Open Forum

Mykola Makhortykh*, Victoria Vziatysheva and Maryna Sydorova

Generative AI and Contestation and Instrumentalization of Memory About the Holocaust in Ukraine

https://doi.org/10.1515/eehs-2023-0054
Received November 13, 2023; accepted November 13, 2023; published online November 27, 2023

The recent rise of generative artificial intelligence (AI) can lead to fundamental transformations in the field of Holocaust remembrance. Earlier non-generative forms of AI, which focused on identifying and retrieving historical information via search engines and content recommenders systems[1] have already caused profound changes in how individuals learn about the Holocaust. By deciding which information sources and individual content items to prioritise to their users (Makhortykh et al. 2023a), these systems shape the visibility of individual Holocaust heritage institutions in digital environments and influence how successful or unsuccessful the online educational and commemorative campaigns are. At the same time, occasional evidence of non-generative AI systems amplifying the visibility of antisemitic and denialist content[2] also raises concerns about the possibility of these systems undermining the institutional efforts to counter Holocaust denialism and hate

[1] In addition to non-generative AI systems used by commercial platforms such as Google (Makhortykh, Urman, and Ulloa 2021) or TikTok (Divon and Ebbrecht-Hartmann 2022), these platforms were also adopted by Holocaust heritage institutions. Some examples include “Let Them Speak” project (https://lts.fortunoff.library.yale.edu/anthology) or “Dimensions in Testimony” (https://iwitness.usc.edu/dit/pinchas).

[2] For some examples, see the case of Google prioritising the images of ovens in response to the search queries inquiring about Jewish baby strollers (Keyser 2020) or DuckDuckGo and Yahoo search engines highlighting antisemitic memes in response to general queries about the Holocaust in Russian (Makhortykh, Urman, and Ulloa 2021).

***Corresponding author: Mykola Makhortykh**, Institute of Communication and Media Studies, University of Bern, Bern, 3012, Switzerland, E-mail: mykola.makhortykh@unibe.ch. https://orcid.org/0000-0001-7143-5317
Victoria Vziatysheva and Maryna Sydorova, Institute of Communication and Media Studies, University of Bern, Bern, Switzerland

speech, in particular in the light of the low levels of Holocaust knowledge around the world (e.g. Claims Conference 2020, 2021).

In contrast to non-generative AI systems, foundation models, such as GPT or Stable Diffusion, allow their users not only to find and retrieve relevant content but actually generate it from scratch. It enables new possibilities for Holocaust remembrance, including a representation of the past from a multitude of perspectives and multiple scales, as well as a broader access to lesser known aspects to it (Kansteiner 2022), in addition to new possibilities for detecting distortion and denial of historical facts (Makhortykh et al. 2023b). However, there are also growing concerns about generative AI contributing to the distorted representation of the Holocaust due to non-intentional hallucinations and fabrications of the AI models (Alkaissi and McFarlane 2023) or intentional jailbreaking by malicious actors (Tucker 2023). These concerns are amplified by the difficulties in differentiating between human- and AI-made content (Susnjak 2022) that can facilitate the production of non-authentic historical content (e.g. fake Holocaust testimonies) that can not only mislead individual users of AI systems but also contribute to the manipulation of the public opinion, in particular, if such content is used for instrumentalising the past.

The concerns are particularly pronounced in the case of Eastern European countries, where Holocaust memory is often instrumentalised by a diverse range of political actors. Many of these countries constitute fractured memory regimes (Kubik and Bernhard 2014, 17), which are distinguished by the tendency of their political elites to monopolise public memory practices which are then used to legitimise elites' decisions, including potentially unpopular moves. Often, such instrumental uses of the past are accompanied by the use of online media for reinforcing dominant representations of the past and suppressing alternative interpretations to eradicate memorial dissent. While most existing studies looking at these "web wars" (Rutten et al. 2013) focus on human users and their efforts to contribute to the instrumentalisation of the past or counter it (e.g. Barna and Knap 2023; Gaufman 2015; Kalinina and Menke 2016; Khlevnyuk 2019; Kulyk 2016), a few recent studies also highlight the growing role of AI systems in this context (Makhortykh, Urman, and Ulloa 2022; Zavadski and Toepfl 2019).

Among many instances of instrumentalisation of Holocaust memory in the region, the use of it by the Kremlin in the context of the ongoing aggression against Ukraine stands out. Since the beginning of the aggression in 2014, the pro-regime Russian institutions and activists extensively relied on Second World War references for mobilising the Russian public and constructing negative identities of Ukrainians resisting the aggression (Gaufman 2015; Makhortykh 2018; Pakhomenko et al. 2018). The instrumental use of war memories, including the ones related to the Holocaust, has increased following the large-scale invasion in 2022 (Ferraro 2023; Shevtsova

2022) with the primary aim of demonising Ukrainians among Russian and Western audiences. Often, such instrumentalisation involves the distortion of historical facts related to the Holocaust in Ukraine: a telling example is the speech of Vladimir Putin in September 2023, in which the Russian president claimed that Ukrainian collaborators were the main perpetrators of the Holocaust in Ukraine and performed cruelties of which SS soldiers were not capable of, which resulted in the death of a quarter of all Holocaust victims (Meduza 2023).

Despite the intense instrumentalisation and distortion of Holocaust memory by the Kremlin, there is limited understanding of how it can be amplified or countered by generative AI. Recent studies (e.g. Urman and Makhortykh 2023) provide evidence that in some cases, outputs of generative AI-powered conversational agents, such as Google Bard, can be systematically skewed in favour of Putin's regime, including the case of the prompts dealing with the Russian aggression against Ukraine. However, it is not clear to what degree these observations are applicable to Holocaust memory and its instrumental distortion by the Kremlin. In order to address this gap, we audited the performance of two major AI-powered conversational agents – ChatGPT and Google Bard – regarding information on the Holocaust in Ukraine.

For this aim, we used a selection of 74 statements translated into English, Russian, and Ukrainian languages; the statements were inquiring information about the aspects of the Holocaust, which are commonly instrumentalised and often distorted by pro-Kremlin propaganda. Examples of such aspects include the attribution of blame for the Holocaust and its individual episodes to Ukrainians in general, and to specific Ukrainian wartime groups (e.g. the Organization of the Ukrainian Nationalists or the *Nachtigall* battalion) and the details regarding specific instances of the Holocaust (e.g. the number of victims and the ways of killing). We also included a few prompts dealing with the mass atrocities targeting non-Jewish communities in Ukraine, in particular ethnic Ukrainians and Poles, for comparative purposes.

To evaluate the performance of the two conversational agents, we compared the responses provided by them to the predefined baseline (i.e., what we would perceive as the correct answer to the question based on existing research regarding the Holocaust in Ukraine). The baseline for the prompts varied from simple "yes" or "no" answers to the range of possible options (e.g. in the case of the debated number of victims for specific instances of the Holocaust). For some questions, where there is no consensus among historians, we identified as a baseline an answer mentioning that the question is unclear or debated.

While we still need to finalise the analysis and ensure the agreement between the coders evaluating the outputs of conversational agents, our preliminary results indicate several troubling findings regarding the performance of agents in relation to information about the Holocaust in Ukraine. Overall, our analysis shows a relatively low accuracy of the agent outputs, with the proportion of correct answers varying between 30 % and 55 % depending on the conversational agent and language. In addition, we considered about 18–30 % of the outputs to be partially correct. We found that Google Bard was approximately 1.5 times more accurate than ChatGPT in English and Ukrainian, but not in Russian. The latter can be explained by a striking amount of non-responses for Bard in Russian (over 30 % of cases), while in the other languages, this number was almost negligible (around 1 %).

The proportion of outputs considered inaccurate varied between 20 % and 45 %, with the largest number of inaccurate responses produced by ChatGPT in Ukrainian.[3] In a number of cases, these inaccuracies were related to Ukrainians and Ukrainian organisations being incorrectly presented as Holocaust perpetrators. In certain cases, the attribution of blame was rather absurd: for instance, conversational agents claimed that the mass murder of Ukrainians in Koryukivka in 1943 was conducted by Ukrainian nationalistic groups despite it being the sole responsibility of the Hungarian occupation forces allegedly assisted by Russian *Hiwis*. Another example relates to outputs (e.g. of ChatGPT in Russian) to prompts regarding the *Nachtigall* battalion, which persistently claimed that the battalion did not exist during the Second World War, and was instead formed in 2014 and participated in the Russian-Ukrainian war.

In addition to the fundamentally incorrect responses (i.e. the ones directly contradicting the baseline established by human experts), we observed multiple instances when conversational agents distorted historical details regarding the Holocaust. In some cases, such distortions were related to factual errors such as incorrect naming of the location of the atrocity (e.g. a Ukrainian village, Koryukivka, was referred to as a location in the Kursk region of Russia), or the way Jewish victims were killed in the course of a specific atrocity (often, such details were based on a few well-known instances of the Holocaust, such as Babyn Yar massacres, and then were attributed to other Holocaust episodes by the agents). However, in other cases, agents went as far as inventing historical personalities (e.g. non-existing Ukrainian and German perpetrators who assumingly committed atrocious actions) or even quotes from eyewitness testimonies hallucinated by the models powering the agents. In a number of instances, these testimonies included fake claims about the involvement of Ukrainians in the Holocaust, thus aligning with the distortion of the historical facts by the Kremlin.

This small-scale investigation of the performance of generative AI-powered conversational agents has several implications. The first is the rather urgent problem

[3] It is important to note that the accuracy of the agents in this study was evaluated primarily on the basis of the general consistency of the output with the baseline. We have not yet systematically examined the entirety of the output for factual accuracy, which means that even responses that were considered accurate may have had some factual errors while meeting the general baseline.

of the poor quality of information generated by AI models regarding the Holocaust in Ukraine. While the distortion of historical facts does not always align exclusively with the Kremlin's interest and, in some cases, mistakenly denies the involvement of Ukrainian perpetrators in the Holocaust, it also often amplifies the Russian regime's efforts to use the past to demonise Ukrainians. Even in those cases, when distortion of the historical facts does not directly serve the instrumental purposes of external actors, it amplifies the epistemic uncertainty regarding the Holocaust in Ukraine, and can facilitate the manipulation of facts in the long term.

The second implication relates to the data the agents are trained on and their ability to assess their own limitations. The large number of inaccurate responses can be explained by the limited access of conversational agents to information about certain episodes of the Holocaust, as well as their inability to correctly retrieve information in relation to more niche Holocaust-related topics. Under these circumstances, agents often produce inaccurate descriptions of events that may include details mistakenly drawn from other Holocaust-related events or even invented by the agent. Instead of pointing out their limited knowledge of the subjects, about which they are prompted, conversational agents in most cases either refuse to respond without providing a clear reasoning for it or, which is more concerning, still provide an inaccurate answer with a large degree of confidence.

The third implication concerns the importance of following the earlier calls (e.g. Walden 2023) to conceptualise the role of digital technology in general and generative AI in particular for the future of Holocaust memory both in Eastern Europe and worldwide. The inability of some of the most used AI-powered conversational agents to provide accurate information about the historical mass atrocities, including the well-documented instances of the Holocaust, is concerning. Not only can it enable new possibilities for distorting historical facts and manipulating the public opinion, but also it can challenge the core mission of Holocaust heritage institutions aiming to preserve memories of the victims and prevent instrumental uses of the past. In addition to exploring possibilities for improving the performance of generative AI models on the level of their design (e.g. by improving the training data regarding information about the Holocaust and other genocides, and optimising guardrails to minimise the likelihood of the model to invent historical facts), it is crucial to increase awareness about the possible risks of AI, and create possibilities for the general public and heritage practitioners to develop AI literacies required to address these risks.

Research funding: This work was supported by Alfred Landecker Foundation's project "Algorithmic turn in Holocaust memory transmission"

References

Alkaissi, Hussam, and Samy I. McFarlane. 2023. “Artificial Hallucinations in ChatGPT: Implications in Scientific Writing.” *Cureus* 15 (2): 1–4.

Barna, Ildikó, and Árpád Knap. 2023. “Analysis of the Thematic Structure and Discursive Framing in Articles about Trianon and the Holocaust in the Online Hungarian Press Using LDA Topic Modelling.” *Nationalities Papers* 51 (3): 603–21.

Claims Conference. 2020. “First-ever 50-state survey on holocaust knowledge of American millennials and Gen Z Reveals Shocking Results.” In *Claims Conference*. https://www.claimscon.org/millennial-study/ (accessed November 10, 2023).

Claims Conference. 2021. “New Study Reveals U.K. Respondents Believe Two Million or Fewer Jews Were Killed in the Holocaust.” In *Claims Conference*. https://www.claimscon.org/uk-study/ (accessed November 10, 2023).

Divon, Tom and Tobias Ebbrecht-Hartmann. 2022. “#JewishTikTok: The JewToks’ Fight against Antisemitism.” In *TikTok Cultures in the United States*, edited by Trevor Boffone, 47–58. London: Routledge.

Ferraro, Vicente. 2023. “The Contradictions in Vladimir Putin’s “Just War” against Ukraine: The Myths of NATO’s Containment, Minority Protection and Denazification.” *SciELO*. https://preprints.scielo.org/index.php/scielo/preprint/download/5486/contradictions-in-putin-arguments-war-in-ukraine-nato-enlargemen (accessed November 10, 2023).

Gaufman, Elizaveta. 2015. “World War II 2.0: Digital Memory of Fascism in Russia in the Aftermath of Euromaidan in Ukraine.” *Journal of Regional Security* 10 (1): 17–35.

Kalinina, Ekaterina, and Manuel Menke. 2016. “Negotiating the Past in Hyperconnected Memory Cultures: Post-Soviet Nostalgia and National Identity in Russian Online Communities.” *International Journal of Media and Cultural Politics* 12 (1): 59–74.

Kansteiner, Wulf. 2022. “Digital Doping for Historians: Can History, Memory, and Historical Theory Be Rendered Artificially Intelligent?” *History and Theory* 61 (4): 119–33.

Keyser, Zachary. 2020. “Google Responds after Search Term Yields Antisemitic Allusion to Holocaust.” In *The Jerusalem Post*. https://www.jpost.com/diaspora/antisemitism/google-responds-after-search-term-yields-antisemitic-allusion-to-holocaust-643811 (accessed November 10, 2023).

Khlevnyuk, Daria. 2019. “Narrowcasting Collective Memory Online: “Liking” Stalin in Russian Social Media.” *Media, Culture & Society* 41 (3): 317–31.

Kubik, Jan and Michael Bernhard. 2014. “A Theory of the Politics of Memory.” In *Twenty Years After Communism: The Politics of Memory and Commemoration*, edited by Michael Bernhard, and Jan Kubik, 7–37. Oxford: Oxford University Press.

Kulyk, Volodymyr. 2016. “Negotiating Memory in Online Social Networks: Ukrainian and Ukrainian-Russian Discussions of Soviet Rule and Anti-soviet Resistance.” In *Disputed Memory: Emotions and Memory Politics in Central, Eastern and South-Eastern Europe*, edited by Tea Sindbæk Andersen, and Barbara Törnquist-Plewa, 273–98. Berlin: De Gruyter.

Makhortykh, Mykola. 2018. “#NoKievNazi: Social Media, Historical Memory and Securitization in the Ukraine Crisis.” In *Memory and Securitization in Contemporary Europe*, edited by Victor Apryshchenko, and Vlad Strukov, 219–47. London: Palgrave Macmillan.

Makhortykh, Mykola, Aleksandra Urman, and Roberto Ulloa. 2021. “Hey, Google, Is it what the Holocaust Looked like? Auditing Algorithmic Curation of Visual Historical Content on Web Search Engines.” *First Monday* 26 (10): 1–24.

Makhortykh, Mykola, Aleksandra Urman, and Roberto Ulloa. 2022. “Memory, Counter-memory and Denialism: How Search Engines Circulate Information about the Holodomor-Related Memory Wars.” *Memory Studies* 15 (6): 1330–45.

Makhortykh, Mykola, Aleksandra Urman, Roberto Ulloa, and Juhi Kulshrestha. 2023a. “Can an Algorithm Remember the Holocaust? Comparative Algorithm Audit of Holocaust-Related Information on Search Engines.” In *Digital Memory: Neue Perspektiven für die Erinnerungsarbei*t, edited by Iris Groschek and Habbo Knoch, 79–93. Göttingen: Wallstein Verlag.

Makhortykh, Mykola, Eve M. Zucker, David J. Simon, Daniel Bultmann, and Roberto Ulloa. 2023b. “Shall Androids Dream of Genocides? How Generative AI Can Change the Future of Memorialization of Mass Atrocities.” *Discover Artificial Intelligence* 3 (1): 1–17.

Meduza. 2023. ““Even the SS troops didn’t consider it possible” Putin says “local nationalists and anti-Semites” killed 1.5 million Jews in Ukraine during WWII.” Meduza. https://meduza.io/en/feature/2023/09/05/even-the-ss-troops-didn-t-consider-it-possible (accessed November 10, 2023).

Pakhomenko, Sergii, Kateryna Tryma, and J’moul A. Francis. 2018. “The Russian–Ukrainian War in Donbas: Historical Memory as an Instrument of Information Warfare.” In *The Use of Force against Ukraine and International Law: Jus Ad Bellum, Jus In Bello, Jus Post Bellum*, edited by Sergey Sayapin, and Evhen Tsybulenko, 297–312. Berlin: Springer.

Rutten, Ellen, Julie Fedor and Zvereva Vera, eds. 2013. *Memory, Conflict and New Media: Web Wars in Post-Socialist States*. London: Routledge.

Shevtsova, Maryna. 2022. “Looking for Stepan Bandera: The Myth of Ukrainian Nationalism and the Russian “Special Operation”.” *Central European Journal of International and Security Studies* 16 (3): 132–50.

Susnjak, Teo. 2022. “ChatGPT: The End of Online Exam Integrity?” *arXiv*: 1–21. https://doi.org/10.48550/arXiv.2212.09292.

Tucker, Joshua A. 2023. “AI Could Create a Disinformation Nightmare in the 2023 Election.” *The Hill*. https://thehill.com/opinion/4096006-ai-could-create-a-disinformation-nightmare-in-the-2024-election/ (accessed November 10, 2023).

Urman, Aleksandra, and Mykola Makhortykh. 2023. “The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat.” *OSF* 1–11, https://doi.org/10.31219/osf.io/q9v8f.

Walden, Victoria Grace. 2023. “Is Digitalization a Blessing or a Curse for Holocaust Memorialization?” *Eastern European Holocaust Studies* 1 (1): 17–22.

Zavadski, Andrei, and Florian Toepfl. 2019. “Querying the Internet as a Mnemonic Practice: How Search Engines Mediate Four Types of Past Events in Russia.” *Media, Culture & Society* 41 (1): 21–37.