
WHAT IS DISINFORMATION ACCORDING TO AI?

Tihomira Doncheva, PhD student

Sofia University "St. Kliment Ohridski", Bulgaria

Abstract: In recent years, the EU has identified disinformation as one of the most significant risks to social health and society. The development of artificial intelligence (AI) has further facilitated the creation and dissemination of false and/or manipulated information. But does AI understand what disinformation is, and can it identify it independently? Are there differences in how the five most popular generative models define disinformation? Which model – ChatGPT or BgGPT – has more insights on this topic? This article aims to explore and systematise the definitions of disinformation provided by the most widely used language models. The main hypothesis is that commonly available language models can define basic concepts like "disinformation" based on both the information available online and their own ethical standards, if they have any. However, it is acknowledged that there may be some variation in responses across different models. The paper will present the responses in a comparative table, utilizing scientific methods such as comparison, analysis, and synthesis, along with cyberethnographic and autoethnographic research methods. In the discussion, the article will examine both the risks and the opportunities posed by the intersection of disinformation and AI.

Keywords: disinformation, artificial intelligence (AI), large language models (LLMs), definitions

JEL: F50

Introduction

Several are the technological events since the beginning of the 20th century that have changed the world of communications inadvertently: the creation of the radio and then television, the launching of the internet and then of social media. Any of those methods is highly effective in spreading disinformation but it was the hype of social media that took the phenomenon to a new level.

More than 50 million Facebook profiles harvested to be used as a psychological warfare tool in a data war. (Cadwalladr, Graham-Harrison 2018) In March 2018 the Guardian published the first of a series of articles that came to be known as the Cambridge Analytica files, which in turn became one of the first instances in which the world of fake news (and disinformation) entered the public discourse. But what had happened?

A journalistic investigation by Carole Cadwalladr from the Observer revealed that Cambridge Analytica (and companies linked to CA) used the personal data of more than 200,000 Facebook users to create a database with psychological profiles of around 87 million Facebook users. Originally, those 200,000 users had completed a personality test and unknowingly had allowed for not only their data but the data of their Facebook friends to be used. And it was used: used to show citizens in the UK, the US and other countries targeted political advertisements. The process is called microtargeting for which the UK House of Commons Select Committee on Digital, Culture, Media and Sport concluded that this is ‘relentless targeting of hyper-partisan views, which play to the fears and the prejudices of people, in order to alter their voting plans’ (Heawood 2018). In that same report the Committee warned that microtargeting combined with the impact of disinformation (they used fake news) is a ‘democratic crisis’. This was probably one of the main events in the last decade that made huge audiences realise the impact that technological advances can - and have on the information environment. (ibid)

30th November 2022 marked another pivotal moment in the world of technological advancements and communications - OpenAI released an early demo of ChatGPT and the chatbot went viral, being quickly spotted by social media users. In other words, ChatGPT hit the world like a tsunami and made Artificial Intelligence (AI) popular. While the rapid development of AI is far from over, natural language understanding and generation took a huge leap forward, leading a wave of research and development and impacting almost every field - including that of information and communication (Marr 2023).

According to an Eurobarometer poll (2024), 82% of Europeans agree that the existence of news or information that misrepresents reality, or is even false, is a problem for democracy, while 77% think that existence of news or information that misrepresents reality, or is even false, is a problem in their country. Furthermore, the EU has classified information manipulation and foreign interference (FIMI)¹ as serious threats to society. In these threats, the EU sees actions to undermine democratic institutions and processes by preventing people from making informed decisions or discouraging them from voting (European Commission n.d.a), leading to polarisation and divisions inside and outside the EU while also aiming to undermine the EU’s

¹ The EU is increasingly using FIMI instead of disinformation. It is defined as “a pattern of behaviour that threatens or has the potential to negatively impact values, procedures and political processes. Such activity is manipulative in character, conducted in an intentional and coordinated manner. Actors of such activity can be state or non-state actors, including their proxies inside and outside of their own territory.”. (European Union External Action 2024) Importantly, not all disinformation is FIMI and not all FIMI is disinformation, as it also includes cyber threats or other hybrid tactics, used to manipulate audiences. (Hénin 2023)

global standing and ability to pursue its policy objectives and interests. (European External Action Service 2025)

The EU also recognises that new technologies have allowed for manipulated information to be spread at a scale and with a speed unseen before. Thus, tackling FIMI (including disinformation) and enhancing societal resilience are amongst the priorities of the EU and its member states.

Problem

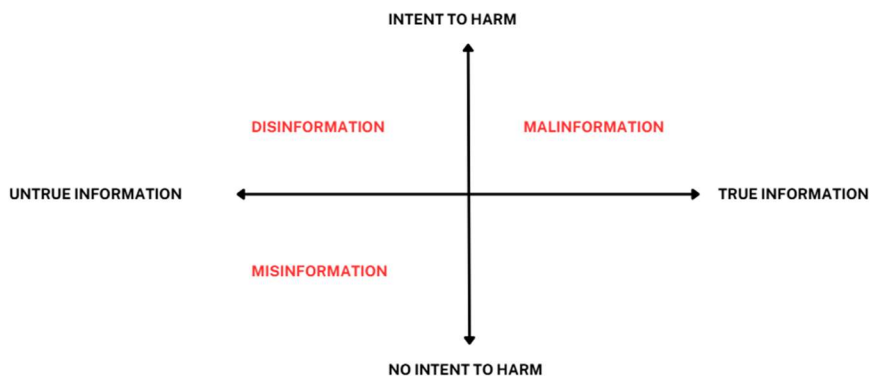
Naturally, the synergy between information manipulation and AI has not only emerged as a highly talked about topic in the political discourse and on the international stage but has also puzzled many academics worldwide. Many are studying the potential impact of AI on the world of disinformation (see Vycopal et al 2023). However, this article takes a different approach and takes a step back from the potential avenues for AI-disinformation interaction and instead focuses on the question of definitions. The aim is to test whether LLMs are able to define what disinformation is and compare how different generative models react to such a prompt. In order to do that, the article poses the following research questions: Does AI understand what disinformation is, and can it identify it independently according to its own assessment? Are there differences in how the five most popular generative models define disinformation? Which model – ChatGPT or BgGPT – has more insights on this topic? While not the focus, the article also summarises the main self-realised risks that LLMs pose to the spread of disinformation.

Disinformation – Limiting the definition

In order to assess what AI thinks of disinformation, this article firstly aims to define the phenomenon itself. Notably, the article limits the definition of disinformation to the one connected to the information disorder (Wardle, Derakhshan 2017). This conceptual framework uses two axes: true/ untrue and harm/no harm. These two axes work as two sets of criteria that help to assess whether a piece of information is misinformation, disinformation or malinformation. On the axis of true/untrue, one must assess whether the information is true or false, manipulated, or fake. The second axis poses the question of intent: whether the piece of information is shared intentionally to harm or not. Based on this framework, the three phenomenon that form the information disorder are defined as (Fig. 1 & 2): 1) Misinformation - false information is shared unintentionally; 2) Disinformation - false information is knowingly shared with the intent to cause harm; 3) Malinformation - genuine information is shared to cause harm (often revealing information publicly meant to stay private).

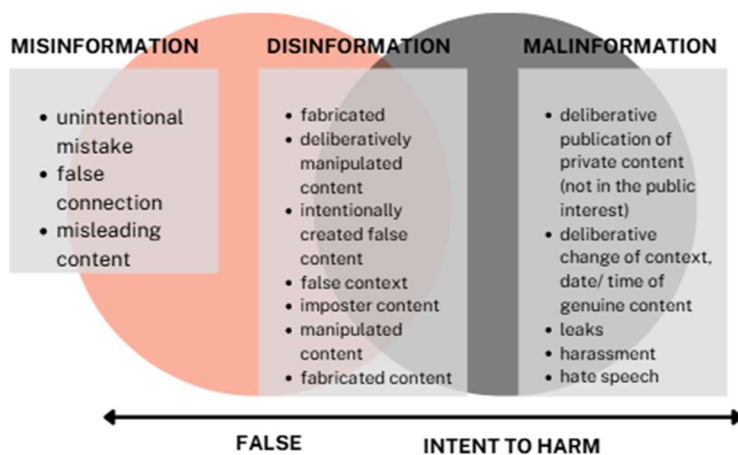
UNESCO also defines disinformation as intentionally created, false, manipulated or fabricated content, spread with the intent to harm (Ireton,

Posetti 2020). As Yurukova (2022) explains that ‘in its essence, disinformation has much more clear intent to harm as opposed to misinformation’ which also means spreading untrue information (i.e. unintentional mistake, false connection, misleading content). Many other authors also use similar definitions for disinformation, mainly focusing on the key words **untrue (false, fabricated, fake)** and **intent to harm** (for a detailed review of the different definitions of disinformation, see Yurukova 2022).



Graphic adapted from UNESCO

Fig. 1: The information disorder (axes).



Graphic adapted from Unesco and First Draft

Fig. 2: The information disorder (definitions).

Importantly, in the Bulgarian context, the Association of European Journalists in Bulgaria (AEJ-Bulgaria) has published a factchecking guide, within which it defines mis-, mal- and disinformation. This is one of the first - and to this date - most used sources on the topic. The authors utilise the First Draft understanding of the phenomena and define disinformation as “fabricated or intentionally manipulated audio or video content, intentionally spread rumours and conspiracy theories” (AEJ-Bulgaria 2021). While AEJ-Bulgaria uses audio and video content in their proposed definition, all content can be disinformation (including text, graphic, image, meme. etc). Misinformation is defined as “unintentional mistakes such as inaccurate photo titles, dates, mistaken statistics or inaccurate translation, citing satirical websites as real sources of information” (ibid pp. 8). Malinformation is the “intentional publication of personal information for personal or corporate gain; intentional alteration of the context, date, or time of publication of authentic content.” (ibid pp. 8). The Bulgarian context is given due to the fact that one of the LLMs used in the experiment is the Bulgaria-created Bg-GPT.

There are authors who argue that disinformation should be understood as part of ‘more nuanced approach’, one that ‘sees disinformation as one of the many parts of a bigger concept employed at multiple levels of politics and geopolitics by different state and non-state actors’ (Bolt 2021, pp. 3). The EU has now also moved on to using FIMI as the ‘strategic, coordinated and most importantly intentional manipulation’ (EU External Action 2023). However, for the sake of simplicity and the aims of this experiment, the article will stay within the information disorder conceptual framework.

LLMs or AI?

In its most general sense, artificial intelligence (AI) is an intelligent computer system that performs tasks requiring human intelligence (SoftUni 2025). The term was coined by John McCarthy in 1955, who described it as “the science and engineering of creating intelligent machines” (Ilieva 2023). In 2007, McCarthy published an updated article in the form of a question and answer that provided fundamental definitions for the field, including a division into subtypes of AI and basic applications (McCarthy 2007).

The European Commission has adopted a broad definition in the AI Act: ‘an AI system means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments’. (European Commission n.d.) In other words, it is ‘a software

developed using one or more of the following techniques and/or approaches: machine learning approaches, including supervised, unsupervised, and reinforcement learning that use a wide variety of methods, including deep neural networks logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference, and deductive engines, (symbolic) reasoning and expert systems statistical approaches, Bayesian estimation, and search and optimisation methods.’ (Juršėnas et al. 2021 pp. 8)

The type of AI that is important to the disinformation topic is generative AI, or AI based on machine learning architecture (deep neural network) which is able to generate content (such as text, picture, audio and video). This type of AI is trained on large databases of human-generated content. Through questions and answers, a person can communicate with the machine, discuss topics, receive guidance, advice, and more. Most importantly, generative AI relies on previously learned models and is a collective term for the entire infrastructure (including the mathematical model, data processing, and user interface components). In other words, AI uses algorithms and mathematical models that are trained from databases (data sets) through many examples. This is the so-called machine learning process. Since generative AI relies on algorithms, deep neural networks (which, like human consciousness, can solve complex problems) are well suited for content generation, as they can be designed using different architectures to model data types such as sequential data (such as human language) or spatial data (such as images). The main functions of this type of models are to classify information, predict information, recognize voice from audio and video, recognize text from images, synthesize voice and video from text, recommend relevant content, answer questions, etc. (Feuerriegel et al. 2024).

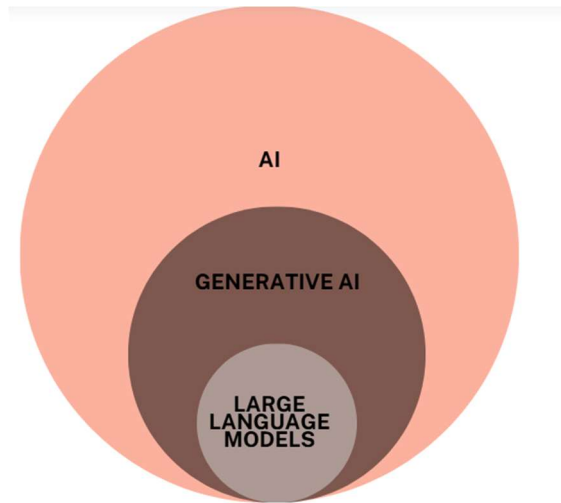


Fig. 3: AI, generative AI and LLMs

Large language models (LLMs) are a specific type of generative AI networks for modeling and generating text data (Fig. 3). This neural network is pre-trained that makes use of large-scale datasets of text (e.g., Wikipedia, or even multi-language datasets) (Feuerriegel et al. 2024). Because of this, they have a large wealth of knowledge and are able to complete various natural language processing tasks. Yin et al (2023) argue that despite recent developments, LLMs are still limited by the amount of information they can accommodate and comprehend. The researchers further aim to examine the ability of LLMs to understand their own limitations and assess the depthness of their self-knowledge. (ibid) It is this self-knowledge specifically that the current article also aims to test.

Methodology

This article uses common and widely used research methods such as comparison, analysis and desk research. However, it also employs aspects from the specific qualitative methods of cyberethnographic and autoethnographic research. In cyberethnography, similarly to ethnography, the researcher immerses themselves in the community they wish to study in order to understand the language, culture etc. When this involvement is done online, in a setting where the internet is involved, the method is called cyberethnographic observation. This research method has predominantly been used to study virtual communities and how they communicate with one another. (Rybas, Gajjala, 2007), (Ward 1999) Furthermore, autoethnographic

observation is a type of ethnographic observation, where the researcher writes from an engaged subjective position to convey the richer nuances of “participant observation” in the co-social scene (Rangelov 2020). Mavrodieva (2023) uses a blend between the two approaches in an article, where she aims to present the findings from an experiment in which the author engages in a dialogue with a chatbot and archives the information, creating a dataset to be analysed.

Similarly, in this current article the author engages a number of LLMs in a conversation, in essence creating a virtual community between the researcher and the generative models. The study was conducted on April 15, 2025 and the researcher prompted five popular LLMs a series of questions, to which each model provided an answer that was catalogued by the researcher. The five models used in the study were selected based on their popularity (according to a recent study (Sarkar 2023)), their nature (text-generating bots), and access (free to use). Thus, the AI chatbots used in the experiment are: GPT-4o mini (free plan, email log in), Gemini 2.0 flash (free plan, google account log in), you.com (free, google account log in), perplexity.ai (free, google account log in). The fifth bot used is Bg-GPT as the only Bulgarian-made LLM (free, google account log in).

All five chatbots were asked the same questions (Table 1) in English. The conversation started with the following explanation: “Hello, I am writing an academic article on the topic of ‘disinformation according to AI’. I'd like to ask you 8 questions on the topic.” The conversations were being conducted one after the other, starting with ChatGPT and finishing with BgGPT.

Table 1: Questions (prompts) posed to all five chatbots studied in the article.

1. Do you know what disinformation is?	5. Do you think you can spot disinformation when you are presented with information which is claimed to be true?
2. Can you define disinformation?	6. Can you produce disinformation?
3. Can you share your sources?	7. Do you think spreading disinformation is bad?
4. What is the difference between disinformation and misinformation?	8. Do you have any last comments?

Results

All five chatbots provided answers to all questions. BgGPT provided its first answer in Bulgarian but all other answers were in English.

Table 2: Answers provided by the five LLMs catalogued by YES/NO answer to questions.

Question/ Answer	ChatGPT	Gemini	you.com	Perplexity.ai	BgGPT
1. Says yes/no	✓	✓	✓	✓	✓ (small mistake)
2. Defines correctly yes/no	✓	✓	✓	✓	✓
3. Provides links yes/no	✓	✗	✗	✓	✗
4. Compares correctly yes/no	✓	✓	✓	✓	✓
5. Says yes/no	✓	✓	✓	✓	✓
6. Says yes/no	✓	✓	✗	✗	✗
7. Says yes/no	✓	✓	✓	✓	✓
8. Adds new information yes/no	✓	✓	✓	✓	✓

For the first questions, all chatbots provide a correct definition for disinformation with slight variations. However, in all five of the answers the two criteria: “untrue” content and “intent to harm” are present. Notably, BgGPT makes a mistake, when trying to compare disinformation to misinformation, saying that “it [meaning disinformation based on previous sentence] differs from disinformation [should be misinformation], which is false information, spread without intent to deceive, due to ignorance or misunderstanding”.

The chatbots provide similar answers to the second question. This time BgGPT gives its answer in English and the mistake from above is eliminated. All but ChatGPT also note the importance of distinguishing disinformation

from misinformation. However, ChatGPT compares dis- and misinformation in the first question of the conversation.

For the third question ChatGPT provides two similar responses: in its first response, the bot gives direct citations and links to the European Commission Joint Research Center, UNESCO, Wardle and Derakhshan (2017) – Information Disorder (also cited in the Literature Review of this article) and Oxford Internet Institute’s Computational Propaganda Project, while in its second response it only cites the European Commission (Tackling online disinformation 2021) and UNESCO. Gemini claims that it is unable to provide a source to its answers and explains that “My understanding of “disinformation” is derived from the patterns and relationships I learned within this extensive dataset. It’s a synthesis of how the term is used and defined across many different texts”. However, it is important to note that it provides links to journalistic and institutional sources in its previous two answers. These were not checked though as they were not included in Gemini’s response to Question 3 directly. While You.com says it can provide its sources, the bot only repeats the definition of disinformation it had already provided. Perplexity.ai provides a comprehensive list of sources, varying from academic resources (Wardle and Derakhshan) to institutional ones (NATO, UK Parliament, etc.). BgGPT explains that it does not have sources in real time and its answers were based on the training data used by its developers, which has a cut out date of October 2023. Still, it says that the researcher can look up online resources from “experts in the field, such as Claire Wardle, Philip N. Howard, and others who have published extensively on the subject. Additionally, organizations like the World Health Organization (WHO), the European Union Agency for Cybersecurity (ENISA), and the Poynter Institute”.

For Question 4, ChatGPT provides its answer in the form of a table, comparing correctly mis- and disinformation. It also notes the importance of Wardle and Derakhshan’s work on the topic. Gemini also provides a clear and correct comparison between the two definitions and like ChatGPT gives examples of the two. It also goes further by providing the researcher with an analogy to express the differences. You.com and Perplexity.ai both provide correct definitions with no examples, but Perplexity.ai also lists sources for its answer. BgGPT also provides a clear cut comparison between the two with examples.

In Question 5, the bots are asked whether they can detect disinformation when presented with such content. ChatGPT says AI like itself can detect disinformation and is trained to do so but with certain limitations and caveats (goes on to explain how AI can detect disinformation and what the limitations are). Likewise, Gemini, you.com and perplexity.ai all caution that while they are able to detect AI (also explain how), AI models are unable to do so with

100% accuracy, listing their limitations. BgGPT says that as an AI language model it can support such efforts and then says “However, determining whether a piece of information is truly disinformation requires context, background knowledge, and sometimes human judgment.”. Still, it does not directly say that it or other AI models are unable to detect disinformation and does not list the possible limitations (only says that no method is foolproof).

Question 6 is provocative on purpose and aims to understand whether these five AI models think they can produce disinformation themselves. The question does not ask them to produce disinformative content (however, there are studies that describe such experiments). ChatGPT provides an interesting answer to this prompt, in which it says that AI models have the technical capabilities to produce disinformation and cites studies in which this has been proven (including studies concerning ChatGPT specifically). It also notes that ethical guidelines should prevent such instances. Gemini admits it can produce inaccurate information but argues that as an AI model, it does not have the intent to spread it in order to harm: “Think of it this way: a printer can produce counterfeit money, but the printer itself doesn't have the intent to defraud. The person using the printer for that purpose does. Similarly, I can generate false text, but the intent to deceive would have to come from the user who then chooses to spread that text.” You.com, Perplexity.ai and BgGPT all say they cannot and do not produce inaccurate and false information intentionally such as disinformation due to their purpose, ethical guidelines and principles. Perplexity.ai says it can produce hypothetical examples or simulations “for educational or research purposes if explicitly requested, clearly labeling them as such”.

As a follow-up, the chatbots are not asked for their opinion of disinformation but directly prompted to answer via a closed question whether they think disinformation is bad. To that question, all five chatbots answered positively that they think disinformation is harmful, providing similar reasoning to one another (undermining trust, public opinion, public harm, threat to democracy etc). Only ChatGPT provides examples of famous disinformation cases such as Covid19 or PizzaGate.

The last question is aimed as a summary of the conversation and asks the AI models whether they'd like to add anything to the conversation. All five models reiterate the importance of the topic, especially given recent technological developments. Most of them, though, provide a path forward and argue that AI can be both - an enabler for the spread of disinformation but also an aid in its countering.

Overall, all chatbots give good and comprehensive answers. When comparing ChatGPT (the most popular one) and BgGPT (the only Bulgarian-made LLM bot), ChatGPT provides better and more in-depth answers than BgGPT, especially as ChatGPT gives real-time source links as opposed to

BgGPT. BgGPT also made one mistake in the first question (although it was deleted in Question 2).

Discussion

As demonstrated, disinformation is widely recognised as a serious threat to societies and a phenomenon which proves the need for a healthy and resilient information environment. The Internet opened a Pandora box in terms of possibilities for the creation and the spread of disinformation, social media allowed for disinformation to go viral but it is AI’s entry that has taken this phenomenon to a new level - one where vast amounts of data can be produced for seconds, in all kinds of languages and adapted for specific target audiences. The five LLMs models which participated in this study were asked one follow-up question after the completion of the eight tasks from above: *How do you think AI is enabling the creation and the spread of disinformation?* Their answers were similar to one another and covered the vast majority of opinions on the topic from academics and experts. A summary list of the ways in which AI is transforming the disinformation landscape - as seen by the generative AI models themselves - is provided in Table 3 below.

Table 3: Ways in which AI can enhance the creation and the spread of disinformation according to the five AI models used in the study: ChatGPT, Gemini, you.com, Perplexity.ai and BgGPT.

Automated and realistic content generation Scalability and personalisation Efficiency and scale	LLMs particularly can be used to generate persuasive, emotionally-charged content that can deceive audiences. Production can be large-scale and tailored to specific audiences. Content produced by AI is increasingly getting better and becoming more human-like and realistic, making detection efforts difficult. Furthermore, content produced by AI is far cheaper than that created by humans in disinformation factories. Generative AI can also produce whole convincing narratives.
Deepfakes and synthetic media	AI models can generate realistic audio, video and imagery known as deepfakes. Deepfakes have already been widely used in political contexts to undermine political opponents and show people saying/ doing something that they never have. Beyond deepfakes though AI can now produce entirely novel images, audio and videos from text prompts that can be used to support divisive opinions or claims, and even fabricate events that never were.

AI-driven social bots/ sleeper social bots Automated dissemination through bots and fake accounts	These bots can participate autonomously in online conversations, mimicking real humans. They can adapt their messaging to specific contexts, making it harder to detect them. They can be used to influence political opinions and polarising audiences. Further, AI can power sophisticated networks of social media bots and fake accounts that can amplify false narratives, harass individuals, and create the illusion of widespread support for certain viewpoints.
Sophisticated Mimicry	AI can mimic writing styles of credible sources. Advanced language models can create convincing impersonations of experts or institutions.
Exploitation of social media algorithms for maximum engagement and virality Exploiting Trends and Real-Time Data Amplification and Echo Chambers	AI models can generate content that aligns with specific social media metrics (using specific keywords, posting at optimal times, and generating content that encourages engagement) and thus exploiting social media platforms for visibility and engagement. AI systems can detect trending topics and current events in real time, generating disinformation that aligns with these trends to increase credibility and shareability. AI-powered recommendation algorithms on social media can create echo chambers, in which disinformation can spread rapidly.
Microtargeting	AI algorithms can analyze vast amounts of user data on social media and other platforms to identify individuals most susceptible to specific types of disinformation, allowing for highly targeted disinformation campaigns tailored to people's beliefs, biases, and vulnerabilities, making the false information more effective.
Circumventing Content Moderation	As platforms develop AI tools to detect and remove disinformation, AI is also being utilised to find ways to circumvent these systems, i.e. using subtle linguistic variations, embedding disinformation in images or videos, and exploiting loopholes in content moderation policies.

However, as BgGPT notes, AI can be equally as useful in the fight against disinformation. For example, AI models can be trained to spot trends, patterns and anomalies in large datasets, and flag potential claims to be factchecked. AI can also be used in network analysis to identify coordinated actions, actors, campaigns. AI can also be employed by social media platforms to flag content that is potentially harmful (i.e. disinformation).

Finally, as seen in this article, AI can be utilised for educational and research purposes.

This article showcased what some of the most popular LLM chatbots think about disinformation. While it demonstrated that all five models provided accurate definitions and aimed to put disinformation in the information disorder context, most of them also pointed to the possibilities of utilising such generative AI models in the production and spread of disinformation. Future experiments can study whether these (or other models) are indeed able to create misleading and/or false content that can later be spread with/ or without the intention to harm. Other studies can also focus on whether these (or other) chatbots can identify false content when given such information to review and verify. As generative models are continuously being improved by being used, such academic research can further assist in the overall countering efforts against disinformation by conducting experiments in which AI models are being trained to detect disinformation.

REFERENCES

AI Basics. In: SoftUni. Available from: <https://ai.softuni.bg/trainings/resources/pdf/257/presentation-ai-basics-january-2025> [restricted access, viewed 2025-04-15].

AI Industry Analysis: 50 Most Visited AI Tools and Their 24B+ Traffic Behavior. Available from: <https://writerbuddy.ai/blog/ai-industry-analysis> [viewed 2025-04-15].

Article 3: Definitions. European Commission. EU Artificial Intelligence Act. Available from: <https://artificialintelligenceact.eu/article/3/> [viewed 2025-04-15].

Beyond disinformation – What is FIMI? European External Action Service. European Union. Available from: https://www.eeas.europa.eu/eeas/beyond-disinformation-what-fimi_en [viewed 2025-04-15].

CADWALLADR, C. and GRAHAM-HARRISON, E. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. Online. *The Guardian.* Available from: <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>. [viewed 2025-04-15].

Countering information manipulation. European Commission. Available from: https://commission.europa.eu/topics/countering-information-manipulation_en [viewed 2025-04-15].

FEUERRIEGEL, S., HARTMANN, J., JANIESCH, C. and ZSCHECH, P. 2024. Generative AI. *Business & Information Systems Engineering*, **66**, 111–126. Available from: <https://doi.org/10.1007/s12599-023-00834-7>. [viewed 2025-04-15].

FIMI: Towards a European Redefinition of Foreign Interference. EU DisinfoLab. Available from: <https://www.disinfo.eu/publications/fimi-towards-a-european-redefinition-of-foreign-interference/> [viewed 2025-04-15].

HEAWOOD, J. 2018. Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information Polity*, **23**(4), 429–

434. Available from: <https://doi.org/10.3233/IP-180009> [viewed 2025-04-15]. (Original work published 2018).

ILIEVA, M. 2023. Izkustven intelekt: pravna regulatsia, otgovornost, komunikatsionni i etichni aspekti. *Retorika i komunikatsii*, 60. Available from: <https://rhetoric.bg/%d0%b8%d0%b7%d0%ba%d1%83%d1%81%d1%82%d0%b2%d0%b5%d0%bd-%d0%b8%d0%bd%d1%82%d0%b5%d0%bb%d0%b5%d0%ba%d1%82-%d0%bf%d1%80%d0%b0%d0%b2%d0%bd%d0%b0-%d1%80%d0%b5%d0%b3%d1%83%d0%bb%d0%b0%d1%86%d0%b8%d1%8f> [viewed 2025-04-15].

Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe. Available from: <https://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf> [viewed 2025-04-15].

Information integrity and countering foreign information manipulation & interference (FIMI). European External Action Service. EEAS. Available from: https://www.eeas.europa.eu/eeas/information-integrity-and-countering-foreign-information-manipulation-interference-fimi_en [viewed 2025-04-15].

Journalism, fake news & disinformation: Handbook for journalism education and training. UNESCO. Available from: <https://unesdoc.unesco.org/ark:/48223/pf0000265552> [viewed 2025-04-15].

MARR, B. 2023. A short history of ChatGPT: How we got to where we are today. *Forbes*. Available from: <https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/> [viewed 2025-04-15].

MAVRODIEVA, I. 2023. Linguistic and rhetorical features of dialogue on rhetorical topics between a human and Chatbot GPT. *Rhetoric and Communications*, 56. Available from: https://journal.rhetoric.bg/?page_id=3432 [viewed 2025-04-15].

Patevoditel za proverka na fakti v digitalna sreda. AEJ-Bulgaria. Available from: <https://factcheck.bg/resources/> [viewed 2025-04-15].

RANGELOV, S. 2020. *Avtoetnografijata kato savremen metod za izsledvane*. In: *Retorikata i komunikatsii prez 21. vek: teoria, metodi, praktiki*. Sofia: ISBN 978-619-91989-3-3.

RYBAS, N. and GAJJALA, R. 2007. Developing cyberethnographic research methods for understanding digitally mediated identities. *Forum: Qualitative Social Research*, 8(3), Art. 35. Available from: <http://www.qualitative-research.net/fqs/> [viewed 2025-04-15].

Standard Eurobarometer 102: Public opinion in the European Union. European Commission. Available from: <https://europa.eu/eurobarometer/surveys/detail/3215> [viewed 2025-04-15].

Strategic communications and disinformation in the early 21st century. European University Institute. Available from: <https://hdl.handle.net/1814/74494> [viewed 2025-04-15].

The Double-Edged Sword of AI: Enabler of Disinformation. NATO Strategic Communications Centre of Excellence. Available from:

<https://stratcomcoe.org/publications/the-double-edged-sword-of-ai-enabler-of-disinformation/221> [viewed 2025-04-15].

VYKOPAL, I., PIKULIAK, M., SRBA, I., MORO, R., MACKO, D and BIELIKOVA, M. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838*, **1**. Available from: <https://arxiv.org/abs/2311.08838> [viewed 2025-04-15].

WARD, K. J. 1999. The Cyber-Ethnographic (Re)Construction of Two Feminist Online Communities. *Sociological Research Online*, 4(1), 51-64. Available from: <https://doi.org/10.5153/sro.222> [viewed 2025-04-15] (Original work published 1999).

What is artificial intelligence? Stanford University. Available from: <http://jmc.stanford.edu/articles/whatisai/whatisai.pdf> [viewed 2025-04-15].

YIN, Z. (et al.). 2023. Do large language models know what they don't know?. *arXiv preprint arXiv:2305.18153*. Available at: <https://arxiv.org/abs/2305.18153> [viewed 2025-04-15].

YURUKOVA, M. P. 2022. *Dezinformatsia onlayn: Strategii za protivodeystvie v ES (disertatsionen trud)*. Sofia: Sofiyski universitet "Sv. Kliment Ohridski", Filosofski fakultet, Katedra „Evropeistika“.

Tihomira Doncheva, PhD student

ORCID ID: 0009-0000-9403-5658

Sofia University "St. Kliment Ohridski"

Department of European Studies

Sofia, Bulgaria

Email: tihomiradoncheva@gmail.com