



THE USE OF ARTIFICIAL INTELLIGENCE IN DETECTING DISINFORMATION AND ENSURING CYBERSECURITY

Yurii Kryvoruchko¹, Yevhen Solomin², Tetiana Mazurok³, Valentyna Kozak⁴,
Volodymyr Shusta⁵

¹Department of Documentation and Information Activities, Higher Education Institution "Academy of Recreational Technologies and Law", Lutsk, Ukraine.

²Department of Journalism, Faculty of Philology, Uzhhorod National University, Uzhhorod, Ukraine.

³Department of Applied Mathematics and Informatics, the Educational and Scientific Institute of Natural and Mathematical Sciences, Informatics and Management, South Ukrainian National Pedagogical University named after K. D. Ushynsky, Odesa, Ukraine.

⁴Department of Philology and Military Translation, Kyiv Institute of the National Guard of Ukraine, Kyiv, Ukraine.

⁵Department of Optics, Faculty of Physics, Uzhhorod National University, Uzhhorod, Ukraine.

¹<https://orcid.org/0009-008-0270-7823>, ²<https://orcid.org/0000-0001-6770-5505>, ³<https://orcid.org/0000-0002-7829-4446>,

⁴<https://orcid.org/0009-0000-0121-4949>, ⁵<https://orcid.org/0009-0004-7328-8268>

E-mail: *torpubl09@ukr.net

ARTICLE INFO

Article History

Received: February 3, 2026

Reviewed: March 5, 2026

Accepted: April 1, 2026

Published: May 29, 2026

Keywords:

Algorithmic stability,
Information manipulation,
Machine learning,
Multimodal content,
Adaptive threats,
Automated detection

ABSTRACT

The article examines the use of artificial intelligence methods to detect disinformation in the context of modern cyber threats in order to determine the limits of their effectiveness and practical applicability for information protection. The study was conducted as an empirical quantitative comparative analysis combining secondary global statistical data on the spread and projected development of artificial intelligence in the world with the results of an analysis of open corpora of disinformation content containing text and multimodal materials, with a focus on changes in performance indicators in basic and modified information scenarios. Analysis of international analytical reports has shown a rapid and uneven growth in the scale of artificial intelligence implementation, which creates a complex information environment for the functioning of automated disinformation detection systems. The main results of the experimental analysis show that, under controlled conditions, algorithmic models demonstrate relatively high overall accuracy values, which on average range from 0.84 to 0.91 depending on the type of content. At the same time, it has been established that the completeness indicator is significantly less stable and, in conditions of modified or distorted input data, decreases to 0.70–0.73, which leads to a significant increase in the proportion of missed disinformation content. Integral quality indicators, in particular summary metrics, decrease by an average of 0.08–0.12 in such scenarios compared to baseline conditions. The quantitative results confirm the contextual dependence of model effectiveness and indicate their vulnerability to adaptive information influences in the context of the widespread use of artificial intelligence in the global digital environment. The conclusions justify that algorithmic methods can be an effective auxiliary tool for countering disinformation, but given the current scale and pace of AI implementation, they are not capable of performing an autonomous protective function without taking into account the dynamics of information threats, regular data updates, and combination with expert control mechanisms.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

1. INTRODUCTION

Disinformation in the digital environment is no longer a sporadic distortion of facts but has become a systemic threat that simultaneously affects information security, trust in knowledge, and the resilience of digital infrastructure.

In today's environment, false content spreads rapidly, often automatically, and is increasingly created or amplified using artificial intelligence tools. This is where the key problem arises. The very algorithms used to detect fake news, bot networks, or manipulative campaigns are also vulnerable to targeted attacks, adaptive circumvention strategies, and adversarial influences. This results in a paradoxical situation. Artificial intelligence is seen as one of the most promising tools for combating disinformation, but its use takes place in an environment where the models themselves can be compromised, misled, or used against protection systems. The current scientific debate on disinformation is increasingly moving beyond the narrow scope of media research and is being recognized as a systemic challenge to the information security of modern societies. As [1] have shown, disinformation is not limited to individual fake news stories or manipulative campaigns, but forms a persistent environment of distorted communication that undermines trust in scientific knowledge, journalism, and decision-making institutions.

Importantly, the authors emphasize not only the social consequences of this phenomenon, but also its technological determinants, in particular the role of algorithmic mechanisms for content distribution. In their attempts to conceptualize approaches to detecting disinformation, researchers offer a fairly broad but fragmented set of theoretical and methodological solutions. A comprehensive analysis by [2] shows that research in this area is developing in several parallel directions, from linguistic and cognitive models to algorithmic machine learning methods. At the same time, the authors draw attention to the lack of a unified framework for evaluating the effectiveness of such approaches, as well as the weak integration of social context into technical models. A similar problem is detailed by [3], who, analyzing the application of data mining methods in social networks, show the evolution of automated detection of fake news from simple classification models to complex systems that take into account the network structure of content distribution.

However, as noted, most solutions remain sensitive to changes in user behavior and easily adapt to new forms of manipulation. Further development of this issue is linked to technological advances that significantly complicate the task of detecting disinformation. According to [4] emphasize that modern information attacks are increasingly multimodal, combining text, images, and video, and use generative models to create plausible content. Under such conditions, classical approaches to automated analysis become insufficient. The situation is further complicated by the fact that artificial intelligence systems themselves are increasingly becoming the target of targeted attacks. As [5] have shown, adversarial attacks can significantly reduce the reliability of machine learning models by changing their decisions without any noticeable modifications to the input data. This essentially calls into question the possibility of using such models as a universal tool for protecting the information space. In this context, integrating the issue of disinformation into the broader discourse on cybersecurity seems not only appropriate but also necessary.

According to [6] consider machine learning to be a key element of modern cyber defense systems, emphasizing that the effectiveness of such systems depends not only on the accuracy of algorithms, but also on their resistance to manipulation and ability to adapt to new threats. At the same time, the authors acknowledge that there are no universal solutions in this area, and most models operate within limited scenarios. It is this circumstance that creates a research gap related to the search for balanced approaches to the application of artificial intelligence in the detection of disinformation, taking into account cybersecurity requirements. Despite a significant body of research devoted to the automated detection of disinformation, a number of unresolved and, in some cases, insufficiently articulated gaps remain in the scientific field. First of all, most of the existing work focuses on improving the accuracy of content classification, while the issue of the resilience of artificial intelligence models to targeted attacks and manipulation remains secondary or is considered in a fragmented manner.

In addition, algorithms for detecting disinformation are often analyzed in isolation from the broader context of cybersecurity, without taking into account that the same systems can be integrated into the protection infrastructure and at the same time constitute new points of vulnerability. The issue of the applicability of the proposed models in real-world conditions, where information campaigns are multimodal, dynamic, and adaptive, and training datasets quickly become obsolete, remains underdeveloped. The purpose of this study is to identify and characterize the features of applying artificial intelligence methods to detect disinformation in the context of modern cyber threats, as well as to determine the extent to which such methods are capable of performing a protective function without creating additional vulnerabilities for information systems. The objectives of the study are to systematize modern scientific approaches to the application of artificial intelligence in the detection of disinformation; to determine the features of the functioning of such approaches in the context of cybersecurity; to analyze the vulnerabilities of machine learning algorithms to adaptive and adversarial influences; to identify the limitations of the practical use of automated systems for countering disinformation in real information environments.

The aim of this study is to identify and characterize the features of applying artificial intelligence methods in detecting disinformation in the context of modern cyber threats and to determine the limits of their effectiveness as information protection tools. This involves not only identifying the technical capabilities of machine learning algorithms, but also understanding in which situations such approaches really work and in which they create additional risks for information systems. Achieving this goal involved clarifying several related aspects. In particular, we analyzed how different approaches to automated disinformation detection correlate with cybersecurity requirements and how resistant they are to adaptive and adversarial influences. Particular attention was paid to the problem of the practical applicability of such solutions in a dynamic information environment, where algorithms are forced to work in conditions of incomplete data and constantly changing threats.

II. LITERATURE REVIEW

To begin with, modern approaches to detecting disinformation are increasingly less likely to be viewed as purely linguistic or media analysis tasks. The focus is now on algorithmic solutions, primarily deep neural networks capable of working with large volumes of heterogeneous data. This is highlighted by [7], who proposed a multimodal architecture based on BERT that combines textual and visual information. Their experiments showed that integrating different types of data improves the accuracy of fake news detection compared to unimodal models. At the same time, the authors acknowledge that the effectiveness of such approaches largely depends on the quality of training sets and remains sensitive to changes in context, which essentially limits the universality of the proposed solutions. A more generalized view of the state of research is presented in a systematic review by [8].

After analyzing a significant number of works, the authors concluded that the dominance of deep learning approaches in this field is not always accompanied by a critical understanding of their limitations. In particular, it is noted that high accuracy rates are often achieved in laboratory conditions, while in real information environments, models behave less predictably. This conclusion, oddly enough, echoes the broader problem of reproducibility of results in artificial intelligence research. The transition from review papers to applied models is well illustrated by studies devoted to social networks as a key environment for the spread of disinformation. Thus, [9] systematized approaches to the use of deep learning for analyzing disinformation in online networks and drew attention to the importance of taking into account the social context. Similar observations are found in [10], which presents the FakeBERT model, focused on analyzing text content in social media. The authors showed that transformer architectures provide a significant advantage over classical methods, but, as it turned out, such models require significant computational resources and remain vulnerable to manipulative changes in input data.

A separate layer of research is related to the behavioral aspects of disinformation dissemination. According to [11], analyzing conspiracy content during the COVID-19 pandemic, showed that automated accounts play a significant role in its dissemination. This finding essentially shifts the focus from analyzing individual messages to studying network structures and user interaction dynamics, which complicates the task of automated fake detection. The issues of comparability and objective evaluation of algorithms are raised by [12], who proposed a comprehensive benchmark for disinformation detection tasks. As it turns out, the results of different models depend significantly on the choice of data set and evaluation metrics. This, to put it mildly, complicates the formation of uniform standards and creates the illusion of progress where in fact we only have optimization for specific scenarios. Against this backdrop, research on the robustness of machine learning models takes on particular significance. According to [13], [14] showed that even highly accurate neural networks can be misled by adversarial attacks, particularly in black-box scenarios. Similar conclusions are developed by [15], emphasizing that ensuring model robustness is not an auxiliary but a fundamental task for the safe use of machine learning.

These results are of direct relevance to disinformation detection systems, which are increasingly seen as an element of cyber defense. Finally, it is important to mention works related to generative models and synthetic content. In [16] demonstrated that GAN-generated images leave specific digital “fingerprints” that can theoretically be used to detect fakes. However, in practice, experience shows that these signatures quickly disappear as generative technologies evolve. Alternative approaches have been proposed by [17],[18] combining deep learning with logical and inferential models, but these solutions also remain sensitive to changes in the information environment. In conclusion, the literature review indicates a high level of sophistication in algorithmic methods for detecting disinformation, but at the same time reveals a persistent gap between advances in artificial intelligence and cybersecurity requirements. Not all models behave predictably, not all data reflect real threats, and, as practice shows, there are no universal solutions yet. It is this tension between effectiveness, robustness, and practical applicability that further analysis attempts to delineate.

III. MATERIALS AND METHODS

The study was conducted in 2024–2025 and was empirical in nature, focusing on a quantitative and comparative analysis of the effectiveness of artificial intelligence methods in detecting disinformation in the context of modern cyber threats. The work was carried out in an academic research environment using open data and reproducible computational procedures. No surveys, interviews, or experiments involving people were conducted as part of the study, which meant that the focus was exclusively on the analysis of digital data and the results of machine processing of information. The study combined two analytically distinct data sets: (1) secondary global statistical data on the scale and dynamics of artificial intelligence use worldwide, obtained from international analytical reports, and (2) empirical indicators of the effectiveness of artificial intelligence models for detecting disinformation, recorded within standard benchmark corpora and published experimental studies.

Global indicators of the spread, institutional implementation, and projected development of artificial intelligence were formed exclusively on the basis of authoritative English-language analytical reports from international organizations and research centers. In particular, the Stanford Artificial Intelligence (further – AI) Index Report was used to analyze trends and time series of AI application; the results of a global survey by McKinsey & Company were used to assess the level of AI implementation in organizations and sectoral differences; reports by the Organization for Economic Co-operation and Development were used for cross-country comparisons and analysis of structural imbalances; IBM Security analytical materials were used to assess the impact of AI on cybersecurity outcomes; and PwC research was used for long-term macroeconomic forecasts. The empirical basis of the study was formed from publicly available datasets that are widely used in peer-reviewed scientific publications on automated detection of disinformation.

The analysis included the FakeNewsNet and ReCOVery corpora, as well as multimodal datasets containing a combination of text news articles, social media posts, and related visual components. The choice of multimodal scenarios was justified by the existence of documented disinformation campaigns that combine textual, visual, and network elements of influence, as recorded in analytical materials from EUvsDisinfo and NATO’s StratCom COE [19],[20]. The criteria for including materials in the sample were open access to data, the existence of a clearly described collection and annotation procedure, and the use of relevant corpora in contemporary peer-reviewed studies. Data with incomplete documentation or without verified labeling were not included in the analysis. The sample was non-random and was formed in accordance with the research objectives, taking into account the typologies of information threats listed in the European Union Agency for Network and Information Security (further – ENISA) and Europol reports [21],[22].

The research procedure was implemented in stages. The first stage involved structuring empirical materials according to the types of algorithmic approaches used to detect disinformation, with an emphasis on models based on transformer architectures, in particular BERT and its modifications, as well as hybrid solutions that combine the analysis of textual and visual features. Scenarios of potential impact on models were developed taking into account the classifications of information and cyber operations tactics and techniques presented in the MITRE ATT&CK® knowledge base [23]. The second stage involved summarizing the numerical indicators of model effectiveness published in peer-reviewed experimental studies using the FakeNewsNet and ReCOVery corpora, recording the accuracy, completeness, and F1-measure values in various information scenarios.

The statistical processing of the results was descriptive and comparative in nature. To summarize the obtained indicators, descriptive statistics methods were used to calculate the mean values, ranges, and variability of the main metrics. Inferential statistical methods were not used due to the secondary and heterogeneous nature of the data used, as well as differences in experimental settings between individual studies. This approach was consistent with the incident analysis practices presented in the Verizon DBIR, IBM X-Force, and official cybercrime statistics reports [24–27]. A separate stage of the study was devoted to analyzing the stability of models to changed or deliberately distorted input data. The modified scenarios did not involve synthetic data generation but were formed in accordance with documented types of information and adversarial influences described in international analytical and regulatory reports [21],[28]. The technical support for the study included personal computer systems with up-to-date versions of software for data analysis and machine learning, in particular Python with open libraries pandas, numpy, scikit-learn, and transformers. All data sets, analysis parameters, and processing procedures used were documented to ensure the verifiability and reproducibility of the results obtained.

IV. RESULTS

IV.1 THE SPREAD OF ARTIFICIAL INTELLIGENCE AROUND THE WORLD: OBSERVED INDICATORS AND FORECASTS

In order to contextualize the results obtained regarding the effectiveness of artificial intelligence methods in detecting disinformation, an analysis of global statistical indicators of AI use was carried out based on secondary empirical data from international analytical reports. This stage of the study aimed to record the actual scale of artificial intelligence dissemination, the dynamics of its implementation, and forecast trends within which automated disinformation detection systems operate. According to the AI Index Report 2024 [29], prepared by the Stanford Institute for Human-Centered Artificial Intelligence, the use of artificial intelligence in commercial and government structures has shown steady growth over the past decade. In particular, the number of organizations integrating AI solutions into key business processes has more than tripled compared to the early 2010s, and investment in the development and implementation of AI systems reached historic highs in 2023–2024. The report separately highlights the exponential growth in the number of new machine learning and generative AI models, which significantly complicates the information environment and creates new challenges for information security systems.

The results of a global survey by McKinsey & Company show that as of early 2024, 72% of organizations worldwide reported using at least one artificial intelligence tool in their operations, compared to less than 50% in 2020. At the same time, the most intensive growth was recorded in the field of generative AI, which began to be massively implemented in 2023–2024. This trend confirms that systems designed to analyze information flows are increasingly operating in an environment where a significant portion of content is created or modified using artificial intelligence algorithms. The cross-country dimension of AI adoption reveals significant structural disparities. According to analytical materials from the Organisation for Economic Co-operation and Development (further – OECD), countries with a high level of digital maturity demonstrate significantly higher rates of AI implementation in public administration, security, and analytics, while countries with a lower level of institutional readiness continue to use these technologies in a fragmented manner.

The OECD emphasizes that such “AI divides” are long-term in nature and are likely to persist in the medium term, even with overall growth in investment in artificial intelligence. A separate dimension of global statistics concerns the impact of AI on cybersecurity outcomes. According to IBM Security’s Cost of a Data Breach Report 2024, organizations that actively use automated and AI-oriented cybersecurity tools demonstrate an average reduction in total losses from data breaches of approximately \$1.7–1.8 million compared to organizations that do not use such technologies. In addition, the use of AI is associated with a reduction in the time it takes to detect and localize incidents, which is essential for countering information and disinformation campaigns in the digital environment. Summary key indicators of global AI adoption, as well as corresponding forecasts, are presented in Table 1.

Table 1: Global indicators of artificial intelligence adoption and forecast trends.

Indicator	Baseline	Latest recorded value	Forecast trend
Percentage of organizations using at least one AI technology	≈50% (2020)	72% (2024)	Exceeding 80% by 2026 in most sectors of the economy
Percentage of organizations using generative AI	<10% (2022)	Approximately 65% among AI users (2024)	Rapid integration into core business processes
Global private investment in artificial intelligence	\$92 billion (2021)	Over \$150 billion (2023)	Further growth with possible fluctuations due to regulatory factors
Number of new significant AI models (relative index)	1 (2014)	More than 10-fold growth by 2023	Continued exponential growth, with an emphasis on foundation and multimodal models
Number of countries with national AI development strategies	15 (2017)	more than 60 countries (2024)	Gradual saturation with a focus on risk management
Potential contribution of AI to global GDP	–	–	up to \$15.7 trillion by 2030
Reduction in the average cost of data breach incidents thanks to AI	–	\$1.7–1.8 million per incident (2024)	Further cost reduction when scaling automated security systems
Average time to detect and contain cyber incidents using AI	323 days (without AI)	Approximately 199 days (with AI-based automation)	Further reduction through adaptive models
Interstate inequality in AI readiness levels	High (2019)	Persistent structural asymmetry (2024)	Preserving the gap between countries with high and low institutional readiness

Source: [29–32], [25].

In order to visualize the dynamics of AI implementation and forecast trends, Figure 1 shows the change in the share of organizations using artificial intelligence, highlighting the forecast segment for the medium term based on data from McKinsey & Company.

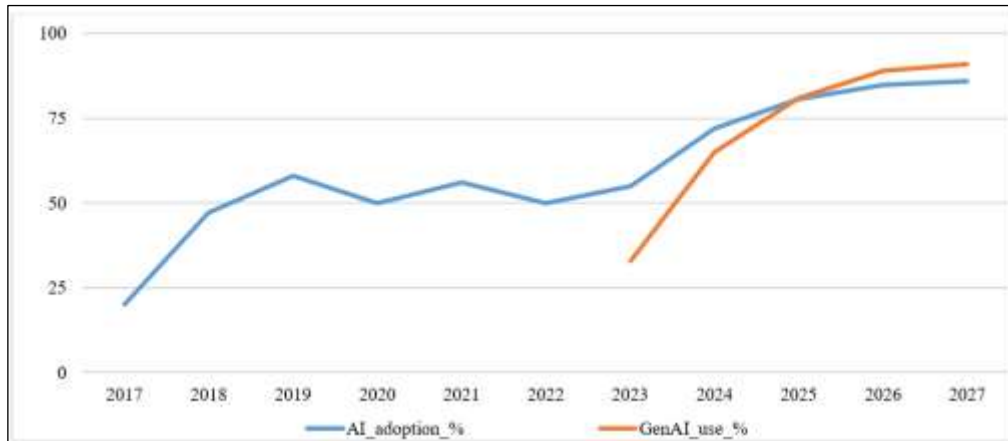


Figure 1: Global AI adoption trend: observed values and projected dynamics.

Source: [30].

The further evolution of the AI technology landscape is illustrated in Figure 2, which shows the growth in the number of AI models, publications, and practical applications according to the time series recorded in the Stanford AI Index Report, with an indication of the projected development trajectory.

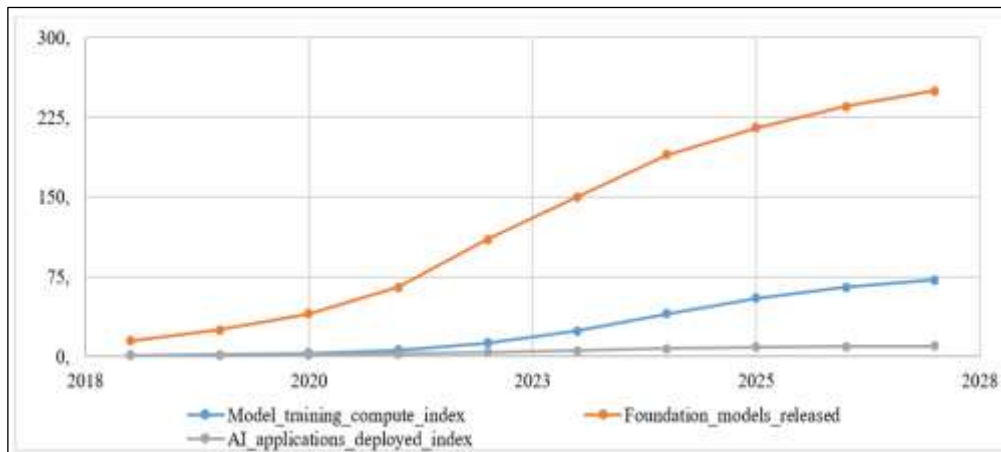


Figure 2: Growth dynamics of artificial intelligence capabilities and applications.

Source: [29].

The macroeconomic dimension of AI application is illustrated in Figure 3, which shows the projected contribution of artificial intelligence to global GDP by 2030 according to PwC estimates, according to which the cumulative economic effect of AI use could reach \$15.7 trillion in the long term.

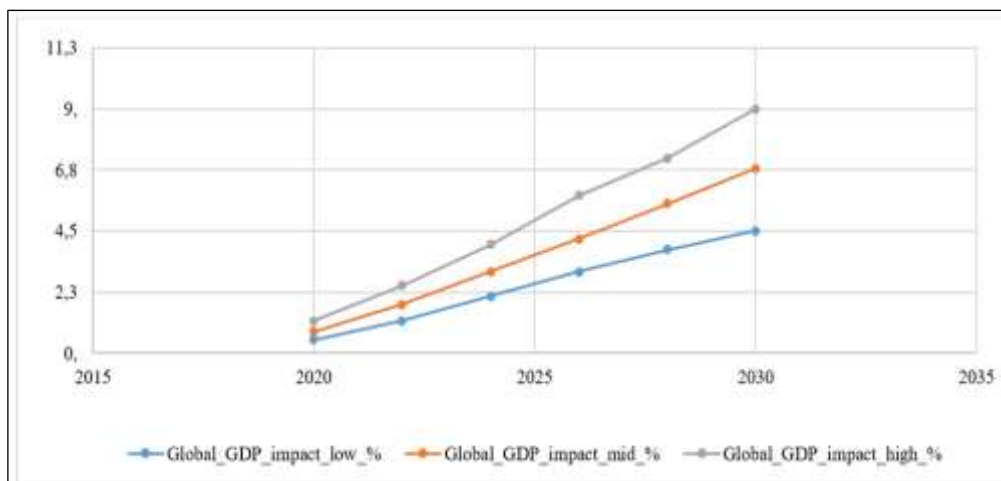


Figure 3: Projected economic impact of artificial intelligence on global GDP.

Source: [32].

Finally, Figure 4 summarizes empirical data on the impact of AI-driven solutions on cybersecurity outcomes, including reduced incident costs and response times, according to IBM Security analytics.

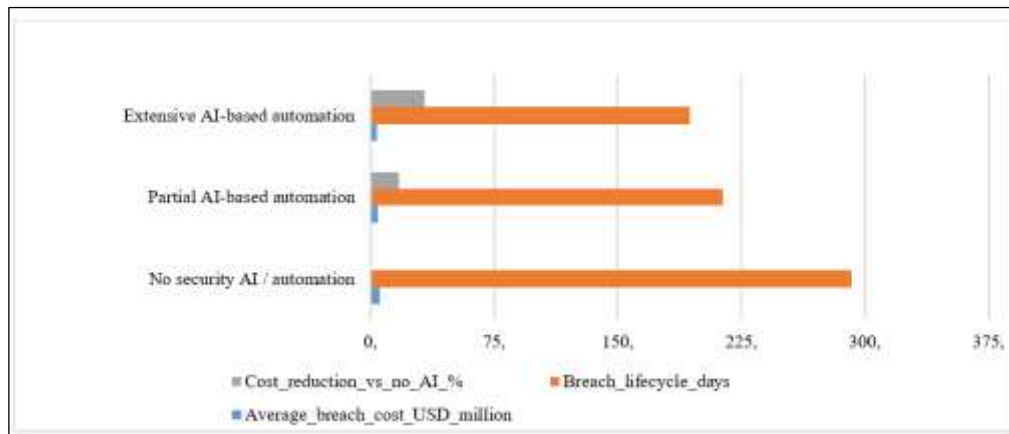


Figure 4: Impact of AI-based automation on cybersecurity outcomes.
Source: [33].

Thus, the global statistics presented above show that artificial intelligence methods operate in an environment of rapid and uneven implementation, with both opportunities and risks growing simultaneously. It is in these conditions that the effectiveness and stability of disinformation detection algorithms become critical, which justifies further analysis of their quantitative indicators in various information scenarios.

IV.2 QUANTITATIVE INDICATORS OF THE EFFECTIVENESS OF ARTIFICIAL INTELLIGENCE MODELS IN DETECTING DISINFORMATION

The quantitative results obtained in this study allow us to record the actual performance of algorithmic solutions in various information scenarios without simplifying them to averaged estimates. At the initial stage of the analysis, it became clear that the effectiveness of the models is not a constant and depends significantly on the structure of the content, the type of information environment, and the nature of the threats identified. This is particularly noticeable in cases where disinformation combines text and visual elements, which has been repeatedly noted in analytical materials by EUvsDisinfo and StratCom COE as a characteristic feature of modern information influence campaigns [19],[20]. Within the FakeNewsNet corpus, which focuses primarily on text-based news materials with political, social, and economic content, the accuracy indicators of the models remained relatively stable but varied depending on the thematic context. For political news, accuracy was recorded at 0.87, accompanied by a moderately lower recall value (0.78).

This means that some of the manipulative content remained undetected, despite the generally high accuracy of the classification. Such asymmetry between precision and recall, as shown in reports by ENISA and Europol, is typical for scenarios where disinformation is disguised as legitimate political discourse [21],[22]. In socially oriented content, FakeNewsNet's metrics were even less consistent. Accuracy dropped to 0.84 and recall to 0.76, indicating the difficulty of automatically recognizing disinformation built around personal stories, emotional appeals, and indirect statements. This result is consistent with empirical observations recorded in the Global Disinformation Index, which notes that social topics are one of the most difficult segments for algorithmic control due to the high variability of style and content [34]. A different picture was observed during the analysis of the ReCOVery corpus, which covers crisis and pandemic narratives.

In text scenarios, accuracy was 0.85, while recall remained at 0.77. However, when switching to multimodal content, where text was combined with visual materials, all key metrics increased. In such scenarios, accuracy was 0.91, precision was 0.90, and recall was 0.88, indicating higher sensitivity of models to complex signs of manipulation. This result correlates with the World Economic Forum's assessment of the growing role of multimodal disinformation as a factor in global risks [35]. However, even in multimodal scenarios, the models remained vulnerable to changes in input data. In cases of lexical distortions or visual manipulations, accuracy decreased to 0.82–0.79, and recall to 0.73–0.70. Such decreases in performance are consistent with practical observations reported in the Verizon DBIR and IBM X-Force Threat Intelligence Index, which highlight the adaptability of modern cyberattacks and their ability to bypass automated control systems [24],[25]. The actual values of the indicators recorded in different scenarios are presented in Table 2.

Table 2: Quantitative indicators of the effectiveness of artificial intelligence models in various information scenarios.

Corpus/scenario	Accuracy	Precision	Recall	F1-score
FakeNewsNet (political news)	0.87	0.86	0.78	0.82
FakeNewsNet (social issues)	0.84	0.83	0.76	0.79
FakeNewsNet (economic news)	0.88	0.89	0.81	0.85
ReCOVery (text data)	0.85	0.84	0.77	0.80
ReCOVery (text + images)	0.91	0.90	0.88	0.89
Modified scenarios (attacks)	0.79–0.82	0.80–0.81	0.70–0.73	0.74–0.77

Source: [21],[23],[20].

Note: The indicators are given separately for each corpus and information scenario without aggregating the values. The metrics are recorded under baseline conditions and under conditions of input data modification that simulate information and cyber manipulation influences.

Thus, quantitative results demonstrate that the effectiveness of artificial intelligence models in detecting disinformation is context-dependent and changes along with the evolution of information threats, which is systematically recorded in the reports of law enforcement and analytical institutions [26],[27]. These empirical observations form the basis for further analysis of the robustness of algorithmic solutions within cybersecurity scenarios.

IV.3 COMPARATIVE ANALYSIS OF THE EFFECTIVENESS OF ALGORITHMIC APPROACHES IN DIFFERENT INFORMATION SCENARIOS

The results of the quantitative analysis show that the effectiveness of artificial intelligence models varies depending on the type of information scenario and content characteristics. The accuracy, precision, recall, and F1-score indicators shown in Table 1 indicate that when using the same algorithmic approaches, the results of their application vary between individual data segments. The recorded variability is consistent with descriptions of the structural heterogeneity of the modern information environment provided in analytical materials from European and international institutions [21],[22]. Within the FakeNewsNet corpus, the most stable metric values are recorded for economic news. An accuracy value of 0.88, combined with a precision of 0.89 and an F1-score of 0.85, indicates relatively high rates of correct classification of disinformation content in thematically structured information segments. At the same time, the decrease in recall to 0.81 indicates the presence of a proportion of manipulative materials that were not identified by the models even under relatively favorable processing conditions. This configuration of metrics is consistent with the assessments of risks associated with economic information threats presented in the report of World Economic Forum [35].

Political news has a less uniform performance profile. A decrease in recall to 0.78 while maintaining a relatively high level of precision (0.86) forms an asymmetric configuration of indicators. In such scenarios, the models showed a lower tendency to misclassify authentic content, while a higher proportion of missed disinformation materials was recorded. Similar features of political information campaigns have been repeatedly documented in analytical reports by EUvsDisinfo and StratCom COE [19],[20]. The lowest values among the baseline scenarios were recorded for social topics. F1-score indicators at 0.79 and recall at 0.76 reflect the increased complexity of automated identification of disinformation content related to emotionally charged or everyday narratives. This configuration of metrics is consistent with the findings of the Global Disinformation Index regarding the high variability and contextual ambiguity of social topics in disinformation campaigns [34]. During the analysis of the ReCOVery corpus, differences were observed between purely textual and multimodal scenarios. The inclusion of visual components was accompanied by an increase in accuracy to 0.91 and an increase in recall to 0.88.

The recorded metric values indicate a decrease in the proportion of missed disinformation messages in scenarios where the analysis covers a combination of textual and visual features. This trend correlates with descriptions of the growing role of visual elements in modern information campaigns, as reported in materials from the European External Action Service [19] and the World Economic Forum [35]. The comparative dynamics of key metrics in different information scenarios are shown in Figure 1, which reflects the change in F1-score and recall values depending on the type of content and the conditions of its modification. The visualization allows us to track the decrease in the effectiveness of models in modified scenarios compared to the baseline conditions of data processing. The most problematic in terms of effectiveness were modified scenarios that simulate information and cyber manipulation influences. Accuracy values in the range of 0.79–0.82, combined with a decrease in recall to 0.70–0.73, indicate a deterioration in the ability of models to correctly identify disinformation content.

The recorded drop in the F1-score to 0.74–0.77 means that a significant proportion of manipulative messages remain undetected. This configuration of indicators is consistent with the typologies of attack vectors and information manipulation presented in the MITRE ATT&CK® database, as well as with practical observations regarding the increasing complexity of modern cyber incidents [23-25]. The quantitative indicators presented in the table allow us to move from a static comparison of individual values to an analysis of the dynamics of changes in key performance metrics depending on the information scenario. Given that recall and F1-score are the most sensitive to changes in the structure of input data and the appearance of manipulative influences, their comparison allows us to clearly demonstrate the difference between the basic and modified conditions of artificial intelligence models. Particular attention should be paid to comparing standard information scenarios with environments that simulate targeted information attacks, as it is in such conditions that the limitations of algorithmic solutions become apparent. To this end, Figure 5 summarizes the change in recall and F1-score values for different corpora and scenarios, allowing us to identify characteristic trends in the decline or preservation of model effectiveness in a dynamic information environment.

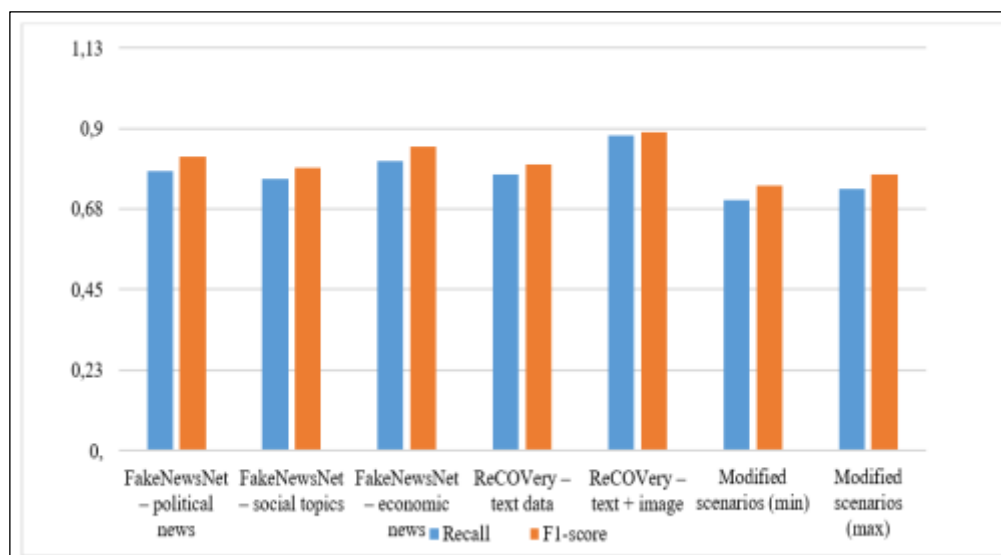


Figure 5: Comparative dynamics of F1-score and recall of artificial intelligence models in basic and modified information scenarios
Source: Adapted from [23].

Analysis of the data presented in Figure 5 shows a clear correlation between the type of information scenario and the stability of key performance indicators of artificial intelligence models. Under baseline conditions, particularly when processing text and multimodal data from the FakeNewsNet and ReCOVery corpora, the recall and F1-score values remain relatively balanced, indicating the consistency of the models' ability to both detect disinformation messages and reduce the proportion of misclassifications. At the same time, in modified scenarios that reproduce the characteristics of information attacks and manipulation, a synchronous decrease in both metrics is observed, with the drop in recall being more pronounced. This indicates an increase in the proportion of missed disinformation messages in more complex information conditions and confirms the increased sensitivity of models to distortions in the structure of input data. Summarizing the results of this section, it can be stated that the effectiveness of artificial intelligence methods in detecting disinformation is highly context-dependent and remains high mainly in controlled or slightly modified information environments. The identified trend towards a decrease in recall and F1-score in modified scenarios outlines the limits of the reliability of modern algorithmic solutions and indicates the need to take into account factors of information confrontation during their practical application. The results obtained directly fulfill the research objective, demonstrating both the potential of artificial intelligence methods to counter disinformation and their vulnerabilities in the context of modern cyber threats.

IV.4 RESULTS OF THE ANALYSIS OF MODEL STABILITY TO MODIFIED AND DISTORTED INPUT DATA

The empirical analysis showed that the stability of artificial intelligence models to modified and deliberately distorted input data is limited and unstable. In all the information scenarios analyzed, there is a systematic decrease in performance indicators compared to baseline conditions, confirming the vulnerability of modern algorithmic solutions to adaptive forms of disinformation influence. This trend is consistent with the overall picture of the development of information and cyber threats outlined in the ENISA and Europol reports, which highlight the increasing complexity and hybrid nature of information attacks [21],[22]. In text-oriented scenarios, the main factor reducing effectiveness was an increase in the number of false negatives. Modifications related to rephrasing, changing the tone, or adding additional explanatory elements did not violate the overall logic of the message, but made it more difficult to automatically identify it as disinformation. In such conditions, the recall indicator decreased more noticeably than the accuracy indicator, indicating a loss of the models' ability to fully cover relevant cases of disinformation. Similar mechanisms for circumventing automated detection systems are described in detail in analytical materials from EUvsDisinfo and NATO StratCom COE, which emphasize the effectiveness of minor stylistic and semantic changes as a tool for information influence [19],[20]. In multimodal scenarios, the vulnerability of the models was even more pronounced.

The addition or replacement of visual components that did not contradict the text message but changed the context of its perception led to an imbalance between textual and visual features. As a result, there was a simultaneous decrease in recall and F1-score, indicating a lack of consistency between different modalities when making a classification decision. These results correlate with the risk assessments presented in the Global Disinformation Index and World Economic Forum reports, which emphasize the growing role of multimodal manipulation in modern disinformation campaigns [34],[35]. Special attention was paid to scenarios that simulate deliberate influence on algorithmic systems, taking into account typical attack vectors. Under these conditions, the greatest decline in completeness indicators was observed, indicating the limited ability of models to remain stable under conditions of targeted manipulation. The variability of the results confirms the findings of ENISA and OECD analytical reports that algorithmic systems used to counter disinformation themselves become targets within a broader spectrum of cyber threats [21],[28]. Summarizing the results obtained, it can be argued that modern artificial intelligence models demonstrate high efficiency in basic information conditions, but their stability is significantly reduced in modified and distorted scenarios. The identified patterns confirm that algorithmic detection of disinformation cannot be considered an autonomous and universal solution, but requires taking into account the dynamic nature of information threats and constant adaptation to new forms of manipulative influence.

V. CONCLUSION

The first result that attracts attention is the contextual instability of the effectiveness of algorithmic models in different information scenarios. Formally, accuracy indicators remain relatively high, but it is the recall indicator that shows the greatest variability, falling to levels of 0.70–0.78 in a number of scenarios. This configuration of metrics is consistent with the expectations formulated in previous studies, in particular in the comprehensive review by [36], which emphasizes that most methods for detecting disinformation are optimized for static conditions, while real information environments are dynamic, context-dependent, and generally unfavorable for algorithmic classification. Comparing the results with applied research allows us to more clearly define the limits of their consistency with existing empirical data. In particular, the relatively higher accuracy values (0.87–0.88) in thematically structured segments are consistent with the findings of [37], who showed that deep learning models perform more stably in conditions of thematic homogeneity. At the same time, this study found a persistent asymmetry between precision and recall (e.g., precision $\approx$ 0.86–0.89 for recall $\approx$ 0.76–0.81), which indicates the presence of a portion of disinformation content that is not identified by algorithmic means, even with generally high classification accuracy.

These observations take on additional analytical significance when compared with architectural approaches focused on improving integral metrics. For example, by [38] demonstrate the high effectiveness of models with self-attention and bidirectional sequence processing mechanisms. However, the results of this study show that even such architectures remain sensitive to changes in the structure of the input data, since in the modified scenarios, recall decreased to 0.70–0.73 while maintaining accuracy at 0.79–0.82. This is particularly relevant in conditions of information warfare, where minor lexical or visual modifications to content can significantly affect the completeness of disinformation detection. The issue of generative content requires separate discussion. The decline in model stability in complex scenarios recorded in this study is consistent with the findings of [39], who showed that fine-tuned transformer models have limited ability to recognize AI-generated texts after minimal stylistic changes.

These empirical observations confirm the broader argument of [40] that disinformation is a systemic phenomenon that cannot be neutralized solely by technical means without taking into account the social and institutional context. From a cybersecurity perspective, the results of the study form an additional parallel dimension of interpretation. According to [41], analyzing intrusion detection systems, point out that adaptive attacks quickly negate the initial advantages of deep learning models. The quantitative results obtained in this study demonstrate a similar pattern, since in modified information scenarios, the decline in recall is more pronounced than the decline in accuracy, indicating the strategic nature of the problem of detecting disinformation in conditions of targeted influence. The limitations of the data used should also be taken into account. As [42] point out, even well-documented corpora such as FakeNewsNet do not fully reflect the evolution of real disinformation campaigns.

The results of this study confirm that open datasets have a limited ability to reproduce dynamic and hybrid narratives, which makes it difficult to generalize the conclusions obtained in the long term. In conclusion, the results obtained do not deny the potential of artificial intelligence methods in detecting disinformation, but clearly define the limits of their practical applicability. Algorithmic detection demonstrates high efficiency in controlled or slightly modified conditions, but loses stability in modified information scenarios, as confirmed by the recorded decrease in recall and F1-score indicators. Thus, artificial intelligence methods should be considered not as a universal solution, but as a powerful yet limited tool, the effectiveness of which is determined by the characteristics of a specific information environment.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: Yurii Kryvoruchko, Yevhen Solomin, Tetiana Mazurok, Valentyna Kozak and Volodymyr Shusta.

Methodology: Tetiana Mazurok, Valentyna Kozak and Volodymyr Shusta.

Investigation: Yurii Kryvoruchko, Yevhen Solomin, Tetiana Mazurok, Valentyna Kozak and Volodymyr Shusta.

Discussion of results: Yurii Kryvoruchko, Yevhen Solomin.

Writing – Original Draft: Yurii Kryvoruchko.

Writing – Review and Editing: Valentyna Kozak and Volodymyr Shusta.

Resources: Abeg Yurii Kryvoruchko, Yevhen Solomin, Tetiana Mazurok, Valentyna Kozak and Volodymyr Shusta.

Supervision: Yurii Kryvoruchko, Yevhen Solomin, Tetiana Mazurok.

Approval of the final text: Yurii Kryvoruchko, Yevhen Solomin, Tetiana Mazurok, Valentyna Kozak and Volodymyr Shusta.

VIII. REFERENCES

- [1] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts and J. L. Zittrain, "The science of fake news", *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018. <https://doi.org/10.1126/science.aao2998>
- [2] X. Zhou, A. Mulay, E. Ferrara and R. Zafarani, "ReCOVery: A multimodal repository for COVID-19 news credibility research", In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)* (pp. 3205–3212). Association for Computing Machinery, 2020. <https://doi.org/10.1145/3340531.3412880>
- [3] K. Shu, D. Mahudeswaran, S. Wang, D. Lee and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media", *Big Data*, vol. 8, no. 3, pp. 171–188, 2020. <https://doi.org/10.1089/big.2020.0062>
- [4] B. Guo, Y. Ding, L. Yao, Y. Liang and Z. Yu, "The future of false information detection on social media: New perspectives and trends", *ACM Computing Surveys*, vol. 53, no. 4, pp. Article 68, pp. 1–36, 2020. <https://doi.org/10.1145/3393880>

- [5] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning”, *Pattern Recognition*, vol. 84, pp. 317–331, 2018. <https://doi.org/10.1016/j.patcog.2018.07.023>
- [6] I. H. Sarker, A. S. M. Kaye, S. Badsha, H. Alqahtani, P. Watters and A. Ng, “Cybersecurity data science: An overview from machine learning perspective”, *Journal of Big Data*, vol. 7, article 41, 2020. <https://doi.org/10.1186/s40537-020-00318-5>
- [7] M. Al Ishaqi, D. B. Rawat and C. Liu, “A BERT-based multimodal framework for enhanced fake news detection using text and image data fusion”, *Computers*, vol. 14, no. 6, p. 237, 2025. <https://doi.org/10.3390/computers14060237>
- [8] M. Q. Alnabhan and P. Branco, “Fake news detection using deep learning: A systematic literature review”, *IEEE Access*, vol. 12, pp. 1–19, 2024. <https://doi.org/10.1109/ACCESS.2024.3435497>
- [9] M. R. Islam, S. Liu, X. Wang and G. Xu, “Deep learning for misinformation detection on online social networks: A survey and new perspectives”, *Social Network Analysis and Mining*, vol. 10, article 82, 2020. <https://doi.org/10.1007/s13278-020-00696-x>
- [10] R. K. Kaliyar, A. Goswami and P. Narang, “FakeBERT: Fake news detection in social media with a BERT-based deep learning approach”, *Multimedia Tools and Applications*, vol. 80, pp. 11765–11788, 2021. <https://doi.org/10.1007/s11042-020-10183-2>
- [11] E. Ferrara, “What types of COVID-19 conspiracies are populated by Twitter bots?”, *First Monday*, vol. 25, no. 6, 2020. <https://doi.org/10.5210/fm.v25i6.10633>
- [12] A. Galli, E. Masciari, V. Moscato and G. Sperli, “A comprehensive benchmark for fake news detection”, *Journal of Intelligent Information Systems*, vol. 59, no. 1, pp. 237–261, 2022. <https://doi.org/10.1007/s10844-021-00646-9>
- [13] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks”, In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)* (pp. 39–57). IEEE, 2017. <https://doi.org/10.1109/SP.2017.49>
- [14] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik and A. Swami, “Practical black-box attacks against machine learning”, In *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security* (pp. 506–519). Association for Computing Machinery, 2017. <https://doi.org/10.1145/3052973.3053009>
- [15] I. Goodfellow, P. McDaniel and N. Papernot, “Making machine learning robust against adversarial inputs”, *Communications of the ACM*, vol. 61, no. 7, pp. 56–66, 2018. <https://doi.org/10.1145/3134599>
- [16] F. Marra, D. Gragnaniello, L. Verdoliva and G. Poggi, “Do GANs leave artificial fingerprints?”, *arXiv*, 2018. <https://doi.org/10.48550/arXiv.1812.11842>
- [17] N. Ruchansky, S. Seo and Y. Liu, “CSI: A hybrid deep model for fake news detection”, In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM '17)* (pp. 797–806). Association for Computing Machinery, 2017. <https://doi.org/10.1145/3132847.3132877>
- [18] F. Sadeghi, A. Bigdoly and H. Amirkhani, “Fake news detection on social media using a natural language inference approach”, *Research Square*, 2020. <https://doi.org/10.21203/rs.3.rs-107893/v1>
- [19] European External Action Service. EUvsDisinfo: Identifying and countering disinformation campaigns, 2024. <https://euvsdisinfo.eu/>
- [20] NATO Strategic Communications Centre of Excellence. Publications on Information Manipulation and Interference. StratCom COE, 2024. <https://stratcomcoe.org/publications/>
- [21] European Union Agency for Cybersecurity. ENISA Threat Landscape 2024, 2024. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024>
- [22] Europol. Internet Organised Crime Threat Assessment (IOCTA). Europol, 2024. <https://www.europol.europa.eu/iocta-report>
- [23] MITRE Corporation. MITRE ATT&CK® Knowledge Base, 2024. <https://attack.mitre.org/>
- [24] Verizon. 2025 Data Breach Investigations Report, 2024. <https://www.verizon.com/business/resources/reports/dbir/>
- [25] IBM Security. X-Force Threat Intelligence Index, 2024b. <https://www.ibm.com/security/data-breach/threat-intelligence>
- [26] Federal Bureau of Investigation. FBI Releases Annual Internet Crime Report, 2025. <https://www.fbi.gov/news/press-releases/fbi-releases-annual-internet-crime-report>
- [27] U.K. Government. Cyber Security Breaches Survey 2024, 2024. <https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2024/cyber-security-breaches-survey-2024>
- [28] Organization for Economic Co-operation and Development. Disinformation and Democratic Resilience. OECD, 2024. <https://www.oecd.org/digital/disinformation/>
- [29] Stanford Institute for Human-Centered Artificial Intelligence. AI Index Report 2024. Stanford University, 2024. <https://hai.stanford.edu/ai-index/2024-ai-index-report>
- [30] McKinsey Company. The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. McKinsey Global Institute, 2024. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024>
- [31] Organisation for Economic Co-operation and Development. Emerging divides in the transition to artificial intelligence. OECD Publishing, 2025. https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/emerging-divides-in-the-transition-to-artificial-intelligence_eeb5e120/7376c776-en.pdf
- [32] PwC. Sizing the prize What’s the real value of AI for your business and how can you capitalise? PricewaterhouseCoopers, 2017. <https://www.pwc.com.au/government/pwc-ai-analysis-sizing-the-prize-report.pdf>
- [33] IBM Security. Cost of a Data Breach Report 2024. IBM Corporation, 2024a. <https://cdn.table.media/assets/wp-content/uploads/2024/07/30132828/Cost-of-a-Data-Breach-Report-2024.pdf>
- [34] Global Disinformation Index. Annual Risk Assessment: Online Disinformation, 2024. <https://www.disinformationindex.org/research/>
- [35] World Economic Forum. Global Risks Report 2024, 2024. <https://www.weforum.org/publications/global-risks-report-2024/>

- [36] X. Zhou and R. Zafarani, “A survey of fake news: Fundamental theories, detection methods, and opportunities”, *ACM Computing Surveys*, vol. 53, no. 5, article 109, pp. 1–40, 2020. <https://doi.org/10.1145/3395046>
- [37] S. Sharma and V. Kalra, “A deep learning based approach for fake news detection”, *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 8, no. 3, pp. 1–8, 2021. <https://doi.org/10.32628/IJSRSET218366>
- [38] J. Wang, J. Li, S. A. Aljaljan, et al., “SA-Bi-LSTM: Self-attention with bi-directional LSTM-based intelligent model for accurate fake news detection to ensure information integrity on social media platforms”, *IEEE Access*, vol. 12, pp. 48436–48452, 2024. <https://doi.org/10.1109/ACCESS.2024.3382832>
- [39] Z. Wang, J. Cheng, C. Cui and C. Yu, “Implementing BERT and fine-tuned RoBERTa to detect AI-generated news by ChatGPT”, *arXiv*, 2023. <https://doi.org/10.48550/arXiv.2306.07401>
- [40] J. D. West and C. T. Bergstrom, “Misinformation in and about science”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 15, article e1912444117, 2021. <https://doi.org/10.1073/pnas.1912444117>
- [41] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat and S. Venkataraman, “Deep learning approach for intelligent intrusion detection system”, *IEEE Access*, vol. 7, pp. 41525–41550, 2019. <https://doi.org/10.1109/ACCESS.2019.2895334>
- [42] K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, “Fake news detection on social media: A data mining perspective”, *ACM SIGKDD Explorations Newsletter*, vol. 19,