The AI Paradox: Balancing Free Speech with Online Safety in the Era of Automated Moderation

Cosmin PROȘCANU[1]
Miruna PROȘCANU[2]

Abstract

There have been numerous reports of news articles that exhibit predominantly negative quality and tone. These types of incidents happen rather often. However, there are certain situations where the news might be highly detrimental, even though it remains essential and continues to hold importance in our lives. Our approach entails leveraging programming skills to gather news stories, which are subsequently assessed using a toxicity meter and analyzed. This operation is undertaken with the purpose of enriching the data that is fed into the programming solution. Improving the current solutions for detecting toxic news can only help to better understand the patterns and the purpose of this news. Moreover, improving this level of understanding can be achieved by further processing the resulting data with new technologies. For the time being the proposed solution helps with the classification of the most toxic news.

Keywords: *AI, Toxicity, News, Toxicity Score, Machine Learning, Python*

JEL classification: C10, C88, C89

DOI: 10.24818/RMCI.2025.1.145

1. Introduction

The integration of Artificial Intelligence (AI) into the realm of online content moderation has emerged as a critical area of focus for social media companies. This development is particularly significant in the context of addressing toxic language and mitigating the proliferation of harmful content. Social media platforms face mounting scrutiny for their role in amplifying polarizing and divisive discourse, underscoring the urgent need for robust and reliable moderation mechanisms. Consequently, attention has increasingly shifted toward evaluating the efficacy of AI-assisted moderation systems in managing and curbing such challenges. These systems hold the potential to significantly enhance the detection and regulation of toxicity in online interactions, thereby contributing to healthier digital environments (Platnick et al., 2024; Udupa et al., 2022).

[1] Cosmin Proșcanu, The Bucharest University of Economics Studies, cosmin.proscanu @csie.ase.ro
[2] Miruna Proșcanu, The Bucharest University of Economics Studies, miruna.proscanu @csie.ase.ro

The news environment has changed massively in the past few years. Misinformation has become a very important tool in manipulating the news world-wide. The toxicity level of the news has reached an incredible level. This problem is very pressing because it influences the world in negative ways and actually, we can't even quantify how much harm the news toxicity really does.

Advances in artificial intelligence (AI) power tools like deep fakes, automated content generation, and biased recommendation engines. In the wrong hands, these capabilities facilitate the creation of highly deceptive news content, amplify hateful narratives, and manipulate public opinion. The consequence is societal harm, including reduced trust in media, political polarization, and potentially even incitement to violence.

The critical importance of addressing the pervasive issue of misinformation and fake news is underscored by a growing body of scientific research dedicated to understanding and mitigating this phenomenon. This academic focus is complemented by significant policy measures, such as the allocation of nearly €5 million by the European Union in 2018 to combat misinformation and disinformation campaigns. Such initiatives reflect a concerted effort to implement practical solutions, including the development of a rapid alert system to identify and address instances of disinformation effectively (Marin, 2018).

Research suggests that while AI-generated news articles may demonstrate reduced bias in certain areas compared to human-authored content, they are not without significant challenges. These include the risks of disseminating misinformation and reinforcing pre-existing societal prejudices. Such limitations underscore the complexity of integrating AI into journalistic practices and highlight the need for comprehensive strategies to mitigate these risks. A comparative analysis of AI policies and guidelines across 52 news organizations worldwide provides valuable insights into how different entities are addressing these concerns and adapting to the evolving landscape of AI-driven journalism (Merrefield, 2023).

Extensive research and the development of multiple Python scripts were necessary to establish communication with various APIs and retrieve the required information. Yet another issue arose with the manner in which the news was uncovered. There were various formats used, such as JSON and csv. It was necessary to standardize it into a single format.

After obtaining the enriched data, our next goal was to perform a basic statistical analysis to uncover intriguing findings. Ultimately, our main goal was to develop a streamlined framework capable of integrating all of these components seamlessly.

To analyze AI's influence on news toxicity, this paper investigates two news sources. Using Perspective API, we will quantify the toxicity of news articles in a comprehensive manner. While challenges in obtaining, managing, and analyzing diverse news sources exist, this project offers a critical step towards understanding how AI has exacerbated misinformation and negativity in news. The hypothesis is that together with the evolution of AI and its application in the world of news, the quantity of toxic news has increased over the last years.

2. Literature review

The proliferation of news on online platforms has brought about a new aspect of toxicity in our modern society. Extensive research has been conducted on the issue of online toxicity in social media comments and user-generated content. However, recent studies have shed light on how news articles can also contribute to the problem by perpetuating harmful language, biased reporting, or deliberately misleading narratives. (Quattrociocchi et al., 2022,) This particular scenario exemplifies how derogatory remarks made online have the potential to incite a decline in civility within American society.

Also, the acceptance and dissemination of fake news have become a significant concern on the policy agendas of major European countries. This heightened focus stems from the recognition of the profound risks posed by the spread of misinformation. These impacts can adversely affect professional, economic, and social activities. Furthermore, the phenomenon poses broader societal threats by creating social imbalances, shaping public opinion, and interfering with critical legislative and democratic processes. (Pop and Ene, 2019).

As the news industry continues to adopt AI-powered tools for various tasks, such as content generation, summarization, and news writing, it becomes crucial to examine the potential for these AI systems to inherit biases, perpetuate harmful stereotypes, or even create entirely new forms of toxicity. (Hede et al., 2021)

In this particular scenario, algorithms that rely on neural networks have the potential to exhibit bias due to the data they are exposed to. To address this issue, I will present a few potential solutions that can help mitigate this bias. The proposed solutions bear resemblance to other existing machine learning algorithms, albeit with a touch of refinement. (Hede et al., 2021)

The potential issues arising from the use of artificial intelligence are incredibly extensive, to the point where certain organizations actively monitor and document the most significant incidents involving AI. An organization called AIAAIC from the USA is dedicated to identifying the most significant AI incidents worldwide. They have successfully developed a comprehensive database that is freely accessible to all. (AIAAIC, 2024)

Many researchers are currently dedicated to enhancing algorithms for toxicity detection. The piece that follows specifically examines Deep Learning models in relation to online comments. There are certain limitations to consider in this case. The primary focus is on comments rather than news, which may restrict the scope of analysis. Additionally, the language barrier poses a challenge when trying to narrow down the geographical areas of interest.

It has been noted in certain academic papers that the toxicity of news extends beyond just the English language, encompassing other languages as well. (Dessì et al., 2021)

Another perspective considers the issue of identifying different types of toxicity in Bulgarian news articles, where natural language processing tools are not as advanced as they are for English. In order to address this issue, the authors have developed a distinctive dataset consisting of Bulgarian news articles that have been

manually categorized based on the type of toxicity. They conducted experiments with various feature representations, such as ELMo, BERT, XLM, and domain-specific ones. They then trained a multi-class classifier and opted for a meta-classifier approach due to the dataset's size. Their findings showcase the possibility of identifying toxicity in Bulgarian news and establish a standard for future research in this field. (Dinkov et al., 2019)

Hopefully, at a point in time there will be a chance for uniformity in the approach towards identifying and solving AI bias, in a standardized way. In conclusion, there is a lot of work to be done in this field, but there are some promising results with the likes of AIAAIC who are gathering valuable information for a public database, or researchers who try to upgrade the process as much as possible. (Schwartz et al., 2022)

An alternative perspective explores the identification of news articles on social networks that can potentially generate toxicity. It provides a comprehensive analysis of the subject matter, examining it through the lens of computer science and distinguishing it from other related concepts such as hate speech. The study utilizes a dataset sourced from the Stop PropagHate project, wherein comments are categorized as toxic through the application of the Perspective API. When the average toxicity of comments surpasses or equals the median toxicity (11.1%), news is deemed to be generating toxicity. The study seeks to forecast the generation of toxic news and gain insights into the factors that contribute to it. The most optimal model, which incorporates both meta-data and news content features, attained an impressive F1 score of 0.74. Factors that play a crucial role in classification are the quantity of comments a news article receives and the presence of title keywords associated with contentious social issues. (Braga da Cruz, 2020)

3. Methodology

3.1 Perspective API

The Perspective API, created by Jigsaw (a subsidiary of Alphabet Inc.), employs advanced machine learning algorithms to detect and classify offensive language. It is a machine learning-powered tool designed to aid in the moderation of online content by identifying comments that may be toxic or abusive. It analyzes text input and assigns a toxicity score on a scale from 0 to 1, providing moderators and developers with a quantitative assessment of user-generated content. Generally, content is classified as toxic when its score falls within the range of 0.5 to 0.7 or higher. (Nogara et al., 2023).

This tool among others represents innovative methodologies for enhancing the navigation and interaction with online news articles. These tools leverage conceptual mapping and categorization techniques grounded in sociological theories, enabling the creation of interactive representations that significantly improve user experiences. By fostering a balance between the democratic ideals of diversity and individual autonomy in news exposure, such systems aim to empower users in making informed decisions about the content they consume. Furthermore, this approach contributes to fostering public discourse by ensuring users retain control over their news preferences while being exposed to a plurality of

perspectives. Insights from real-world projects have demonstrated the importance of overcoming challenges in API integration through effective strategies, further validating the potential of these tools to transform news consumption (Crudu & MoldStud Research Team, 2025).

Natural language processing techniques can be utilized to decode the meaning and sentiment of text. The program utilizes extensive datasets of text that have been identified as poisonous by human reviewers, which aids in training its models to identify patterns of toxicity. Perspective API may take into account contextual clues beyond individual words or sentences, hence enhancing its accuracy. Despite its capabilities, Perspective API, like any other AI system, is not flawless. The presence of biases inherent in its training data or the possibility of overlooking nuances underscores the necessity of human judgment in conjunction with its utilization.

Perplexity API is a tool that uses Machine Learning algorithms to analyze text. If news data or comments or any other type of text is fed into this API, it will try to understand how toxic it is from a human perspective.

The most important aspect about this API is the internal algorithm that has been pre-trained with data validated by humans as being toxic. This feature is extremely valuable and that is the main reason why it was chosen against other solutions. The models used for this analysis score a phrase based on the perceived impact it may have on an individual. It can provide more than a toxicity score, but for the purpose of this article, we only considered this score.

Some other metrics that can be generated are insult, profanity or threat scores, and the list goes on. The machine learning models they use are experimental so all the information should be taken with a grain of salt. The scores are probability scores with a range from 0 to 1, where 1 is the most toxic text one could receive.

In order to connect to this API and in order to feed it the news from our sources, a python solution has been created with a couple of considerations. The calls towards Perplexity API needed to be done via a security token that can be acquired by registering on their official website. The script needed to connect to the news sources and take the relevant text out of them. Next step was to iterate over each text and make a call towards the Perplexity API as it can be seen in Figure 1.

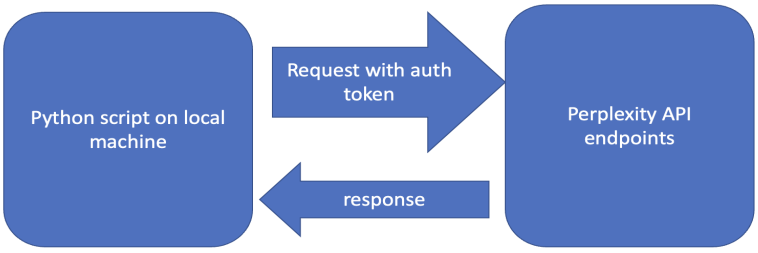

Figure 1. Basic communication between the local python script and Perplexity API
Source: Authors' own creation.

However, this approach has a few limitations because of the risk of flooding Perplexity API with too many calls. For this situation an upgrade for the solution was needed, in the form of a delay mechanism, that could feed the API in a linear predictable manner. Naturally, this process would mean that a lot more time would be necessary to process all the news from the data sources. The simple architecture of the full solution implemented in this article is detailed in Figure 2.

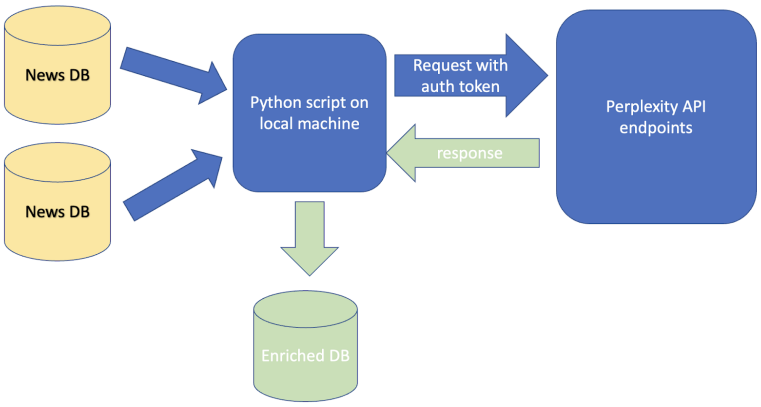

Figure 2. The architecture of the designed solution for enrichment of the news databases
Source: Authors' own creation

This enrichment pipeline can be used for multiple sources of data as it was designed to be a generic mechanism.

For the purpose of this research, I supplemented a dataset of AI incidents from the United States of America with information from the Perspective API in order to assess the potential for toxicity in the reporting of AI issues. This application programming interface (API) assigns toxicity scores to text, facilitating the detection of malevolent language or emotion in the dataset. By employing this methodology, potential correlations between the toxicity levels of news coverage and various types of incidents are sought after. This observation may suggest the presence of biases in reporting pertaining to artificial intelligence.

After the enriched data has been processed, the resulting data manages to capture a couple of behaviors. Since one of the data sets comprises news starting with 2012, carefully curated to display only AI incidents that would naturally have intrinsic toxicity, the values collected with Perplexity API show indeed interesting levels of toxicity.

3.2 Data source

The data has been taken mainly from AIAAIC project which is an independent, non-partisan, public interest initiative that examines and makes the case for real AI, algorithmic, and automation transparency and openness. The news found in the public database curates over 1000 AI incidents starting with 1999.

Another source of news is News API, but that source is not as reliable. The information has a certain element of randomness, and it does not provide news older than 30 days. The interesting aspect is that data from News API must be extracted with Python, as this data is generated in JSON format via a REST API. In this article the main focus was on the AIAAIC news data source, as it was much more reliable and it had the best toxicity scores. Another aspect worth mentioning is that for a small period of 30 days, we could not find very relevant AI toxicity news cases. There were rare cases where we had a score bigger than 0.1. (AIAAIC, 2024)

Identifying trustworthy news sources in Romania has distinct difficulties. This environment is complicated by a mix of restricted access to archives, the emergence of political internet platforms, and the widespread use of social media as the main source of news. Dependable news archives that cover a wide range of historical periods may be fractured or not available in digital format, which makes it difficult to do research spanning several time periods. Distinguishing factual reporting from opinion or purposefully misleading content is challenging due to the abundance of online news channels with various levels of journalistic integrity. In addition, the extensive utilization of social media for news consumption in Romania results in consumers frequently encountering material within echo chambers, lacking the editorial supervision typically found in established news institutions.

4 Results and discussion

4.1 Statistical analysis of the generated results

The data we have enriched using the methods described above is quite fascinating. We began by generating descriptive statistics to gain an understanding of the data's organization. Upon initial examination, it becomes evident that there have been 1367 news reports documenting incidents related to AI. Among these incidents, the highest recorded toxicity level was 0.69. Upon analyzing the data set, one notable news article that caught my attention is titled "Peer-reviewed journal publishes AI-generated rat penis."

Another news item that garnered significant attention was the discovery that Google Autocomplete was suggesting derogatory terms for certain groups, with a toxicity score of 0.63.

Additionally, there is news that is not deemed as harmful as the one mentioned earlier. As an illustration, consider the news titled "Safe Kerala AI camera traffic surveillance," which obtained a toxicity score of 0.01. A discernible pattern begins to emerge. The machine learning algorithms utilized by Perplexity API have a tendency to assign higher scores to news articles that contain more assertive language. When it comes to the application of AI cameras for traffic surveillance, it is crucial to consider the potential negative consequences that may arise if these cameras are misused to track individuals in an unethical manner.

Descriptive statistics of the enriched news data set

Table 1

Toxicity index	
Mean	0.11
Standard Error	0.00
Median	0.05
Mode	0.00
Standard Deviation	0.12
Sample Variance	0.01
Kurtosis	2.18
Skewness	1.56
Range	0.69
Minimum	0.00
Maximum	0.69
Sum	148.73
Count	1367
Confidence Level (95,0%)	0.01

Source: Authors' processing.

The kurtosis value of 2.18 suggests that the data exhibits a distribution that is more peaked compared to a normal distribution. In a normal distribution, there is an increased concentration of data points around the mean, while the tails have a relatively smaller number of data points.

With a skewness value of 1.56, it can be inferred that the data exhibits a positive skew. The tail of the distribution extends more towards the right side than the left side.

The 10 biggest toxicity scores out of all the news data set

Table 2

News Category	Biggest toxicity scores	Headline with max toxicity score
Research/academia	0.68	Peer-reviewed journal publishes AI-generated rat penis
Technology	0.63	Facebook labels black men 'primates'
Politics	0.63	Google Autocomplete suggests Jews, women are 'evil'
Entertainment	0.61	Roblox Condo nazi sex parties
Consumer goods	0.58	Amazon Alexa plays child pornography
Multiple	0.56	GPT-3 advises patient to kill themselves
Business/professional services	0.47	Google ads for Blacks suggest criminal records
Private	0.47	Google Autocomplete links French user to rape
Banking/financial services	0.45	Allstate car insurance 'suckers list' overcharging
Education	0.40	Almendralejo hit by AI naked child images

Source: Authors' processing.

As we can observe from Table 2, the data set presents the top 10 news articles with the highest toxicity levels. With a maximum score of 0.69, and despite the seemingly exaggerated titles, one can only speculate on the potential for further deterioration. However, it is gaining momentum, the notion that machine learning models are more influenced by the intensity of the language used rather than the overall message.

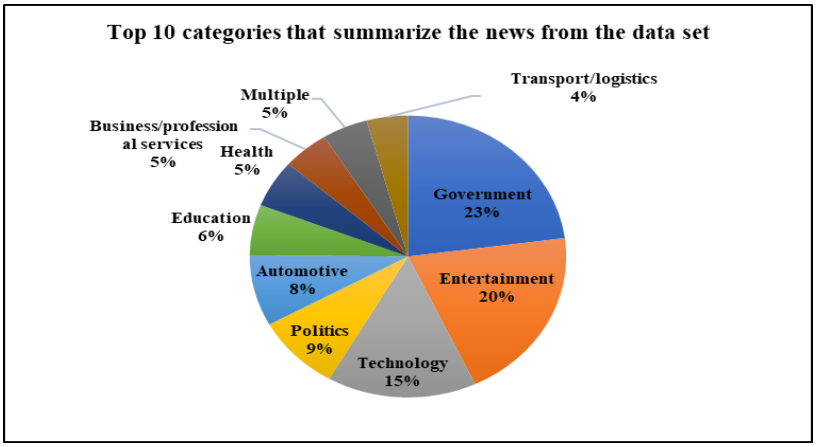

Figure 3. Top 10 categories that summarize the news from the data set
Source: Authors' own research results/contribution

In terms of AI incidents, the Business/professional services category had the highest number, followed by technology and transport/logistics. On the other hand, categories like Health, Education, and Politics accounted for smaller percentages.

Additionally, we have uncovered a minor pattern regarding the toxicity of news articles and the timeframe in which they were published. It is interesting to observe that a significant number of highly concerning AI incidents have been reported between the years 2006 and 2008. It is possible to speculate that the economic crisis of 2007 may have played a role in this aspect. However, it is unfortunate that we are unable to establish a statistical correlation.

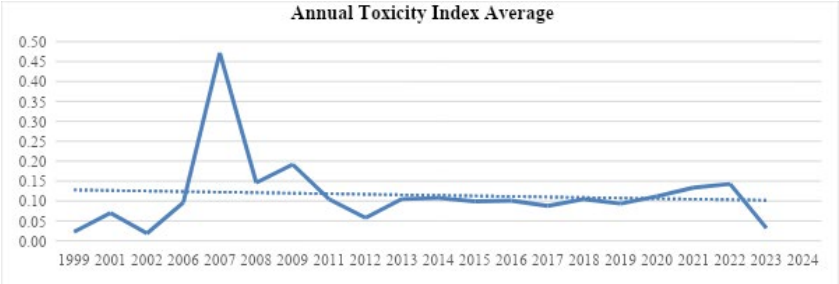

Figure 4. Annual Toxicity Index Average
Source: Authors' own research results/contribution

During that period, there were already discussions in the media about AI systems capable of monitoring individuals' activities, such as the "VioGén gender domestic violence protection system." In 2008, there was a particularly high toxicity score recorded when a French user was linked to a rape incident through Google Autocomplete, scoring 0.47.

5. Conclusions

The subject matter discussed in this article is both intellectually demanding and remarkably varied. The prevalence of AI-driven toxic news has become commonplace, necessitating a nuanced understanding of its impact on individuals. An alternative approach would involve utilizing a significantly larger database of news articles spanning a broader timeframe, with a focus on specific local regions.

The inclusion of the toxicity score and the detailed information about the news allows for a deeper understanding of the impact of toxicity on different regions of the planet. Additionally, this data may reveal previously unnoticed correlations. One of the challenges lies in gathering the necessary data, as there is a scarcity of sources capable of handling large volumes of data. This issue becomes particularly pronounced in less developed regions. Another issue that arises is the exorbitant cost of data, as numerous providers charge exorbitant fees for accessing news databases. One area that could benefit from improvement is the architecture of the solution.

An essential consideration revolves around the limitation imposed by Perspective API. This is due to the fact that the machine learning algorithms employed are not entirely infallible, and currently, the intricate technical intricacies behind them remain unknown. An optimal strategy would involve the development of a novel solution based on existing ones, although this endeavor necessitates both a considerable investment of time and financial resources.

An area of architecture that could greatly benefit from improvement is the data collection process. One might consider developing scrapers to gather daily news and build a comprehensive database of AI toxicity-related articles. The data can be utilized to train the models employed in this solution.

This initial endeavor represents an attempt to explore a fresh perspective on the subject of AI toxicity in the news. This text aims to encompass various aspects of enhancing the process of uncovering pertinent information on this particular subject matter. One of the initial steps involves identifying and uncovering pertinent news. The second step involves conducting machine learning analysis on the identified news articles. Additionally, there is the matter of consolidating all of these articles within a single architecture, which has the potential to enhance itself through the acquisition of news and subsequent algorithm training.

As previously discussed, there are numerous areas for improvement and the statistical analysis section revealed fascinating insights. Upon conducting a thorough examination of the existing literature, it becomes evident that this particular subject warrants significantly greater scrutiny than it has received thus far. The analysis of

news in various regions is a topic that piques our interest. Surprisingly, there seems to be a dearth of relevant research on this subject, particularly in the context of Romania.

Acknowledgements

This work was funded by the EU's NextGenerationEU instrument through the National Recovery and Resilience Plan of Romania - Pillar III-C9-I8, managed by the Ministry of Research, Innovation and Digitalization, within the project entitled „CauseFinder: Causality in the Era of Big Data and AI and its applications to innovation management", contract no. 760049/23.05.2023, code CF 268/29.11.2023

References

1. AIAAIC. (2024). *AIAAIC*. Retrieved March 15, 2024, from http://aiaaic.org.
2. Braga da Cruz, L., (2020, July). *Prediction of toxicity-generating news using machine learning*. Available at: https://hdl.handle.net/10216/128539.
3. Dessì, D., Recupero, D. R., & Sack, H. (2021, March). *An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection within Online Textual Comments*. *Electronics*, (10). Available at: https://doi.org/10.3390/electronics10070779.
4. Dinkov, Y., Koychev, I., & Nakov, P. (2019, September). *Detecting Toxicity in News Articles: Application to Bulgarian*. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 247-258. Available at: https://doi.org/10.26615/978-954-452-056-4_029.
5. Hede, A., Agarwal, O., Lu, L., Mutz, D. C., & Nenkova, A. (2021, April). *From Toxicity in Online Comments to Incivility in American News: Proceed with Caution*. Association for Computational Linguistics, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2620-2630. Available at: https://doi.org/10.18653/v1/2021.eacl-main.225.
6. Morzhov, S. (2020). *Avoiding Unintended Bias in Toxicity Classification with Neural Networks*. Conference of Open Innovations Association (FRUCT), pp. 314-320. Available at: https://doi.org/10.23919/FRUCT48808.2020.9087368.
7. Quattrociocchi, A., Avalle, M., Etta, G., Cinelli, M., & Quattrociocchi, W. (2022, October). *Reliability of News and Toxicity in Twitter Conversations*. Social Informatics: 13th International Conference, SocInfo 2022, Glasgow, UK, Proceedings, pp. 245-256. Available at: https://doi.org/10.1007/978-3-031-19097-1_15.
8. Schwartz, R., Vassilev, A., Greene, K. K., Perine, L., Burt, A., & Hall, P. (2022, March). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. National Institute of Standards and Technology, Gaithersburg, MD. Available at: https://doi.org/10.6028/NIST.SP.1270.
9. Platnick, D., Alirezaie, M., & Rahnama, H. (2024). *Enabling Perspective-Aware AI with Contextual Scene Graph Generation*. Flybits Labs, Creative School, Toronto Metropolitan University, Toronto, ON, Canada; MIT Media Lab, Cambridge, MA, USA. *Information*, 15(12), p. 766. Available at: https://doi.org/10.3390/info15120766.
10. Udupa, S., Maronikolakis, A., Schütze, H., & Wisiorek, A. (2022). *Ethical Scaling for Content Moderation: Extreme Speech and the (In)Significance of Artificial Intelligence*. June 9. (Fall 2021 Joan Shorenstein Fellow). Available at: https://shorensteincenter.org/ethical-scaling-content-moderation-extreme-speech-insignificance-artificial-intelligence/.

11. Crudu, V. & MoldStud Research Team (2025). *Overcoming Challenges in API Integration Through Effective Strategies Derived from Real-World Projects*. Published on 19 February. Available at: https://moldstud.com/articles/p-overcoming-challenges-in-api-integration-through-effective-strategies-derived-from-real-world-projects
12. Merrefield, C. (2023). *Researchers compare AI policies and guidelines at 52 news organizations around the world*. *The Journalist's Resource*, 12 December. Available at: https://journalistsresource.org/home/generative-ai-policies-newsrooms/.
13. Marin, V. (2018). *UE se dotează cu un sistem de alertă rapidă pentru combaterea dezinformării*. *Adevarul.ro*. Available at: https://adevarul.ro/international/europa/ue-doteaza-sistem-alerta-rapida-combaterea-dezinformarii1_5c08c66cdf52022f752be8b6/index.html.
14. Pop, M.I. & Ene, I. (2019). *Influence of the educational level on the spreading of Fake News regarding the energy field in the online environment*. Proceedings of the International Conference on Business Excellence, The Bucharest University of Economic Studies, 13(1), pp. 1108–1117. Available at: https://doi.org/10.2478/picbe-2019-0097.
15. Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., & Giordano, S. (2023). *Toxic Bias: Perspective API misreads German as more toxic*. Proceedings of the 19th AAAI International Conference on Web and Social Media (ICWSM'25). Available at: https://doi.org/10.48550/arXiv.2312.12651.