriz manipülasyonu aracılığıyla uluslararası güvenlik sisteminin meşruiyet zeminini nasıl zayıflattığını ortaya koymaktadır. Bu nedenle geleceğin caydırıcılık ve kriz yönetimi

stratejileri, yalnızca silah sistemlerine değil; içerik menşei işaretleme ve hızlı teyit ağları gibi güçlü bir bilgi bütünlüğü altyapısına da dayanmak zorundadır.

YZ, siber saldırıları daha düşük maliyetle ve yüksek ölçekte otomatize ederken savunma kalkanlarını aşma hızını da artırmakta; bu hızlanma 'gri bölge' eylemlerinin tespit-atıf-yanıt döngüsünü sıkıştırarak diplomasiye bırakılan süreyi daraltmakta ve kriz iletişimi ile gerilimi düşürme kanallarını zayıflatmaktadır.

Sonuç

Çalışma kapsamında üç temel bulgu ön plana çıkmaktadır. İlk olarak YZ ekosistemindeki altyapı ve hesaplama gücü yoğunlaşması, uluslararası sistemdeki hiyerarşik yapıyı derinleştirmektedir. Güç rekabeti, model performansından ziyade kritik girdilere erişim rejimleri ve akış kontrolü etrafında şekillenmektedir. İkinci olarak hız ve otomasyon etkisi karar sürelerini daraltmakta ve belirsizliği artırmaktadır. Bu durum kriz anlarında yanlış hesaplama ve tırmanma riskini büyüterek stratejik istikrar pratiklerini baskılamaktadır. Üçüncü olarak çift kullanımlılık, özel sektör ağırlığı ve büyük güç rekabeti gibi unsurlar bağlayıcı rejimlerin inşasını zorlaştırmaktadır. Bu süreçte yönetişim araçları modüler ve parçalı bir hatta ilerlemektedir. Söz konusu eğilim, güvenlik mimarisinin doğrulama, hesap verebilirlik ve meşruiyet kapasitesini sınamaktadır. Bu üç temel bulgu makalenin ana argümanını doğrulamaktadır. YZ, tekil bir kapasite artışından ziyade güvenlik mimarisinin tüm katmanlarına eş zamanlı basınç uygulayan bir "çarpan teknoloji" niteliğindedir. Aşağıda, bu üç bulgunun (H1–H3) güvenlik mimarisinin normatif, kurumsal ve operasyonel katmanlarında ürettiği başlıca sonuçlar ve bunlara karşılık gelen risk azaltımı seçenekleri özetlenmektedir.

Birinci bulgu, güvenlik mimarisinin kurumsal katmanında karar ve kapasite üretimini dar bir sağlayıcı/ittifak ekosistemine bağımlı kılmakta; normatif katmanda ise "eşit egemenlik" ve "adil erişim" ilkeleriyle fiili altyapı hiyerarşisi arasındaki gerilimi görünürleştirmektedir.

İkinci bulgu, operasyonel katmanda karar döngülerini sıkıştırarak yanlış anlama ve tırmanma riskini büyütmekte; bu nedenle kriz yönetimi, caydırıcılık ve komuta-kontrol pratikleri üzerinde baskı üretmektedir.

Üçüncü bulgu, normatif ve kurumsal katmanlarda evrensel mutabakatın yerini standartlar, ittifak içi düzenlemeler ve tedarik zinciri üzerinden kurulan fiili uyum baskılarının aldığı daha parçalı bir yönetişim hattına işaret etmektedir.

Bu çerçevede risk azaltımı ve uyum için öneriler, doğrudan bulguların işaret ettiği mekanizmalara bağlanmalıdır. (H1) Altyapı/ hesaplama kapasitesi yoğunlaşmasına karşı, uluslararası düzeyde asgari şeffaflık ve karşılaştırılabilir raporlama (kritik girdiler, tedarik zinciri kırılganlıkları, büyük ölçekli veri-merkezi kapasitesi) ile kapasite geliştirme araçlarının birlikte tasarlanması; bağımlılıkları azaltırken "fiili hiyerarşi"nin meşruiyet maliyetini azaltma potansiyeli taşır. (H2) Hız ve otomasyon kaynaklı tırmanma riskine karşı, kriz iletişim hatlarının güncellenmesi, YZ destekli karar süreçlerinde "anlamlı insan kontrolü" eşiğinin operasyonel olarak tanımlanması ve olay/arıza bildirimi gibi güven artırıcı önlemler, stratejik istikrar pratiklerini destekleyebilir. (H3) Evrensel bağlayıcı rejimlerin tıkandığı koşullarda ise modüler düzenleme yaklaşımı daha uygulanabilir bir hat sunar: teknik standartlar, birlikte-çalışabilirlik ölçütleri, denetim/sertifikasyon ve ittifak içi ortak ilkeler yoluyla, en azından belirli risk sınıflarında doğrulama, hesap verebilirlik ve uyum kapasitesi artırılabilir. Bu modüler hat, parçalanmayı ortadan kaldırmasa da mimarinin 'işleyen asgari ortak zemini'ni üretmeye dönük pragmatik bir ara çözüm olarak görülebilir.

Dolayısıyla sonuçlar, YZ'nin güvenliği yalnızca yeni sistemler üzerinden değil; altyapı hiyerarşisi (hesaplama kapasitesi), karar hızlanması (otomasyon) ve yönetişim biçimi (modülerleşme) üzerinden eşzamanlı biçimde yeniden yapılandırdığını göstermektedir. Bu da makalenin ana iddiasıyla tutarlıdır: YZ, tekil bir askeri kapasite artışından ziyade güvenlik mimarisinin tüm katmanlarında basınç üreten bir çarpan teknolojidir; bu nedenle çözüm arayışı da tek bir rejime indirgenmekten çok, mekanizma-temelli ve katman-duyarlı bir risk azaltımı mimarisi olarak kurgulanmalıdır.

Bu bulgular, güvenlik mimarisinde uygulanabilir müdahale alanlarının ideal çözümlerden çok risk azaltımı ve doğrulanabilirlik odaklı, kademeli araçlara dayandığını göstermektedir. Bu yönelim, çalışmanın üçüncü hipotezinin (H3) işaret ettiği üzere, büyük güç rekabeti altında evrensel bağlayıcı rejimlerin sınırlı kalması nedeniyle yönetişimin daha çok modüler ve doğrulanabilir araçlara kaymasıyla uyumludur.

Bu yeni dönemde güvenlik mimarisini YZ'nin hız, ölçek ve belirsizlik etkilerine karşı dayanıklı kılmak, kuralların ve kurumsal süreçlerin yeniden yapılandırılmasının gerektiğine işaret etmektedir. Bu dayanıklılığı inşa etmenin ilk adımı, askeri YZ sistemleri için kullanım bağlamı sınırlarını, eğitim verisi standartlarını ve doğrulama kayıtlarını içeren kapsamlı bir yaşam döngüsü denetimi ve emniyet dosyası standardı geliştirmektir.

Bununla birlikte, kriz anlarında tırmanma riskini ve sistem davranışlarındaki belirsizliği azaltmak amacıyla güven artırıcı önlemlerin teknoloji odaklı bir yaklaşımla güncellenmesi ihtiyacını doğurmaktadır. Bu çerçevede; karşılıklı bilgi paylaşımı, ortak terminoloji oluşturulması ve kaza senaryolarına yönelik ortak tatbikatlar stratejik istikrarın korunmasında kritik rol oynamaktadır.

Güvenlik mimarisinin merkezindeki bir diğer unsur ise insan yargısının doktrin ve eğitimle kurumsallaştırılmasıdır. YZ'nin önerdiği hedefleme verilerinin kanıt standartlarının belirlenmesi ve komutan sorumluluğunu netleştiren raporlama mekanizmalarının tesisi, teknolojinin etik ve hukuki sınırlar içinde tutulmasını sağlar. Ayrıca, veri setlerindeki önyargıların uluslararası insancıl hukuk uyumunu sistemik olarak aşındırmasını önlemek için tedarik süreçlerinde bağımsız denetim mekanizmalarının kurulması ve hukuk uzmanları ile veri bilimcilerin eşgüdümlü çalışması önem kazanmaktadır.

Ne var ki çalışmanın önceki bölümlerinde vurgulanan büyük güçler arası rekabet ve veto kaynaklı kurumsal tıkanıklıklar, tam da bu tür yaşam döngüsü denetimi ve bağımsız denetim önerilerinin uygulanabilirliğini sınırlayan temel siyasal engeli oluşturmaktadır. Rekabet koşullarında denetim rejimleri, çoğu zaman ortak güvenlik ihtiyacından ziyade göreli kazanımlar mantığıyla okunmaktadır. Taraflar, denetimin, kendi inovasyon hızını yavaşlatırken rakibin kapasitesini görünür kılacak bir 'asimetrik şeffaflık' aracı olmasından çekinmektedir. Veto mekanizmaları ise evrensel ve bağlayıcı standartların -özellikle doğrulama, raporlama ve yaptırım boyutları olan rejimlerin- kabulünü geciktirerek denetimi en düşük ortak paydaya itmektedir. Bu nedenle önerilen denetim araçlarının hayata geçebilmesi, ideal tipte küresel mutabakattan çok, önce dar kapsamlı kademeli ve modüler düzenlemelerle ilerlemeye; ayrıca bağlayıcı rejimlerin tıkandığı yerde çok taraflı teknik standartlar, ittifak içi düzenlemeler ve tedarik zinciri üzerinden 'fiili' uyum baskısı üreten mekanizmaların devreye girmesine bağlıdır.

Nihayetinde dijital uçurum, yalnızca teknik kapasite eksikliği değil; YZ ekosistemindeki hesaplama gücü-veri-altyapı yoğunlaşmasının ürettiği sert hiyerarşi içinde çevre aktörleri kalıcı bağımlılığa iten ve karar/uyum alanlarını daraltarak egemenlik kaybı riskini büyüten yapısal bir tahakküm mekanizması olarak yorumlanabilir. Bu nedenle gelişmekte olan ülkelerin teknik ve hukuki kapasitesinin desteklenmesi, küresel yönetişimin kapsayıcı ve kırılganlıktan uzak bir yapıya kavuşması için stratejik bir zorunluluk olarak öne çıkmaktadır.

Sonuç olarak bulgular, YZ'yi tamamen dışlamaya dayalı yaklaşımların yerine, teknik doğrulama kapasitesi ile şeffaflık/diyalog/çok taraflı iş birliği arasında denge kuran bir uyum hattının daha uygulanabilir olabileceğine işaret etmektedir.

KAYNAKÇA

Acemoglu, D., & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions, Economy and Society, 13*(1), 25–35. https://doi.org/10.1093/cjres/rsz022.

African Union Peace and Security Council. (2025, March 20). *Communiqué of the 1267th meeting of the Peace and Security Council on artificial intelligence and its impact on peace, security and governance in Africa.*

African Union. (2002). *Protocol relating to the establishment of the Peace and Security Council of the African Union*(Adopted July 9, 2002).

African Union. (2024). *Continental Artificial Intelligence Strategy* (Adopted July 2024).

Allison, D. M., & Herzog, S. (2025). Artificial intelligence and nuclear weapons proliferation: The technological arms race for (in)visibility. *Risk Analysis, 45*(11), 3839–3859. https://doi.org/10.1111/risa.70105

Beach, D., & Pedersen, R. B. (2019). *Process-tracing methods: Foundations and guidelines* (2nd ed.). University of Michigan Press.

Bennett, A., & Checkel, J. T. (Eds.). (2015). *Process tracing: From metaphor to analytic tool*. Cambridge University Press.

Boulanin, V. (Ed.). (2019). *The impact of artificial intelligence on strategic stability and nuclear risk: Volume I—Euro-Atlantic perspectives*. Stockholm International Peace Research Institute (SIPRI).

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027.

Casey-Maslen, S. (2025). Autonomous weapons systems under international law. In K. S. Talves (Ed.), *Artificial intelligence in military technology* (pp. 161–179). Springer. https://doi.org/10.1007/978-3-031-95578-5_10.

Conference on Security and Co-operation in Europe. (1975). *Helsinki Final Act*. Organization for Security and Co-operation in Europe (OSCE).

Erskine, T., & Miller, S. E. (2024). AI and the decision to go to war: Future risks and opportunities. Australian Journal of International Affairs, 78(2), 135–147. https://doi.org/10.1080/10357718.2024.2349598.

Farrell, H., & Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. *International Security, 44*(1), 42–79. https://doi.org/10.1162/isec_a_00351.

Güneysu, G. (2024). Autonomous weapon systems and the humanitarian principle of distinction. *Annales de la Faculté de Droit d'Istanbul, 74*, 133–153. https://doi.org/10.26650/annales.2024.74.0007

Hawkins, Z. J., Lehdonvirta, V., & Wu, B. (2025). AI compute sovereignty: Infrastructure control across territories, cloud providers, and accelerators. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5312977.

Horowitz, M. C. (2018). Artificial intelligence, international competition, and the balance of power. *Texas National Security Review, 1*(3), 36–57. https://doi.org/10.15781/T2639KP49.

Horowitz, M. C., Kahn, L., & Mahoney, C. (2020). The future of military applications of artificial intelligence: A role for confidence-building measures? *Orbis, 64*(4), 528–543. https://doi.org/10.1016/j.orbis.2020.08.003.

Hynek, N., Gavurová, B., & Kubák, M. (2025). Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European countries. *Journal of Innovation & Knowledge, 10*(5), Article 100782. https://doi.org/10.1016/j.jik.2025.100782.

Johnson, J. (2019a). Artificial intelligence & future warfare: Implications for international security. *Defense & Security Analysis, 35*(2), 147–169. https://doi.org/10.1080/14751798.2019.1600800.

Johnson, J. (2019b). The AI-cyber nexus: Implications for military escalation, deterrence and strategic stability. *Journal of Cyber Policy, 4*(3), 442–460. https://doi.org/10.1080/23738871.2019.1701693.

Johnson, J. (2020). Artificial intelligence: A threat to strategic stability. *Strategic Studies Quarterly, 14*(1), 16–39.

Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights, 29*(1–4), 55–81. https://doi.org/10.1163/18750230-02901005.

Korinek, A., & Stiglitz, J. E. (2021). *Artificial intelligence, globalization, and strategies for economic development*(NBER Working Paper No. 28453). National Bureau of Economic Research. https://doi.org/10.3386/w28453.

Maslej, N., Fattorini, L., Perrault, R., Gil, S., Parli, V., Kariuki, A., Capstick, M., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, H., Santarlasci, S., … Oak, S. (2025). *The AI Index 2025 annual report*. Stanford University—Institute for Human-Centered Artificial Intelligence (HAI). https://doi.org/10.48550/arXiv.2504.07139.

North Atlantic Treaty Organization. (2022). *NATO 2022 Strategic Concept* (Madrid Summit, 29 June 2022).

North Atlantic Treaty Organization. (2024, July 10). *Summary of NATO's revised artificial intelligence (AI) strategy.*

Organization for Security and Co-operation in Europe Parliamentary Assembly. (2024). *Bucharest Declaration* (2024 Annual Session).

Qerimi, Q. (2023). Controlling lethal autonomous weapons systems: A typology of the position of states. *Computer Law & Security Review, 50*, Article 105854. https://doi.org/10.1016/j.clsr.2023.105854.

Sastry, G., Heim, L., Belfield, H., Andres, M., Chan, B., Baker, M., Anderljung, M., Hecht, F., & Coyle, D. (2024). *Computing power and the governance of artificial intelligence* (arXiv:2402.08797) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2402.08797.

Stockholm International Peace Research Institute. (2019). *The impact of artificial intelligence on strategic stability and nuclear risk: Volume I—Euro-Atlantic perspectives*. Stockholm International Peace Research Institute. [Rapor].

U.S. Department of Commerce, Bureau of Industry and Security. (2023, October 25). *Export controls on semiconductor manufacturing items* (88 FR 73424). *Federal Register*.

United Nations. (1945). *Charter of the United Nations.*

United Nations. (1970). *Treaty on the Non-Proliferation of nuclear weapons (NPT).*

United Nations. (1998). *Rome Statute of the International Criminal Court.*

United Nations. (2008). *United Nations Peacekeeping Operations: Principles and Guidelines* (Capstone Doctrine).

United Nations. (2021). *Strategy for the digital transformation of UN peacekeeping.* https://peacekeeping.un.org/en/strategy-digital-transformation-of-un-peacekeeping.

United Nations General Assembly. (2024a). *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development* (A/RES/78/265).

United Nations General Assembly. (2024b). *Artificial intelligence in the military domain and its implications for international peace and security* (A/RES/79/239).

United Nations System Chief Executives Board for Coordination. (2022). *Principles for the ethical use of artificial intelligence in the United Nations system.* https://unsceb.org/principles-ethical-use-artificial-intelligence-united-nations-system.

Zala, B. (2024). Should AI stay or should AI go? First strike incentives & deterrence stability. *Australian Journal of International Affairs, 78*(2), 154–163. https://doi.org/10.1080/10357718.2024.2328805.

DEMOCRACY IN THE AGE OF AI; THE FINE LINE BETWEEN THE KNOWN AND UNKNOWN

Mihai SEBE, Alexandru GEORGESCU and Eliza VAȘ*

Declaration*

We are currently witnessing a structural shift in the democratic societies across the globe that impacts the internal structure of politics. The rapid development of Artificial Intelligence (AI) and the increased level of cyber security threats affect political socialisation in ways that could be observed before in a generation's time, not just in a matter of years as it is the case now. Therefore, we need to start thinking about the rhetorical shift of digitalisation towards AI, given its applicability as tool for political deliberation, preference aggregation and discovery, as well as electoral integrity. The technological advancement can impact social stability, all while not offering antibodies the democratic societies need to deal with this disruption.

The future of politics seems to be more defined by the way in which states will protect the minds, the will and the hearts of their citizens against cognitive warfare and threats. On a decision-making level, this further burdens the process of making the AI an instrument for authentic public participation. Between informing choices and decisions and delegating free will to AI lies a fine line. As such, a new social contract is needed.

In a world dominated by geopolitical tensions, the multilateral approach offers us ideas and ideational tools to draft this new contract. In this context, the United Nations represent a global standard that one cannot ignore.

The malicious use of digital tools is high on the international and national agenda and is set to remain a pressing issue for the upcoming years. In fact, the United Nations went one step

* Mihai SEBE, PhD, lecturer, University of Bucharest. E-mail: mihai.sebe@fspub.unibuc.ro,
Alexandru GEORGESCU, PhD, is an Established Researcher (R3) with the National Institute for Research and Development in Informatics ICI Bucharest, alexandru.georgescu@ici.ro,
Eliza VAȘ is an expert in European affairs and vice-president of the Young Initiative Association. E-mail: eliza.vas@younginitiative.org.
* The opinions expressed belong to the authors only and do not reflect the official position of the institutions they are affiliated with.

further and signalled in 2024 the impact on politics. “Electoral security posed a significant challenge throughout the year, as elections took place in more than 70 states. Disinformation campaigns designed to influence voters, and uses of deepfakes of political figures were observed in the lead-up to national elections in several states. At the same time, interference in critical infrastructure facilitating electoral processes also posed risks.” (United Nations, 2024).

Following this assessment, the UN General Assembly adopted the Pact for the Future (*Resolution 79/1*), including the Global Digital Compact (*Annex I*). It calls for closer international cooperation that reduces all digital divides between and within countries and reiterates that the digital future should be guided by the purposes and principles of the Charter of the United Nations (United Nations General Assembly [UNGA], 2024). Thus, the Human Rights Council chose to anchor the new developments in the framework of the international law stating explicitly this duality: the new and emerging digital technologies can hold great potential for strengthening democratic institutions and the resilience of civil society, but they can also affect the integrity of democratic institutions (*Resolution 59/11, 2025*) (United Nations Human Rights Council [UNHRC], 2025a).

We are now in a situation where the quick evolution of AI needs to meet a series of standards, such as a human rights-based approach and appropriate safeguards and human oversight. Moreover, we are faced with an unprecedented spread of disinformation the AI can only make worse (United Nations Human Rights Council [UNHRC], 2025b).

As a technology, AI comes with a paradigm shift from a security perspective. The cyber security we are used to must make way for new approaches, especially when it comes to generative AI embodied in the vast majority of systems with which average citizens will interact in a media, social and political context. On the one hand, we rely on this technology for AI aggregators of information, AI-based communication, AI pollsters and AI-based societal systems (education, public services etc.). However, AI diverges from traditional software by utilising natural language interfaces and generating probabilistic rather than deterministic outputs. This lack of a predictable baseline makes it difficult to detect anomalies.

Furthermore, because AI resilience is deeply tied to specific data environments and usage patterns, conventional security audits often prove inadequate. Experts are recommending redteaming, which is a structured ethical adversarial testing under real conditions (United

Nations Educational, Scientific and Cultural Organization [UNESCO], 2025). However, there is a lack of specialists, especially within user entities as opposed to AI development entities, leading to a "black box" problem where AI tools become opaque and anomalous behaviours become harder to identify. On the other hand, we have AI systems as tools for cyber attacks on electoral infrastructure (some of which may also be AI-based) or through generation of fake news, of deepfakes, of tailored disinformation and through the manufacturing of public moods and ideas. This is an incredibly pernicious issue. The identification of AI-generated content before destabilising a narrative or an entire society is a difficult task, likely involving AI tools, new skillsets, new modes of societal protection and new tools used ethically under an acceptable and values-based legal and administrative framework.

This complicated relationship between AI and democracy has been further explored in a UNESCO report that underlined both the great expectations and fears of this process, underlying that "the only political certainty we have today is that politics in the future will inevitably be very different from politics in the past". The report warns of the erosion of public discourse, the rise of new intermediaries, and the opacity of algorithmic decision-making. This should be coupled with secular trends involving the shrinking of party membership as principal means of political activity, education and legitimation, as well as the erosion of the authority of public institutions and official narratives in a digitalized society. The digital utopianism needs to make room for more balanced democratic nuances as the AI tends to reflect the values of its creators and the biases in its datasets instead of remaining neutral. The evolution of AI raises two main questions as per this report: "Do the principles of democratic self-governance still hold relevance and significance in a digital, automated public space largely governed by algorithmic systems? Is this a new era that we must simply accept, or does this historical moment bring new opportunities for democratisation?" (Innerarity, 2024).

While decision-makers all around the world are looking for answers to these questions, the European Union (EU) is proposing a risk-based approach. With the AI Act introduced in 2024, the EU mapped out four levels of risk for AI systems. From minimal to unacceptable risk, the purpose was to make sure that AI brings more benefits to the citizens than it takes away freedoms and rights. Under high-risk category we have those situations that can severely affect fundamental rights, safety and health. With respect to democracy, the Regulation foresees that "AI systems intended to be used to influence the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote

in elections or referenda should be classified as high-risk AI systems with the exception of AI systems whose output natural persons are not directly exposed to" (Regulation (EU) 2024/1689 [AI Act], 2024).

Coming back to the will of the people as a core democratic principle to uphold, we invite the readers to reflect on the words of Stephen Hawking, who 10 years ago affirmed that "in the future, AI could develop a will of its own — a will that is in conflict with ours" (University of Cambridge, 2016). Furthermore, in seeking to develop national and international governance frameworks and accepted principles of AI development, we should be way of the differing rates of digitalization throughout various societies. As the famous science fiction author William Gibson once observed, "the future is already here, but it is not evenly distributed" (O'Toole, 2012).

References

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

Innerarity, D. (2024). *Artificial intelligence and democracy*. Paris, France: UNESCO. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000389736

O'Toole, G. (2012, January 24). *The future has arrived — It's just not evenly distributed yet*. Retrieved from http://quoteinvestigator.com/2012/01/24/future-has-arrived/

UNESCO. (2025). *Red teaming artificial intelligence for social good: The playbook*. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000394338.locale=en

United Nations General Assembly. (2024). *Resolution adopted by the General Assembly on 22 September 2024* (A/RES/79/1). Retrieved from https://docs.un.org/en/A/res/79/1

United Nations Human Rights Council. (2025a). *New and emerging digital technologies and human rights* (A/HRC/RES/59/11). Retrieved from https://digitallibrary.un.org/record/4087001?v=pdf

United Nations Human Rights Council. (2025b). *New and emerging digital technologies and human rights* (A/HRC/59/L.14). Retrieved from https://docs.un.org/en/a/hrc/59/l.14

United Nations. (2024). *United Nations disarmament yearbook: Developments and trends 2024*. Retrieved from https://yearbook.unoda.org/en-us/2024/chapter5/

University of Cambridge. (2016, October 19). *"The best or worst thing to happen to humanity": Stephen Hawking launches Centre for the Future of Intelligence*. Retrieved from https://www.cam.ac.uk/research/news/the-best-or-worst-thing-to-happen-to-humanity-stephen-hawking-launches-centre-for-the-future-of

DIGITAL WORLD-SYSTEM HIERARCHIES AND AI-DRIVEN SECURITY COMPETITION

Merve Suna ÖZEL-ÖZCAN[*]

ORCID: 0000-00019027-3990

Declaration[*]

Artificial intelligence is no longer just a new technology; it has become a structural force multiplier that changes everyday life, the ability of the state to do things, security, and even the nature of war. This broad sphere of influence is creating a new concentration of power in the global system, covering everything from healthcare and education to finance, public administration, defense industries, and knowledge production. The main question, though, is how this change affects the system itself. We are now connected to algorithms in almost every area of our lives. Our choices, likes and dislikes, and even the things we do every day go through algorithmic filters. But the most important part is not at the individual level; it is at the global system level, where a new ecological structure is forming.

So the AI and algorithms are no longer just a technical infrastructure; they have created a new system with its own rules, hierarchies, and ways of excluding people. At this point this is not entirely unprecedented in historical terms. I belive that the in global system analyses, particularly in Wallerstein's work, we have previously identified analogous configurations where power becomes centralized, and the periphery is perpetually reproduced. But right now is a very important time that calls for a different way of thinking. Algorithms now directly affect how we define "great power" and "superpower," moving the system into a new historical phase. In this new phase, power is no longer determined solely by possessing capacity or by the ability to absorb costs.

Thus equally decisive is the capacity to govern algorithms in order to control the system itself. In other words, rather than asking who produces more, the key questions have become who calculates, who directs, and who codes decision-making processes. In short, this

[*] Doç. Dr., Kırıkkale Universitesi, Türkiye, mervesuna@kku.edu.tr.

[*] In the preparation of this chapter, AI-based tools were utilized for translation and proofreading purposes. The AI tools were employed solely as supportive instruments to enhance linguistic clarity and coherence. The ideas, analyses, and arguments presented in the chapter remain entirely the responsibility of the author.

transformation strains the explanatory frameworks of classical international relations theories. It reveals the need for a hybrid reading situated between Wallerstein's world-systems theory and Mearsheimer's offensive realism. In the age of artificial intelligence, the global system cannot be explained solely by structural inequalities, nor reduced to pure military power competition. Instead, this new phase, shaped by algorithms, is fundamentally rewriting the definition of power, its modes of use, and its conditions of persistence.

A New Centre in the World-System: The Algorithmic Core

Wallerstein's world-systems theory approaches the global order not as an anarchic structure composed of equal state actors, but as a historically constructed and enduring hierarchical system of production. This structure is essentially read through the historical cycles of capitalism. The center–periphery relationship is shaped not only by political power balances, but also by the temporal and spatial expansion of the capitalist mode of production. When we look at this system, the centre is defined not merely by military or political power, but by its capacity to control capital accumulation, technological production, and flows of knowledge (Wallerstein, 1974 and 2004). According to Wallerstein, the global system is not an arena of anarchy in which states act as equal units; rather, it is a hierarchical order in which structural inequalities are continuously reproduced. In its most basic form, this structure consists of three main zones:

1. The Core, which controls capital accumulation, technological production, and information flows.

2. The Periphery, which is confined to labour, raw materials, and low-value-added production.

3. The Semi-Periphery occupies a balancing and transitional position between these two zones.

But today, the centre within the system is no longer only a geography where factories, financial centres, or military bases are located. The centrecentre has transformed into a cognitive space where algorithms operate, data is interpreted, and decisions are made through artificial intelligence. Therefore, the twenty-first-century world-system should now be read in the following way: Digital Core: actors that control artificial intelligence, big data, cyber security, and digital platforms Digital Semi-Periphery: actors that have partial access to digital capacity but are unable to overcome algorithmic dependency Digital Periphery: spaces

that provide data but are unable to establish sovereignty over that data In this context, the algorithmic core produces not only economic superiority, but also epistemic superiority. In other words, within the world-system, it is no longer only the question of who produces that matters; instead, the decisive questions have become who knows, who calculates, and who decides. So, this approach provides a strong theoretical foundation for discussions of artificial intelligence, as it defines power not solely through military capacity but through production relations and information flows. However, by the second quarter of the twenty-first century, it has become clear that the classical core–periphery–semi-periphery distinction has entered a phase of deepening through an additional layer.

How should this system be read today?

In the contemporary global system, the newly emerging domain is shaped less by physical production and more by data, algorithms, and AI-based decision-making mechanisms. In this sense, the world-system is no longer only an economic or geopolitical hierarchy; it has also become a digital and algorithmic field of domination. It is precisely at this point that algorithms—and the new ecology they generate within themselves—must be incorporated into the equation. We are no longer dealing with a system that works in the traditional way; instead, we are dealing with a new systemic configuration. In this new structure, the algorithmic core gets its power not from military strength but from its ability to control large amounts of data, train AI models, manage cybersecurity systems, and direct global information flows through digital platforms and cloud systems (Özel Özcan, 2025). Thus, the algorithmic core arises not only as a domain that regulates production and capital relations but also as a novel gravitational center that influences decision-making processes, the dissemination of knowledge, and the overall operation of the system. This issue does not imply that Wallerstein's world-systems approach has become obsolete; rather, it suggests that the theory necessitates a reconsideration within a new historical contex. Now is a good time to look at the people who work in the system. Today, algorithms are used more and more to explain how great powers interact with each other and how they can change the system. The meanings of how power is made, kept, and shown are changing. They are shifting from conventional military or economic dominance to algorithmic preeminence. In this context, Mearsheimer's offensive realism offers a practical analytical framework. Mearsheimer asserts that great powers operate within uncertainty, prioritize survival as their primary objective, and view power accumulation as a rational strategy (Mearsheimer, 2001)—competition in the field of artificial intelligence strictly follows this logic. Algorithmic superiority speeds up

decision-making in the military, makes it possible to have autonomous weapons systems, and turns war into a quiet but ongoing field of conflict. The Department of War's Artificial Intelligence Strategy makes it clear that AI is not just a side technology; it is a key part of how the military will be powerful in the future. This plan makes it clear that the US wants to be the AI-first warfighting force.

It wants to make its military more deadly, faster, and more efficient by using advanced AI to help with planning missions, making decisions, and carrying them out. The plan is to quickly use AI in all areas of the mission, with the clear goal of keeping the U.S. ahead of the rest of the world. Russia, on the other hand, is a player who tests this change not only in strategic documents but also on the battlefield. This trend is not limited to Russia; similar AI-enhanced systems are simultaneously being deployed on the Ukrainian side (Özel Özcan, 2025). Actually Russian forces have reportedly begun producing more than 5,000 modernised Geran-2 unmanned aerial vehicles per month at the Alabuga facility in Tatarstan illustrates both the scale and speed of this shift. More significantly, Russia is reported to have upgraded the Geran-2 UAV, based on the Iranian Shahed-136 design, with advanced onboard artificial intelligence (Autonomy Global, 2025). On the other hand, in August 2025, the State Council of China released the "AI Plus" strategy, which saw AI as more than just a technological tool. It was seen as a key tool for boosting the economy, improving social welfare, increasing productivity, and making the government stronger. By 2030, one of the main goals of the strategy is to have a lot of smart terminals and AI agents that are next-generation. This will make the new economy the main force behind national growth. It's also clear that China's "AI Plus" plan isn't just about making money at home. Instead, it wants to be involved in how power works between countries and how systems are shaped. This demonstrates that China's AI policy operates concurrently at both internal and external governance levels (State Council of China, 2025; Ministry of Foreign Affairs of the People's Republic of China, 2025).

At the end, Taken together, these developments compel us to reflect not only on the present but also on a highly consequential future. Competition among great powers is no longer shaped solely by the number of tanks, missiles, or soldiers, but increasingly by the speed of algorithms, decision-making capacity, and the operational effectiveness of autonomous systems on the battlefield.

This dynamic indicates that power distribution is not static; instead, it is continuously reproduced. However, this transformation is not limited to its visible dimensions. The

transition from a classically functioning world-system to a digital and algorithmic world-system now generates intense competition and dense interaction across almost all domains. This transition represents not merely a technical shift but a structural rupture that redefines how power is conceptualised, exercised, and constrained. Accordingly, the emerging picture points to a period in which states are evaluated not solely based on capacity expansion, but increasingly on their ability to adapt, govern, and integrate into new systemic domains.

Refences

Autonomy Global. (2025). What the other guys are doing: Russia mass producing AI-enabled Geran-2 drones. Autonomy Global. https://www.autonomyglobal.co/what-the-other-guys-are-doing-russia-mass-producing-ai-enabled-geran-2-drones/.

Mearsheimer, J. J. (2001). The tragedy of great power politics. New York, NY: W. W. Norton & Company.

Ministry of Foreign Affairs of the People's Republic of China. (2025). Global AI Governance Action Plan. Retrieved from https://www.fmprc.gov.cn/mfa_eng/xw/zyxw/202507/t20250729_11679232.html.

Özel Özcan, M. S. (2025). Recoding rules of war: AI reshapes balance of Russia-Ukraine war. Daily Sabah. https://www.dailysabah.com/opinion/op-ed/recoding-rules-of-war-ai-reshapes-balance-of-russia-ukraine-war.

State Council of China. (2025). Guideline to accelerate "AI Plus" integration across key sectors. Retrieved from https://english.www.gov.cn/policies/latestreleases/202508/27/content_WS68ae7976c6d0868f4e8f51a0.html.

Wallerstein, I. (1974). The modern world-system I: Capitalist agriculture and the origins of the European world-economy in the sixteenth century. New York, NY: Academic Press.

Wallerstein, I. (2004). World-systems analysis: An introduction. Durham, NC: Duke University Press.

ARTICLE AND BOOK REVIEWS / MAKALE VE KİTAP İNCELEMELERİ

HUMANS IN THE CYBER LOOP: PERSPECTIVES ON SOCIAL CYBERSECURITY

Selim Mürsel Yavuz[*]
ORCID ID: 0000-0002-0545-9123

Edited by Dorota Domalewska, Aleksandra Gasztold, and Agnieszka Wrońska, *Humans in the Cyber Loop: Perspectives on Social Cybersecurity* (Leiden | Boston: Brill, 2025), Studies in Critical Social Sciences 317. ISBN 978-90-04-54989-0 (hb), 978-90-04-54990-6 (e-book). DOI: 10.1163/9789004549906.

Declaration[*]

Humans in the Cyber Loop makes a direct claim that many cybersecurity discussions quietly sidestep. Security is not only about protecting machines and networks. It is also about people and the social settings that shape what they notice, trust, and do online. Domalewska, Gasztold, and Wrońska approach cybersecurity as a socio-technical problem tied to community dynamics and to the political economy of platforms. That framing is hard to dismiss once you sit with it, because so many recent "cyber" harms travel through attention and trust, not only through technical compromise.

The authors build the book around the idea that humans are "in the cyber loop" as both vulnerability and resilience. People create openings for harm through ordinary behavior, limited attention, and familiar cognitive shortcuts. Yet people can also become the source of resilience when learning is shared and institutions design for safer routines. I kept thinking of a scene that plays out in almost any organization. Someone is racing to clear an inbox before a meeting. A message arrives that looks like it came from IT, uses plausible language, and offers a quick link "to verify." The login page looks normal. In a hurry, they comply. The error is human, but the surrounding conditions were designed to make the click feel reasonable.

This leads to the book's central contribution, the development of "social cybersecurity" as an analytic lens. Rather than separating cybercrime, cyber conflict, and disinformation from

[*] Presidency for Turks Abroad and Related Communities, PhD Student in International Relations at Social Sciences University of Ankara, Selimmurselyavuz@gmail.com.
[*] The author acknowledges the use of Gemini for proofreading and language editing purposes during the preparation of this manuscript. The final content was reviewed and approved by the author.

platform governance, the authors treat them as connected. Algorithmic curation, surveillance capitalism, and influencer markets sit alongside more conventional security threats because they shape influence and coordination at scale. The book also keeps psychosocial consequences in view, including hate speech, cyber aggression, and problematic internet use.

Conceptually, the early chapters are among the strongest parts of the volume. Drawing on Beskow and Carley's definition, the authors describe social cybersecurity as concerned with cyber-mediated changes in human behavior and socio-political outcomes, while also building the conditions for societal endurance under social cyber threats. This clarifies what is being added to the more familiar cybersecurity agenda. Traditional cybersecurity tends to focus on compromise, systems, and technical controls. Social cybersecurity draws attention to manipulation, influence, and marginalization. It is also explicitly interdisciplinary, and the authors are right to treat that as a requirement rather than a slogan.

The book's structure is pedagogical and cumulative. Across nine chapters it moves from definitions to cyber threats, then to information warfare and disinformation, algorithmic influence, platform political economy, influencer ecosystems, content overload and hate, and finally problematic internet use, before concluding with a synthetic "digital ecosystem" chapter. Chapters 1 to 3 trace a familiar arc from crime to warfare to information operations while keeping the human pathway central. Chapter 2 expands the taxonomy toward cyber war and hybrid warfare, stressing that hybrid operations span infrastructure attacks, social fragmentation, and narrative shaping. Chapter 3 then synthesizes psychological and network mechanisms of disinformation and uses Russia as an illustrative case of tightened media control after 2022.

Chapter 4, "The Power of Algorithms," is the analytic pivot because it makes influence tangible. The authors acknowledge that ranking and recommendation systems can broaden exposure under certain conditions. Still, their emphasis falls on the costs when curation hardens bias, suppresses content, or distorts public reality. They ground the discussion in controversies that will be familiar to many readers, including allegations that TikTok suppressed LGBTQ-supportive content in some contexts. The unsettling point is not only bias in a narrow sense. It is how easily "visibility" becomes a political variable, adjusted quietly and at scale.

What keeps this chapter from becoming a simple critique is the authors' contrasting case. Estonia is presented as a governance model in which AI-enabled public services and the #KrattAI initiative are framed as public-good deployment, with human oversight positioned as a final layer of accountability. The point is not that Estonia has solved automation. It is that objectives and incentives matter. Systems built for engagement and monetization invite one set of outcomes. Systems built for public service, coupled with oversight that is at least visible to the public, invite a different set of outcomes.

Chapters 5 to 7 widen the focus to the socio-economic systems that scale social cyber threats. Chapter 5 treats digital marketing as a paradox. Personalization can feel empowering, yet it is often built on surveillance and behavioral steering. The authors draw on the language of surveillance capitalism and describe "hypernudges" as subtle ways platforms shape choices. Chapter 6 turns to influencer ecosystems and, in the discussion of "kidfluencers," highlights both the strain of persistent visibility and the weak protections around labor and privacy when content is produced in identifiable home settings. Chapter 7 connects content abundance to overload, hate speech, and cyber aggression, and it calls for multidisciplinary counter-hate strategies. The thread remains consistent. When incentives reward outrage and speed, the information environment becomes easier to weaponize.

Chapter 8 shifts to "digital dependency," and it changes the book's tone in a useful way. The authors outline symptoms such as loss of control, tolerance-like escalation, and withdrawal analogues. They stress that full abstinence is unrealistic, so the practical goal is balanced use framed as "conscious computing." They also summarize research on brain structure and function associated with addictive patterns and treat these findings as security-relevant insofar as they affect impulse control and decision-making. For security studies readers, this chapter expands what vulnerability can mean. It is not only weak passwords or unpatched systems. It can also be fatigue, compulsive checking, and attention fragmentation that make manipulation easier and self-control harder.

The concluding chapter consolidates an ecosystemic view through a socio-ecological approach that places the human being at the center of the digital ecosystem. Attention markets, micro-targeting, automated decision-making, and surveillance-based personalization are treated as forces shaping perceptions, relationships, and democratic stability. The authors also resist technological determinism. Manipulation and power-seeking predate Web 2.0.

Platforms amplify and accelerate those dynamics, which is precisely why governance and resilience still matter.

As an academic contribution, *Humans in the Cyber Loop* succeeds most as synthesis and orientation. Its chief strength lies in integrating multiple levels of analysis, from cognition and dependency to platform governance and hybrid conflict, supported by accessible cases and examples. For teaching and for interdisciplinary conversations, that coherence is valuable. The main limitation is the same breadth that makes the book readable and usable. Because the authors aim to provide representative examples rather than sustained case studies, the analysis sometimes reads as a well-organized tour. Readers looking for operationalization will find fewer concrete measures and research designs than the definition of social cybersecurity might suggest.

Two final points are worth noting. First, the book's normative stance is explicit and generally well defended, especially where it criticizes moderation practices that hide vulnerable users rather than confronting harassment. Still, the policy discussion could go further by treating regulatory tradeoffs more systematically, including transparency versus security, moderation versus speech, and privacy versus personalization. Second, the authors disclose using AI tools, including ChatGPT, Paperpal, Grammarly, and DeepL, followed by human proofreading. In a book about human agency inside algorithmic systems, that disclosure is quietly instructive.

Overall, *Humans in the Cyber Loop* is a timely entry point into social cybersecurity as both an interdisciplinary research area and a policy-relevant lens on contemporary digital threats. Its greatest utility is as a framework builder. It helps readers see that disinformation, algorithmic curation, surveillance economies, influencer-driven persuasion, and digital dependency are not separate problems. They interact within a single ecosystem, and that interaction is where security debates now need to live.

NOTES FOR AUTHORS / YAZARLAR İÇİN NOTLAR

We would like to thank you for choosing to submit your paper to *Cyberpolitik*. In order to fasten the process of reviewing and publishing please take try to read and follow these notes in depth, as doing so will ensure your work matches the journal's requirements.

All works including research articles, comments and book reviews submitted to *Cyberpolitik* need to be original contributions and should not be under consideration for any other journal before and/or at the same time.

All submissions are to be made online via the Journal's e-mail address: cyberpolitik@gmail.com

The authors of a paper should include their full names, affiliations, postal addresses, telephone numbers and email addresses on the cover page of the manuscript. The email address of the author will be displayed in the article.

Articles should be 1.5-spaced and with standard margins. All pages should be numbered consecutively. Please avoid breaking words at the end of lines.

The articles need to be between 5000 - 7000 words (including footnotes and references); comments between 2000-4000 words (including footnotes and references); and book - article reviews between 500 - 1500 words.

An abstract of up to 150 words should be added during the submission process, along with an average of five keywords.

Authors should make a final check of their article for content, style, proper names, quotations and references.

All images, pictures, maps, charts and graphs should be referred to as figures and numbered. Sources should be given in full for images, pictures, maps, tables and figures.

Comments in Cyberpolitik

A comment is a short evaluation of an expert regarding new issues and/or development in cyberpolitics.

Comments require journal's full reference style.

Book / article Reviews in Cyberpolitik

A book review should provide a fair but critical assessment of a recent (not older than 5 years) contribution to the scholarly literature on the themes and topics relevant to the journal.

A book review for Cyberpolitik:

- Provides complete bibliographical references of the book(s) and articles to be reviewed.
- Summarizes the content and purpose of the book, focusing on its main argument(s) and the theory, methodology and empirical evidence employed to make and support these arguments
- Critically assesses the author(s)' arguments, their persuasiveness and presentation, identifying the book's strengths and weaknesses
- Presents a concluding statement that summarizes the review and indicates who might benefit most from reading the book

Book / article reviews should be preceded by full publication information, in the following form:

Education for Peace: Politics of Adopting and Mainstreaming Peace Education Programs in Post-Conflict Settings by Vanessa Tinker, Academica Press, 2015, $81.62 (Hardcover), ISBN 978-1680530070.

The reviewer's name, affiliation and email address should appear, on separate lines, at the top of the review, right after the bibliography of the book/article.

Journal style

Authors are responsible for ensuring that their manuscripts conform to *cyberpolitik*'s reference style.

Reference style of *Cyberpolitik* is based on APA 6th Edition.