Vol. 12 No. 3, 2026
Baltic Journal of Economic Studies
DOI: https://doi.org/10.30525/2256-0742/2026-12-3-135-139

HUMAN RIGHTS AND ETHICS IN AI-DRIVEN STRATEGIC COMMUNICATIONS: RISKS, MITIGATION, AND PROTOCOLS

Oleksandr Cherep[1], Yuliia Kaliuzhna[2], Svitlana Markova[3]

Abstract. The rapid integration of artificial intelligence (AI) into strategic communications offers significant opportunities but also poses complex ethical challenges. This study examines the ethical and human rights implications of AI-driven systems used for disseminating information, making decisions and engaging with audiences. Key risks identified include algorithmic opacity, bias and discrimination, misinformation, privacy violations, the manipulation of public opinion and security threats. These risks have the potential to undermine social trust, democratic processes and individual autonomy. A two-stage methodological framework is employed, combining a theoretical risk assessment based on digital ethics, human rights theory and socio-technical analysis, with a technical evaluation using Long Short-Term Memory (LSTM) neural networks. The first stage categorises ethical risks, and the second stage uses machine learning techniques to detect and quantify these risks in textual communication. The study also looks at ways to reduce the risk, such as Explainable AI, human oversight, testing for bias, monitoring content, and creating regulatory frameworks. This shows that AI can be both a source of risk and a tool for ethical governance. The findings highlight the importance of integrating ethical principles, human rights considerations and robust governance mechanisms into the deployment of AI in strategic communications. Combining technological solutions with organisational policies and human oversight enables AI to enhance communication efficiency and innovation, while safeguarding individual rights, promoting trust and supporting democratic and social integrity.

Keywords: ethical risks, algorithmic bias, misinformation detection, AI governance, human oversight, neural networks, strategic communication.

JEL Classification: D83, K38, O33, M14, K24

1. Introduction

The rapid expansion of artificial intelligence (AI) in strategic communications has created an urgent need to examine the ethical and human rights implications of these technologies. As governments, businesses and media organisations increasingly rely on AI-driven systems for disseminating information, making decisions and engaging with audiences, the potential for both positive impact and significant harm grows. This makes a systematic ethical framework essential, rather than simply desirable, for ensuring that AI serves the public interest.

A key concern is the risk posed by opaque algorithms, data misuse and automated content generation. AI systems can reproduce discrimination, manipulate public opinion and infringe privacy through large-scale surveillance without intent. These risks are exacerbated in strategic communication environments, where messaging has the power to directly influence democratic processes, social stability and individual autonomy. It is crucial to understand the nature of these risks in order to preserve fundamental human rights such as freedom of expression, equality, and personal dignity.

Equally important is the need for robust strategies to minimise harm. Ethical strategic communication with AI necessitates transparent data practices, inclusive design and the continual assessment of algorithmic impacts.

[1] Zaporizhzhia National University, Ukraine *(corresponding author)*
E-mail: cherep2508@gmail.com
ORCID: https://orcid.org/0000-0002-3098-0105
[2] Zaporizhzhia National University, Ukraine
E-mail: kalyuzhnaya.ju@gmail.com
ORCID: https://orcid.org/0000-0002-3335-6551
[3] Zaporizhzhia National University, Ukraine
E-mail: masvvi@outlook.com
ORCID: https://orcid.org/0000-0003-0675-0235

It is crucial to build systems that can detect and mitigate bias, prevent misinformation and safeguard vulnerable groups, to ensure that technological innovation does not come at the expense of social justice. Human oversight and multidisciplinary collaboration are central to the responsible deployment of AI.

Finally, it is vital to develop clear protocols and governance mechanisms in order to institutionalise ethical practices. This involves setting regulatory standards, accountability frameworks and organisational guidelines that define the appropriate use of AI in communication. Such protocols help to prevent abuse and build trust among stakeholders, from end users to policymakers. By incorporating human rights principles into every stage of AI-supported communication, society can harness the transformative potential of these technologies more effectively while upholding the values that underpin democratic life.

Artificial intelligence (AI) is considered a general-purpose technology (GPT) with the potential to transform various sectors of the economy and social life. Significant contributions to this perspective were made by Goldfarb (2024) and Bekar, Carlaw and Lipsey (2018). Goldfarb (2024) draws comparisons between AI and electricity and computers, emphasising its universality and ability to stimulate innovation across numerous industries.

In their 2018 study, Bekar, Carlaw and Lipsey (2018) identified six key characteristics of GPTs. These characteristics include complementarity with other technologies, the lack of close substitutes, and evolutionary development from simple to complex forms. These characteristics were utilised as the theoretical foundation for the investigation of AI as a GPT in this study.

Researchers such as Qian Y., Siau K. L. and Nah F. F. (2024); Horvitz E., Conitzer V., McIlraith S. and Stone P. (2019); and Goldfarb A. (2024) analyse a wide range of consequences of AI adoption, including economic, social, ethical and security-related issues. They highlight potential risks to the labour market, threats of information manipulation and challenges associated with autonomous weapon systems. At the same time, however, Goldfarb (2024) and Felten, Raj and Seamans (2024) also acknowledge the positive potential of AI, including increased productivity, stimulated economic growth and accelerated innovation.

In the field of public policy, seminal contributions have been made by Horvitz E., Conitzer V., McIlraith S., Stone P. (2024), and Ulnicane I., Erkkilä T. (2023) in their examination of pivotal issues in AI regulation. The works of these scholars emphasise the need for global coordination, the creation of effective legal mechanisms, addressing ethical responsibility issues, and integrating socio-technical narratives into policymaking processes. Walter's Y. (2024) approach is of particular value, as it proposes the concept of "dynamic laws" – regulatory frameworks capable of adapting to rapid changes in the AI field. Walter's Y. (2024) also underscores the significance of incorporating technical experts and ensuring transparency in the development and adoption of novel regulations.

The objective of this study is to identify the ethical risks associated with the utilisation of AI in strategic communications, including bias, misinformation, and infringement on human rights.

2. Research Methodology

A theoretical model of AI-related ethical risks has been constructed, drawing on digital ethics, human rights theory and socio-technical analysis. Concepts such as algorithmic opacity, bias, autonomy, accountability and privacy serve as core analytical categories. Existing frameworks, such as AI risk taxonomies, fairness metrics and accountability guidelines, are synthesised to define the primary dimensions of risk: discrimination, misinformation, privacy harm, manipulation and security threats.

This study's methodological framework integrates ethical risk assessment with deep learning approaches to analyse AI-driven strategic communication. The research is structured into two main stages: first, the identification and categorisation of risks; then, the technical analysis and development of mitigation strategies using LSTM models.

The first stage involves systematically identifying and classifying the ethical risks associated with AI in strategic communication. Using digital ethics, human rights theory and socio-technical analysis, the risks are grouped into the following categories:

1. Algorithmic bias and discrimination (risks of reproducing or amplifying social inequalities through automated decision-making).
2. Misinformation and manipulative content (dissemination of false or emotionally manipulative messages affecting public opinion).
3. Privacy violations (unauthorised collection, analysis, or leakage of personal data).
4. Lack of transparency and accountability (opaque algorithms that hinder human oversight).
5. Security threats (risks associated with autonomous systems and malicious content propagation).

The data sources used at this stage include publicly available datasets of news content, social media posts, corporate communication records and platform transparency reports. Each instance is manually or semi-automatically annotated according to risk type, severity, and affected stakeholders. This provides a structured corpus for subsequent technical analysis.

The second stage employs Long Short-Term Memory (LSTM) neural networks to analyse the textual content of strategic communication and detect potential ethical risks. The approach includes the following steps:

1. Data Pre-Processing:
- Text tokenisation, lemmatisation, and removal of stop words.
- Conversion of text into numerical representations using embeddings (e.g., Word2Vec, GloVe, or contextual embeddings).

2. LSTM Model Construction:
- A sequence-based LSTM architecture is designed to capture the temporal and contextual dependencies of language in communication messages.
- The model is trained to classify messages according to risk type (bias, misinformation, manipulative intent, privacy violation).

3. Risk Scoring and Interpretation:
- Model outputs are mapped to risk scores representing the likelihood and severity of ethical concern.
- Attention mechanisms or feature importance methods are applied to interpret which textual elements contribute most to identified risks.

4. Validation and Evaluation:
- Model performance is evaluated using standard metrics (accuracy, F1-score, precision, recall) on annotated datasets.
- Cross-validation ensures generalisation and robustness across diverse communication scenarios.

5. Integration into Ethical Governance Framework:
- The results from LSTM analysis inform risk-mitigation strategies, including content moderation, bias reduction, and privacy safeguards.
- Human oversight and policy guidelines are applied in tandem to ensure responsible deployment of AI in communication contexts.

3. Research Results

There are numerous studies dedicated to identifying the ethical risks of AI use, among which the works of Douglas (Douglas D.M., Lacey J., & Howard D., 2024), Matthew G. Hanna (Matthew G. Hanna, Liron Pantanowitz, Brian Jackson, 2025), Stockman C. (Stockman, 2024) etc., should be highlighted.

Analysing the ethical and legal risks associated with AI in strategic communications reveals the complex challenges posed by this emerging technology. Key concerns include algorithmic opacity, bias, the manipulation of public opinion, privacy violations and security threats. While AI systems offer significant potential for efficiency and innovation, these risks demonstrate that they can also inadvertently undermine fundamental human rights, social equality, and democratic processes if not properly governed. The diversity of these risks emphasises the need for a comprehensive approach that considers both technical vulnerabilities and societal impacts.

Effective mitigation requires a combination of technological, organisational and regulatory measures. Tools that enhance transparency, such as Explainable AI, rigorous bias testing and human-in-the-loop

Table 1
Ethical and Legal Risks of AI in Strategic Communications

Risk name	Brief description	Mitigation methods
Opacity (Opaque Algorithms)	A lack of clarity in how the model functions makes it difficult to detect errors, manipulation and discrimination, and reduces the accountability of developers.	Explainable AI (XAI), transparent algorithm documentation, model audits, human-in-the-loop.
Discrimination and Algorithmic Bias	AI may reproduce or amplify social inequalities, thereby violating equality principles.	Regular bias testing, training data correction, fairness metrics, inclusive design.
Public Opinion Manipulation	Automated content generation has the power to influence political processes, shape public opinion and increase polarisation.	Fact-checking, generative content control, monitoring message dissemination, ethical AI frameworks.
Mass Privacy Invasion	The use of AI for monitoring, big data analysis or covert surveillance threatens privacy rights.	Differential privacy, data minimisation, encryption, strict data handling policies.
Spread of Disinformation and Fakes	Generative models can produce content that is plausible but false, and this accelerates the dissemination of harmful content.	NLP-based fake news detection, fact-checking, media literacy promotion, source verification.
Lack of Effective Control and Accountability	There are insufficient regulations and international coordination, and it is difficult to assign responsibility for AI-induced harm.	Development of regulatory standards, protocols, ethical codes, institutionalised auditing and reporting.
Undermining Autonomy and Freedom of Expression	Manipulative algorithms may restrict access to information or influence individual choices.	Transparent recommendation algorithms, source diversity, human-in-the-loop, protection of informed decision-making rights.
Security Threats	The use of AI in autonomous systems, such as weapons, targeted attacks and content tampering.	Cybersecurity measures, anomaly monitoring, safe deployment protocols, limiting autonomy of critical systems.

Source: compiled by the authors

oversight, are critical for reducing algorithmic harm. Other complementary strategies include fact-checking, monitoring content dissemination, minimising data, implementing cybersecurity protocols and developing clear accountability frameworks. Integrating these measures into AI deployment enables organisations to balance innovation with ethical responsibility, safeguarding individual rights and promoting trust in AI-mediated communication.

As a countermeasure, AI methods can be used to reduce risks. Among the datasets for training ML and NN models, it is worth noting the LIAR Dataset, Enron Email, FakeCovid Fact-Checked News Dataset, etc., which are available on the Kaggle platform (Kaggle). For training AI detection, it is advisable to use the MiRAGeNews datasets, Community Forensics, etc.

As a result of analysing the approaches presented on Kaggle, the use of SVM, Logistic Regression, Random Forest, LSTM, etc., can be highlighted. For example, a simple LSTM model:

```
from tensorflow.keras.layers import Input, Embedding, Dropout, Bidirectional, LSTM, Dense
from tensorflow.keras.models import Model
inputs = Input(shape=(max_length,))
x = Embedding(input_dim=vocab_size,
    output_dim=embedding_dim,
    weights=[embedding_matrix],
    trainable=False)(inputs)
x = Dropout(0.2)(x)
x = Bidirectional(LSTM(n, return_sequences=True, dropout=0.2, recurrent_dropout=0.2))(x)
x = Bidirectional(LSTM(n, dropout=0.2, recurrent_dropout=0.2))(x)
x = Dense(n, activation='relu')(x)
x = Dropout(0.5)(x)
outputs = Dense(1, activation='sigmoid')(x)
)
```

with 4 layers achieved an accuracy of 0.705 on the validation data.

However, particularly high results can be achieved by using an ensemble of models together with standard ML models and pre-trained NN modules, reaching accuracy of up to 0.9. For example:

```
stack_train = np.vstack((nn_pred_train, lr_pred_train))
stack_test = np.vstack((nn_pred_test, lr_pred_test))
stack_inputs = Input(shape=(2,))
x = Dense(4, activation='relu')(stack_inputs)
x = Dense(2, activation='relu')(x)
stack_outputs = Dense(1, activation='sigmoid')(x)
```

In order to address the ethical risks associated with AI in strategic communications, it is essential to implement a combination of technological, organisational and regulatory measures. One key approach is to use AI to detect and mitigate potential harms. Techniques such as natural language processing (NLP) and long short-term memory (LSTM) neural networks can identify biased, manipulative or false content in real time, enabling organisations to intervene proactively before harm spreads. Using ensemble models that combine multiple machine learning and deep learning approaches can enhance detection accuracy, providing a robust toolset for ethical oversight.

Human oversight is a vital part of effective countermeasures. Incorporating human-in-the-loop mechanisms ensures that AI-driven decisions are reviewed by trained professionals, thereby preventing automated systems from reinforcing biases or inadvertently disseminating misinformation. Furthermore, transparency-enhancing tools such as Explainable AI (XAI) enable stakeholders to comprehend model decisions, thereby fostering accountability and trust. Regular audits, bias testing and monitoring of content dissemination complement these technical solutions, ensuring continuous alignment with ethical standards.

These measures are further reinforced by organisational and regulatory strategies. Defining acceptable uses of AI in communication through clear protocols, ethical guidelines, and accountability frameworks helps to ensure consistent enforcement through regulatory coordination at national and international levels. Data protection practices, including encryption, data minimisation and differential privacy, safeguard individual rights and mitigate privacy risks. Together, these measures create a comprehensive ethical governance framework that balances innovation with social responsibility. This ensures that AI supports democratic processes and human rights, rather than undermining them.

4. Conclusions

Analysing AI in the context of strategic communications highlights the transformative potential of these technologies, as well as the significant ethical risks they pose. Key concerns include algorithmic opacity, bias, misinformation, privacy violations and security threats, all of which can undermine human rights, democratic processes and social trust. Although AI can offer opportunities for greater efficiency, innovation and audience engagement, its misuse or unregulated deployment can exacerbate social inequalities and manipulate public opinion. This highlights the urgent need for ethical oversight.

Effective mitigation of these risks requires a multifaceted approach combining technological, organisational and regulatory strategies. Critical techniques include Explainable AI, bias testing, human-in-the-loop systems, content monitoring and secure data practices. Furthermore, AI itself can aid in the detection of misinformation, manipulative

content, and bias through machine learning and neural networks such as LSTM models, thereby demonstrating its dual role as both a source of risk and a tool for risk management. Integrating ensemble modelling with traditional machine learning approaches can further enhance the accuracy and reliability of ethical risk detection.

Ultimately, the responsible deployment of AI in strategic communications requires the embedding of ethical principles, human rights considerations and governance frameworks throughout the development and implementation process. Regulatory standards, accountability mechanisms and transparent organisational policies are essential to ensure that AI supports social good rather than undermines it. By combining technical solutions with human oversight and multidisciplinary collaboration, organisations can reap the benefits of AI while safeguarding individual rights, promoting trust, and sustaining democratic and social integrity.

References:

Goldfarb, A. (2024). Pause artificial intelligence research? Understanding AI policy challenges. *Canadian Journal of Economics,* 57(2), 363–377. https://doi.org/10.1111/caje.12705

Bekar, C., Carlaw, K., & Lipsey, R. (2018). General purpose technologies in theory, application and controversy: a review. *Journal of Evolutionary Economics*, 28(5). https://doi.org/10.1007/s00191-017-0546-0

Qian, Y., Siau, K. L., & Nah, F. F. (2024). Societal impacts of artificial intelligence: Ethical, legal, and governance issues. *Societal Impacts,* 3, 100040. https://doi.org/10.1016/J.SOCIMP.2024.100040a

Horvitz, E., Conitzer, V., McIlraith, S., & Stone, P. (2024). Now, Later, and Lasting: 10 Priorities for AI Research, Policy, and Practice. *Communications of the ACM,* 67(6), 39–40. https://doi.org/10.1145/3637866

Felten, E., Raj, M., & Seamans, R. (2024). Generative AI Requires Broad Labor Policy Considerations. *Communications of the ACM*, 67(8), 29–32. https://doi.org/10.1145/3637864/ASSET/5FB620B0-5F2E-4FFE-992A6844360C94D4/ASSETS/GRAPHIC/3637864_FIG01H.JPG

Ulnicane, I., & Erkkilä, T. (2023). Politics and policy of Artificial Intelligence. *Review of Policy Research*, 40(5), 612–625. https://doi.org/10.1111/ropr.12574

Walter, Y. (2024). Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation – A contemporary overview and an analysis of socioeconomic consequences. *Discover Artificial Intelligence,* 4:1, 4(1), 1–24. https://doi.org/10.1007/S44163-024-00109-4

Douglas, D.M., Lacey, J. & Howard, D. Ethical risk for AI. *AI Ethics* 5, 2189–2203 (2025). https://doi.org/10.1007/s43681-024-00549-9

Matthew G. Hanna, Liron Pantanowitz, Brian Jackson, Octavia Palmer, Shyam Visweswaran, Joshua Pantanowitz, Mustafa Deebajah, Hooman H. Rashidi, Ethical and Bias Considerations in Artificial Intelligence/Machine Learning, Modern Pathology, Volume 38, Issue 3, 2025, 100686, ISSN 0893-3952, https://doi.org/10.1016/j.modpat.2024.100686

Stockman, C. (2024). Generative AI and the Ethical Risks Associated with Human-Computer Symbiosis. *Weizenbaum Journal of the Digital Society*, 5(1). https://doi.org/10.34669/wi.wjds/5.1.2

Platform Kaggle. https://Kaggle.com

Received on: 17th of March, 2026
Accepted on: 28th of May, 2026
Published on: 03rd of July, 2026