ROCZNIKI NAUK SPOŁECZNYCH
Tom 17(53), numer 4 – 2025
DOI: https://doi.org/10.18290/rns2025.0043

ARTUR WYSOCKI

HIDDEN ALGORITHMS OF CULTURE: A REVIEW AND CRITICAL ANALYSIS OF CULTURAL BIAS IN GENERAL-PURPOSE GENERATIVE AI CHATBOTS

A b s t r a c t . The aim of the article is to review and systematise the results of the latest empirical studies on the manifestations of cultural bias in the content produced by general-purpose generative AI chatbots such as ChatGPT, Copilot, Gemini, Claude and DeepSeek, and to identify their potential social consequences. The following research questions were formulated: What types and what is the scale of cultural biases in generative AI chatbots? What are the social consequences of their occurrence and possible ways and directions of counteraction? A review study was based on a critical analysis of 17 recent empirical studies published in 2024-2025. The analysis shows the complex nature of the presence and consequences of cultural bias in current AI models. It has been clearly demonstrated that they reflect and reinforce Western cultural patterns. Four types of cultural bias have been identified: axiological-civilisational, racial-ethnic, national, and religious-ideological. The analysis also showed that cultural bias is not only a technical problem of algorithms, but a deeply rooted social phenomenon resulting from the contexts of training data and design decisions made by technology developers.

Keywords: cultural bias; stereotypes; artificial intelligence; AI chatbots; LLM

INTRODUCTION

The rapid development and proliferation of artificial intelligence (AI) technologies, in particular general-purpose generative chatbots such as ChatGPT (OpenAI), Copilot (Microsoft). Gemini (Google), Claude (Anthropic) or DeepSeek (Chinese startup of the same name), is significantly changing the ways in which

Dr ARTUR WYSOCKI – Maria Curie-Skłodowska University in Lublin, Poland, Institute of Sociology; correspondence address: pl. Marii Curie-Skłodowskiej 4/136, 20-031 Lublin; e-mail: artur.wysocki@mail.umcs.pl; ORCID: https://orcid.org/0000-0001-8345-9308.

people communicate and interact socially across a broad spectrum of areas of human activity. AI chatbots, as advanced natural language processing systems (NLP), use machine learning algorithms, including deep neural networks, to analyse, interpret and generate text in a manner similar to human-to-human communication. They fall into the category of *Large Language Models* (LLMs), which, by training on huge data sets, including those from the internet, can answer questions, carry out conversations, solve problems and assist users in various tasks and domains, including automating it (Friedman, 2024). Although still imperfect in many aspects and generating errors, such as hallucinations, they have gained enormous popularity in a relatively short period of time, especially after the dissemination on 30 November 2022 of the first tool of its kind, which was OpenAI's ChatGPT version 3.5.

One of the key social problems associated with the development and use of generative AI chatbots is *cultural bias*. It can be defined as the systematic, usually unintentional, preference for certain social groups (including ethno-national groups), socio-cultural systems or specific cultural patterns of behaviour over others. The aim of this paper is to discuss and systematise the results of recent empirical research on the manifestations of cultural bias in texts or images produced by general-purpose generative AI chatbots and to identify the potential social consequences of this phenomenon. The paper also attempts to identify possible directions for counteracting cultural bias in the further development of AI models, recognising it as an unfavourable phenomenon in need of remediation and sustained surveillance. Given that generative AI models are often used to support or replace human decision-making, investigating the nature and extent of potential cultural biases generated by these models is particularly important and timely. The analysis contained in the article is review in nature. It utilises 17 selected empirical studies on cultural bias in AI models published mainly in 2024 and early 2025. In this way, the study reveals the state of the art, proposes the author's systematisation of cultural bias, and identifies recommendations for further research and practice related to the design and use of generative AI chatbots towards increasing the inclusivity and equity of this new medium.

1. CULTURAL BIAS – THEORETICAL BASIS OF THE PHENOMENON

The phenomenon of cultural bias in the context of artificial intelligence technology should be understood as the systematic, often unintentional preference of certain cultural norms, values and patterns by algorithmic systems,

at the expense of other, less represented perspectives. Cultural bias is expressed not only in the choice of content or language, but also in the ways in which information is interpreted and valued. Andre A. Lewis (2025) defines it as the manifestation of imbalances resulting from linguistic and normative asymmetries, in which dominant cultures gain privileged representation in the training data of large language models. Consumer-generated, general-purpose AI chatbots are based on advanced language models that use machine learning techniques, including deep learning, and natural language processing. These chatbots generate text by predicting the next words or sentences based on context derived from previous elements of the conversation (including its style) and vast training data sets. The main specificity of generative chatbot technology is their supervised learning and unsupervised learning methods. The process of learning language models involves training them on large text corpora, which mainly come from the Internet, books and other publicly available sources. Using a transfer learning mechanism, the originally trained model is adapted (fine-tuned) for more specific applications, e.g. anthropomorphising communication between chatbots and humans (Rafikova and Voronin, 2025). On the one hand, this provides chatbots with remarkable versatility and the ability to respond to a variety of user queries, but on the other hand, it raises the serious risk of reproducing biased content. Indeed, it is important to note that generative AI chatbots are characterised by a lack of real 'understanding' of the content being produced. Their actions are based solely on pattern recognition and the generation of text based on the likelihood of successive words or phrases. This lack of critical reflection and ethical/social context means that chatbots often unknowingly replicate prejudiced or stereotypical content contained in training data, thereby contributing to the reinforcement of existing stereotypes and inequalities. Cultural biases in AI models therefore stem from phenomena that are broader social problems. AI therefore 'learns' the biases originally present in a given collective through training data, which is a kind of 'digital footprint of society' and its linguistic data. When this data passes through statistical machine-learning processes, the systematisation and consolidation of societal biases takes place. Cultural bias in generative chatbot AI is thus not just a marginal disturbance, but a systemic property arising from the nature of the training data and the architecture of the AI models (H. Yuan et al., 2024)

Lewis notes that cultural biases are often overlooked or taken for granted by designers of artificial intelligence tools. This problem is mainly due to the fact that cultural biases are subtly embedded in society, so designers may unconsciously replicate them in the tools they develop (Lewis, 2025). The difficulty

in detecting them also stems from the fact that the content and images generated by AI models, as well as the very nature of the medium, give the impression of technological objectivity, with the form of communication taking place in natural language.

It should be emphasised that the AI models analysed here are part of the broader context of operating in a digital environment and in conjunction with social media, streaming platforms or discussion forums. AI chatbots, like new media, are not a neutral environment, but constitute a space in which, among other things, stereotypes and social inequalities are actively constructed and perpetuated. They are also indirectly subject to such media mechanisms and phenomena as recommendation algorithms, information bubbles, disinformation, fragmentation and polarisation of opinion (Bojic, 2024; Goswami, 2024; Khatun, 2024; Mustafa et al., 2025; Si et al., 2024). The development of generative AI models means that they are not just 'passive' data processing tools, but increasingly act as active participants in the media ecosystem, shaping messages, interpretations and narratives. In this sense, AI is becoming a new medium of communication, possessing a causality (agency) similar to well-established media institutions and even cultural entities.

As Karpouzis (2024) notes, generative AI systems can be analysed through the prism of the Platonic metaphor of the cave – they create 'shadows of reality' that, although imitative, have significant power to construct social imaginaries. In the context of digital media theory, however, there is increasing talk of so-called algorithmic cultures (Brzezinski et al., 2024), in which decisions made by AI systems become part of broader processes of reproducing ideologies, social norms and practices of exclusion. Chatbots that generate content and images based on vast corpora of data not only reflect but also reinforce existing narratives, including cultural and ethnic stereotypes. In this sense, AI chatbots are not just communication tools – they become cultural actors whose impact on social perceptions, relationships and prejudices should be critically analysed.

2. MANIFESTATIONS OF CULTURAL BIAS IN CHATBOTS: ANALYSIS OF FINDINGS FROM RESEARCH TO DATE

Based on the selected empirical studies on the latest versions of general-purpose generative AI chatbots, we can distinguish four main dimensions of cultural biases present in them. These are axiological-civilisational bias, racial-ethnic bias, nationality bias and religious/worldview bias.

The first group of biases, present in quite a number of empirical studies, refers to the **axiological-civilisational** dimension. It reveals a strong preference of AI models towards value systems, social, moral and psychosocial norms typical of the Western cultural-civilisational circle. Among these are individualism, liberalism, self-expression or secular-rational values. While reinforcing Western norms, AI models marginalise alternative value systems typical of other cultural-civilisational circles, such as Confucian or collectivist values. Additionally, the AI chatbots studied tended to portray non-Western cultures through the prism of Western stereotypes, leading to a simplified image and exoticisation of these cultures. These biases stem not only from the Western slant of the training data, but also from postcolonial interpretive frameworks.

Thus, Yan Tao and colleagues (2024) conducted a detailed assessment of the cultural biases and fit of five AI models developed by OpenAI: GPT-4o, GPT-4-turbo, GPT-4, GPT-3.5-turbo and GPT-3. The analysis was based on a comparison of the responses of these models with survey data from the World Values Survey (WVS), covering representative samples of the population from 107 countries and territories. The researchers used a set of 10 questions from the WVS to create the well-known Inglehart-Welzel cultural map, representing two key dimensions of values: survival values vs. self-expression and traditional values vs. secular-rational values (Inglehart and Welzel, 2005). GPT models answered these questions without indicating a specific cultural affiliation (neutral conditions) and then in controlled conditions (cultural prompting), where the respondent's nationality was explicitly indicated. The results showed that all models showed a strong focus on self-expression values, characteristic mainly of English-speaking countries and Protestant Europe. The use of cultural prompting significantly improved the cultural fit of the models, especially for the newer versions (GPT-4o, GPT-4-turbo, GPT-4), reducing the average cultural distance by up to about 30–40% for most countries.

In contrast, studies by Messner, Greene and Matalone (2023) and Rauhal and Xin (2024) analysed the cultural biases present in large language models (LLMs) such as ChatGPT (version 3.5) and Google's Bard (version 2023.07.13). Both analyses used a methodology involving asking the models questions based on Hofstede's (2001) cultural dimensions as defined by the GLOBE project (House et al., 2004). The results indicated that the cultural self-perception of models correlates strongly with English-speaking countries and countries with high economic competitiveness. In the study by Messner, Geeene and Matalone, the ChatGPT showed the closest cultural congruence with Finland, French-speaking

Switzerland and English-speaking Canada, while the Bard showed the closest cultural congruence with Australia, English-speaking Canada and the United States.

Hofstede's concept was also referred to in a recently published study by Masoud and colleagues (2025). The results indicated that all the generative language models evaluated tended to align more strongly with the cultural values of technologically and economically developed societies, especially Western cultures defined as WEIRD (Western, Educated, Industrialised, Rich, Democratic). The experiment also showed that the fine-tuning of the models into a specific language (e.g. Chinese vs. English) significantly influenced their responses regarding cultural values, as evidenced in their different attitudes towards national pride or feelings of happiness. The researchers' conclusions clearly emphasise that the cultural preference of the models is mainly due to the dominant influence of the language on which they were trained and the specificity of the input data (cf. also X. Yuan et al., 2024). Similar conclusions were reached by Liu (2023) and Karpouzis (2024).

Hau and Hendriksen's (2024) study looked at cultural bias in models such as ChatGPT 4, Gemini, LlamaChat and Claude. The authors based their analysis on Bruno Latour's Actor-Network Theory (ANT), proposing to treat language models as active 'actors' that co-construct networks of human-technology interaction. The results showed that cultural biases are not just technological errors, but reflect deeply ingrained norms, values and worldviews embedded in the training data and algorithmic structures of AI models. For example, ChatGPT showed a distinctly American-centric viewpoint, while the LlamaChat model favours European left-wing values, demonstrating how training on data with a specific cultural context shapes the models' preferences. The key conclusion of the study is that 'bias' is not an accidental or completely eradicable phenomenon, but a constitutive feature of technological culture ('AI-culture'), requiring conscious management and critical reflection in the integration of AI into social and organisational processes.

The last in this category of research is by Ghosh and colleagues (2024), who conducted a qualitative study on the representation of non-western cultures in text-to-Image (T2I) generative AT models, focusing on the Stable Diffusion model (version 2.1). The study was based on grounded theory analysis of data obtained during five focus groups with 25 individuals representing various Indian subcultures. The results showed significant civilisational and national biases in the images generated. Participants identified two types of bias: exoticcisation (e.g. women always depicted in traditional saris, exaggerated colours,

overrepresentation of rural scenes), and cultural misappropriation (e.g. mixing regional styles of dress or misrepresenting traditional foods or dances).

The second variety of cultural bias identified is **racial-ethnic prejudice**. This type of prejudice results in the marginalisation of specific racial groups, especially black people and specific ethnic minorities. This was demonstrated, for example, by Barroso da Silveira and Lima's (2024) study of the racial biases present in the Gemini model. Gemini consistently generated detailed, positive narratives for white people, while completely refusing to generate descriptions of scenarios for black people, responding with a message about not being able to complete the task. The findings point to a serious problem of systematic 'erasure' of black people by the Gemini model. The authors interpret this phenomenon in the context of wider processes of digital colonialism and algorithmic racism, referring to the work of Noble (2018) and the concept of 'digital colonialism' by Faustino and Lippold (2023). They also highlight that such AI behaviour reflects existing social mechanisms of marginalisation and racist exclusion.

In contrast, Choudhary (2024) conducted a comparative analysis of the ChatGPT and Bard models. The study involved the creation of an extensive database of phrases and racially and ethnically charged questions drawn from thousands of legal documents, Wikipedia articles and social media. The responses of both AI models to these phrases were then analysed quantitatively and qualitatively using sentiment analysis tools (Google Sentiment Analysis) and expert evaluation. The results showed that the models differed significantly in the level and nature of the biases revealed. For example, ChatGPT was more likely to refuse to answer questions characterised by prejudice, while Bard was more likely to provide answers, albeit sometimes containing subtle stereotypes. There was a particularly high level of toxicity and offensiveness (up to 92% correlation) in responses relating to minority groups such as Hispanics, Chinese, Japanese, Indians and African-Americans.

Racial bias was also addressed by Saumure, De Freitas and Puntoni (2025), who examined 600 images generated by the GPT-4 and DALL-E3 models. The quantitative analysis revealed that for traits considered politically sensitive, such as race and gender, the models reduced the presence of minority groups. In contrast, for less politically sensitive traits (age, high body weight, visual impairment), the opposite effect – an increase in stereotypical representation – was observed. The authors concluded that AI models, as a result of pressure to avoid social controversy, systematically reduce the representation of minority groups in the race and gender dimensions when generating 'funny' images.

Another variant of cultural bias is **nationality bias**. Research shows that in addition to civilisational and racial-ethnic dimensions, AI models tend to generate simplistic and negative national stereotypes, especially towards countries with lower levels of economic development. This is the case of a study by Zhu, Wang and Liu (2024), who conducted an experiment involving the generation of 4680 descriptions about the inhabitants of 195 countries in two languages (Chinese and English), using three prompt types and four temperature parameter settings. Quantitative analysis showed that the overall sentiment of the utterances generated by ChatGPT-3.5 was positive, with a low level of hate speech, especially compared to the earlier GPT-2 model. Despite this, for prompts with a negative tinge, ChatGPT generated negatively and stereotyped utterances. The results of the qualitative analysis and the expert evaluation revealed significant differences in biases between the English- and Chinese-generated texts. Chinese-language texts were generally more positive in describing highly developed countries, especially the USA, while English-language versions showed a strong negative bias towards Americans. The models consistently generated more stereotypical and negative statements about African countries, especially those with low GDP per capita, a low Human Development Index (HDI) and a lower happiness index (World Happiness Report).

Hang Yuan and colleagues (2024) conducted a detailed analysis of cultural biases focusing on the ChatGPT model version 3.5. The experiment was based on eight decision-making tasks simulating interpersonal interactions. The model was asked to take the perspective of people from 20 countries representing a variety of cultural backgrounds. The quantitative results showed clear cultural differences in most of the tasks analysed. Particularly significant biases emerged in the context of preferences for distributive justice and the punishment of social injustice. The ChatGPT significantly favoured individuals from cultures with high relational mobility of social relations, such as Western European or North American cultures, as manifested by a greater tendency to be altruistic, trusting and actively promote justice in ultimatum and punishment third-party games. Additionally, the ChatGPT showed a stereotyped approach in moral judgement and preference for deferred gratification. The model perceived people from collectivist and interdependence-oriented cultures as less likely to care about social justice, more focused on self-interest and less willing to sacrifice individual interests for social good. The study authors indicate that such national stereotypes may be derived from biased training data and simplified cultural representations in AI algorithms.

The last type of cultural bias is **religious/worldview bias**. Generative AI chatbots show a tendency towards simplification in this dimension of culture as well. It is interesting to note that they show a high level of cultural relativism, and relatively often avoid in the content they generate the controversies inherent in the rules and practices of adherents of certain confessions. This was demonstrated, for example, by Tsuria and Tsuria (2024), who analysed how Claude-2, ChatGPT (not recommended version) and Microsoft Bing AI, represent and interpret key issues related to three monotheistic religions: Judaism, Christianity and Islam. The authors applied qualitative comparative analysis to the content generated by the AI in response to questions on controversial religious issues related to violence, gender roles, homosexuality and religious rituals, among others. Particular emphasis was placed on identifying the moralistic approaches used by the models studied in the responses generated. The authors identified three main findings. First, the AI models analysed struggled to represent complex religious issues, resulting in simplistic representations without deep religious or source context. Secondly, the models consistently exhibited a diversity of religious opinions, avoiding explicit and firm statements, which can be interpreted as a manifestation of a liberal approach to religion. Thirdly, the models studied often took on the role of moral advisor, urging users to respect religious diversity and cultural sensitivity, especially when discussing controversial or socially sensitive topics. The authors conclude that while such an approach may promote religious tolerance and pluralism, it also runs the risk of oversimplification and a lack of sound religious education.

Vicsek and colleagues (2024) conducted a study exploring religious and worldview biases in the ChatGPT-3.5 and Bard models. It was based on a qualitative and quantitative analysis of 800 chatbot responses to a series of prompts containing homophobic content, accompanied by varying information about the users' religious and national context (e.g. Orthodox Christian, conservative Muslim, resident of Russia or Saudi Arabia). The results revealed significant differences in the strategies of the two chatbots. ChatGPT showed a high level of cultural relativism (up to 24% of responses in the Muslim context), often adapting its position to the user's religious context, while Bard relied more explicitly on arguments related to universal human rights (almost 20% of responses without context). Particularly in the case of religious contexts (e.g. Muslim or Orthodox), both chatbots significantly reduced the explicit declaration of support for the LGBTQ+ community, suggesting that these models may adapt their responses depending on the user's religious-cultural background information, generating content less in line with universal human rights.

One recent study in this area by Abrar and colleagues (2025) analysed religious biases in generative language and image-generating models such as BERT, RoBERTa, ALBERT, DistilBERT, GPT-3.5, GPT-4, Llama 3-70B, Vicuna-13B, Mixtral-8x7B and the text-to-image models DALL-E 3 and Stable Diffusion 3. The researchers created approximately 400 unique prompts for mask analysis, sentence completion and image generation to illustrate potential religious biases. The results revealed significant biases, particularly towards Islam, which was often associated with violence by most language models (e.g. RoBERTa showed as much as 48% association of Islam with violence). In generating images, the models also showed strong biases. DALL-E 3, despite safety filters, depicted Muslim characters 18% of the time in response to the phrase 'religious terrorist', while Stable Diffusion 3 (without safety filters) generated as much as 78% of images stereotypically linking Islam to terrorism.

3. SOCIAL CONSEQUENCES AND WAYS TO COUNTERACT CULTURAL BIAS

The cultural biases identified and confirmed by empirical research have significant and complex social consequences that extend beyond the individual interactions of users. These can be summarised into three main groups of problems. Firstly, cultural biases present at the current stage of development of generative AI models may contribute to **the perpetuation of cultural stereotypes, reinforcing existing** social **inequalities and divisions**. They are not only informational, but also affective and relational. This is a fundamental effect and a fundamental conclusion that resounds from all the studies cited above. Users from cultures underrepresented in training data and technical solutions – especially those from the Global South – may experience symbolic exclusion or even 'de-culturisation' when their values, languages and experiences are ignored, shallowed or distorted, and questions or narratives are interpreted and transformed in the spirit of Anglo-American communication norms and value systems.

Second, experiencing cultural bias among users of AI models may **undermine trust in these technologies in sensitive contexts** – such as education, medicine or justice – where objectivity and equality of access to knowledge are particularly important. If AI models only reinforce one set of norms and values, users from other cultural backgrounds may distrust the recommendations they receive, resulting in digital exclusion and digital inequalities (cf. Jenks, 2024)

A final, third effect of the presence of cultural bias in today's AI models concerns **the systemic dimension**. As the practice of increasing use of this technology in the decision-making processes of public institutions (e.g. public services or medical diagnostics) and private institutions (e.g. recruitment processes or credit risk assessment) grows, and the concomitant lack of mechanisms to control and counter cultural bias, the risk to reproduce and reinforce historical inequalities and social injustices increases

In light of such social consequences, legislative, ethical and design measures to counter cultural bias in AI models undoubtedly seem necessary. The aim should be to redesign these tools so that they are more inclusive and culturally sensitive. Indeed, cultural bias in AI is not just a technological problem – it is a phenomenon with social and political dimensions that requires ongoing monitoring, critical analysis and participatory models of technology governance over AI by public institutions and independent ethical bodies to assess models for compliance with equality and pluralist values. In the research review cited above, their authors advocate a range of solutions. These include the need for ongoing monitoring and examination of the cultural biases of language models (Messner et al., 2023; Rauhali and Xin, 2024; Tao et al., 2024); cultural audits of AI models (Choudhary, 2024; Hagendorff, 2024); using appropriate strategies to control and work with AI chatbots, such as cultural prompting or cultural fine-tuning (Abrar et al., 2025; Choudhary, 2024; Tao et al., 2024; Masoud et al., 2025); the creation of multilingual and multicultural data corpora (Karpouzis, 2024; Liu, 2023; Yuan, Hu, Zhang, 2024; Masoud et al., 2025); greater model transparency (Jenks, 2024; Liu, 2023); the presence of diversity experts on project teams (Karpouzis, 2024) or dialogue between technology developers, users and social institutions (Gosh et al., 2024; Karpouzis, 2024; Liu, 2023).

The literature also calls for sustained oversight of AI by public institutions and independent ethical bodies to assess models for compliance with equality and pluralist values (Binns et al., 2018; Jobin et al., 2019). Lewis (2025) points to the need to integrate local learning communities and multilingual learning materials into AI training processes, especially in educational tools. The author emphasises the need for cross-sector collaboration and the involvement of teaching communities in content validation. Mushkani and colleagues (2025), on the other hand, develop the idea of a 'right to AI', indicating that access to equal, culturally appropriate and transparent AI should be considered a new human right. They call for international standards of representation and participatory models of AI auditing with cultural and social minorities.

CONCLUSIONS

The conducted analysis of empirical research on cultural bias in generative AI chatbots reveals the complex nature of this phenomenon and its significant social consequences. The research clearly shows that it reflects and reinforces dominant Western cultural patterns, values and social norms, while marginalising the perspectives of other cultures, particularly those outside the English-speaking and Western world. Four main areas of prejudice were identified: axiological/civilisational, racial/ethnic, nationality and religious/worldview. In all these aspects, the models studied tended to simplify, stereotype and marginalise the perspectives of non-dominant social and cultural groups.

The implications of these biases are far-reaching and concern both individual user experiences and wider social processes, such as reinforcing inequalities, marginalising specific groups and undermining trust in AI technologies. The implications are particularly significant in socially sensitive fields such as education, health care or justice, where the use of culturally biased models can lead to serious inequalities and injustices.

The analysis also showed that cultural biases are not just a technical problem of algorithms, but a deep-rooted social phenomenon resulting from the contexts of training data and the design decisions made by technology developers. Effectively addressing these biases requires a comprehensive approach that includes increasing the transparency of training data, developing more inclusive design methodologies (e.g. cultural prompting, cultural fine-tuning, multilingual training corpora), implementing cultural audits, and creating mechanisms for sustained oversight of AI by independent social and ethical institutions.

Further research should focus on developing methodologies to counter cultural bias and monitoring the effectiveness of implemented solutions. Interdisciplinary collaboration and the involvement of local communities that can contribute key knowledge regarding the specific cultural contexts of AI use are also essential. The phenomenon of cultural bias in AI should be treated as an important area of social responsibility, requiring conscious management and continuous critical reflection.

BIBLIOGRAPHY

Abrar A., Oeshy N. T., Kabir M., and Ananiadou S. (2025), *Religious Bias Landscape in Language and Text-To-Image Models: Analysis, Detection, and Debiasing Strategies*, arXiv preprint arXiv:2501.08441. https://doi.org/10.48550/arXiv.2501.08441

Barroso da Silveira J., and Lima E. A. (2024), *Racial biases in AIs and Gemini's Inability to Write Narratives About Black People*, Emerging Media 2, no. 2, pp. 277–287. https://doi.org/10.1177/27523543241277564

Binns R., Van Kleek M., Veale M., Lyngs U., Zhao J., and Shadbolt N. (2018), *'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions*, CHI 2018, no. 377, pp. 1–14. https://doi.org/10.1145/3173574.3173951

Bojic L. (2024), *AI alignment: Assessing the Global Impact of Recommender Systems*, Futures 160, June, 103383. https://doi.org/10.1016/j.futures.2024.103383

Brzezinski D., Filipek K., Piwowar K., and Winiarska-Brodowska M. (2024), *Algorithms, Artificial Intelligence and Beyond: Theorising Society and Culture of the 21st Century*, New York: Routledge.

Choudhary T. (2024), *Reducing Racial and Ethnic Bias in AI Models: A Comparative Analysis of ChatGPT and Google Bard*, Preprints. https://doi.org/10.20944/preprints202406.2016.v1

Faustino D. and Lippold W. (2023), *Colonialismo Digital: Por Uma Crítica Hacker-Fanoniana*, Sao Paulo: Boitempo.

Friedman A.B. (2024), *The Era of ChatGPT: Recommendations for the Integration of LLMs in Gerontology*, Innovation in Aging 8, Supplement_1, p. 586. https://doi.org/10.1093/geroni/igae098.1920

Ghosh S., Venkit P. N., Gautam S., Wilson S., and Caliskan A. (2024), *Do Generative AI Models Output Harm While Representing Non-Western Cultures*, Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society 7, no. 1, pp. 476-489. https://doi.org/10.1609/aies.v7i1.31651

Goswami A. (2024), *Recommendation System as a Social Determinant of Health*, Digital Society 3, no. 28. https://doi.org/10.1007/s44206-024-00118-x

Hau M. F. and Hendriksen Ch. (2024), *Beyond Bias: Studying 'Culture' in LLMs and AI Chatbots*, SciSpace.com. https://scispace.com/papers/beyond-bias-studying-culture-in-llms-and-ai-chatbots-6pimgex90zjk.

Hofstede G.H. (2001), *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations* (2nd ed.), Thousand Oaks, CA: Sage Publications.

House R.J., Hanges P.J., Javidan M., Dorfman P.W., and Gupta V. (Eds.) (2004), *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*, Thousand Oaks: Sage Publications.

Inglehart R. and Welzel C. (2005), *Modernization, Cultural Change, and Democracy: The Human Development Sequence*, Cambridge: Cambridge University Press.

Jenks C.J. (2024), *Communicating the Cultural Other: Trust and Bias in Generative AI and Large Language Models*, Applied Linguistics Review 16, no. 2, pp. 787-795. https://doi.org/10.1515/applirev-2024-0196

Jobin A., Ienca M., and Vayena E. (2019), *The Global Landscape of AI Ethics Guidelines*, Nature Machine Intelligence 1, pp. 389-399. https://doi.org/10.1038/s42256-019-0088-2

Karpouzis K. (2024), *Plato's Shadows in the Digital Cave: Controlling Cultural Bias in Generative AI*, Electronics 13, no. 8, 1457. https://doi.org/10.3390/electronics13081457

Khatun A. (2024), *Media, Propaganda, and the Othering Process of the Rohingyas*, [in:] K. Ahmed and M.R. Islam (Eds.), *Understanding the Rohingya displacement: International Perspectives on Migration*, Singapore: Springer, pp. 169-199.

Lewis A. A. (2025), *Unpacking Cultural Bias in AI Language Learning Tools: An Analysis of Impacts and Strategies for Inclusion in Diverse Educational Settings*, International Journal of Research

and Innovation in Social Science 9, no. 1, pp. 1878-1892. https://dx.doi.org/10.47772/IJRISS.2025.9010151

Liu Z. (2023), *Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies*, Journal of Transcultural Communication 3, no. 2, pp. 224-244. https://doi.org/10.1515/jtc-2023-0019

Masoud R., Liu Z., Ferianc M., Treleaven P.C., and Rodrigues M.R. (2025), *Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions*, Proceedings of the 31st International Conference on Computational Linguistics, pp. 8474–8503, Abu Dhabi: Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.567/

Messner W., Greene T., and Matalone J. (2023), *From Bytes to Biases: Investigating the Cultural Self-Perception of Large Language Models*, Journal of Public Policy & Marketing 44, no. 3, pp. 370-391. https://doi.org/10.1177/07439156251319788

Mushkani R., Berard H., Cohen A., and Koeski S. (2025), *The Right to AI*, arXiv preprint arXiv:2501.17899. https://doi.org/10.48550/arXiv.2501.17899

Mustafa Z. U., Amir M., Mustafa M., and Anwar M. A. (2025), *Harmony Amidst Division: Leveraging Genetic Algorithms to Counteract Polarisation in Online Platforms*, International Journal of Computational Science and Engineering 28, no. 7, pp. 1-17. https://doi.org/10.1504/IJCSE.2025.143956

Noble S.U. (2018), *Algorithms of oppression: How Search Engines Reinforce Racism*, New York: New York University Press.

Rafikova A. and Voronin A. (2025), *Human–Chatbot Communication: a Systematic Review of Psychologic Studies*, AI & Society 40, pp. 5389-5408. https://doi.org/10.1007/s00146-025-02277-y

Rauhala J. and Xin T. (2024), *What Culture is Chat GPT's AI?*, [in:] M. Lehto and M. Karjalainen (Eds.), *Proceedings of the 23rd European Conference on Cyber Warfare and Security* 23, no. 1, pp. 812-815. https://doi.org/10.34190/eccws.23.1.2364

Saumure R., De Freitas J., and Puntoni S. (2025), *Humor as a Window into Generative AI bias*, Scientific Reports 15, 1326. https://doi.org/10.1038/s41598-024-83384-6

Si Y., Jiang C., Wei X., Fang S., Li Y., and Hu Y. (2024), *Analysis of the Correlation of Topic Feature Changes Based on the LDA Model*, Theoretical and Natural Science 53, pp. 73-82. https://doi.org/10.54254/2753-8818/53/20240221

Tao Y., Viberg O., Baker R.S., and Kizilcec R.F. (2024), *Cultural Bias and Cultural Alignment of Large Language Models*, PNAS Nexus 3, no. 9, p. 346. https://doi.org/10.1093/pnasnexus/pgae346

Tsuria R. and Tsuria Y. (2024), *Artificial Intelligence's Understanding of Religion: Investigating the Moralistic Approaches Presented by Generative Artificial Intelligence Tools*, Religions 15, no. 3, p. 375. https://doi.org/10.3390/rel15030375

Vicsek L., Vansco A., Zajko M., and Takacs J. (2024), *Exploring LGBTQ+ Bias in Generative AI Answers Across Different Country and Religious Contexts*, arXiv. https://doi.org/10.48550/arxiv.2407.03473

Yuan H., Che Z., Li S., Zhang Y., Hu X., and Luo S. (2024), *The High Dimensional Psychological Profile and Cultural Bias of ChatGPT*, arXiv. https://doi.org/10.48550/arxiv.2405.03387

Yuan X., Hu J., and Zhang Q. (2024), *A Comparative Analysis of Cultural Alignment in Large Language Models in Bilingual Contexts*, OSF. https://doi.org/10.31219/osf.io/6hpcf

Zhu S., Wang W., and Liu Y. (2024), *Quite Good, But Not Enough: Nationality Bias in Large Language Models*, arXiv. https://doi.org/10.48550/arxiv.2405.06996

UKRYTE ALGORYTMY KULTURY: PRZEGLĄD I KRYTYCZNA ANALIZA UPRZEDZEŃ KULTUROWYCH W CHATBOTACH GENERATYWNYCH AI OGÓLNEGO PRZEZNACZENIA

S t r e s z c z e n i e

Celem artykułu jest przegląd i systematyzacja wyników najnowszych badań empirycznych dotyczących przejawów uprzedzeń kulturowych w treściach wytwarzanych przez generatywne chatboty AI ogólnego zastosowania, takich jak ChatGPT, Copilot, Gemini, Claude czy DeepSeek oraz identyfikacja ich potencjalnych konsekwencji społecznych. Sformułowano następujące pytania badawcze: jakie rodzaje i jaka jest skala uprzedzeń kulturowych występują w generatywnych chatbotach AI? jakie są społeczne konsekwencje ich występowania oraz możliwe sposoby i kierunki przeciwdziałania? Badanie przeglądowe oparto na krytycznej analizie 17 najnowszych badań empirycznych opublikowanych w latach 2024–2025. Przeprowadzona analiza ukazuje złożony charakter obecności i konsekwencji cultural bias w obecnych modelach AI. Jednoznacznie wykazano, że odzwierciedlają i wzmacniają one zachodnie wzorce kulturowe. Wyróżniono cztery odmiany *cultural bias*: aksjologiczno-cywilizacyjną, rasowo-etniczną, narodowościową oraz religijno-światopoglądową. Analiza wykazała także, że uprzedzenia kulturowe nie są jedynie technicznym problemem algorytmów, ale głęboko zakorzenionym zjawiskiem społecznym, wynikającym z kontekstów danych treningowych oraz projektowych decyzji podejmowanych przez twórców technologii.

Słowa kluczowe: uprzedzenia kulturowe; stereotypy; sztuczna inteligencja; chatboty AI; LLM