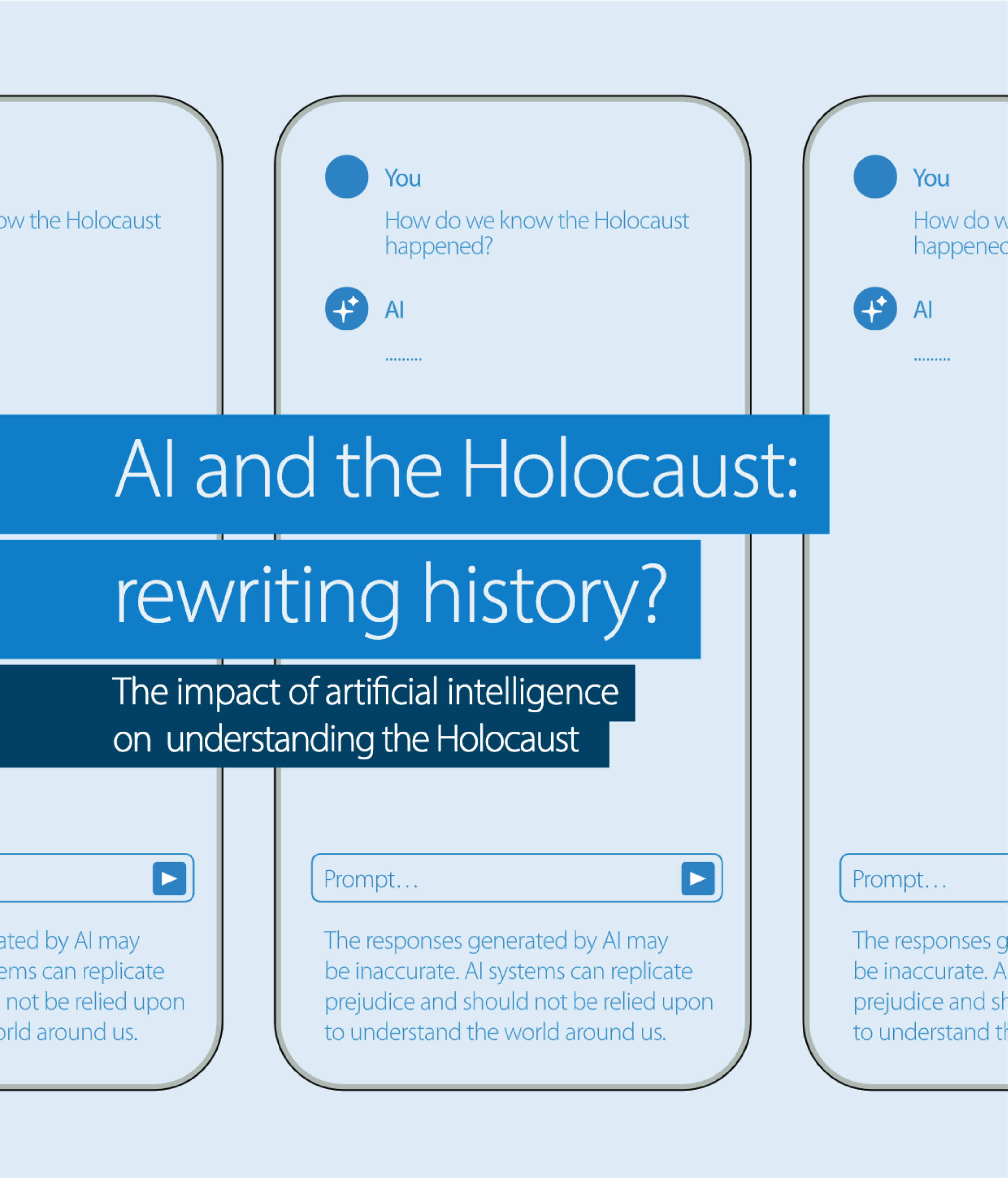
unesco



You

How do we know the Holocaust happened?

AI

....

Prompt…

The responses generated by AI may be inaccurate. AI systems can replicate prejudice and should not be relied upon to understand the world around us.

AI and the Holocaust: rewriting history?

The impact of artificial intelligence on understanding the Holocaust

Education 2030

UNESCO – a global leader in education

Education is UNESCO's top priority because it is a basic human right and the foundation for peace and sustainable development. UNESCO is the United Nations' specialized agency for education, providing global and regional leadership to drive progress, strengthening the resilience and capacity of national systems to serve all learners. UNESCO also leads efforts to respond to contemporary global challenges through transformative learning, with special focus on gender equality and Africa across all actions.

The Global Education 2030 Agenda

UNESCO, as the United Nations' specialized agency for education, is entrusted to lead and coordinate the Education 2030 Agenda, which is part of a global movement to eradicate poverty through 17 Sustainable Development Goals by 2030. Education, essential to achieve all of these goals, has its own dedicated Goal 4, which aims to ***"ensure inclusive and equitable quality education and promote lifelong learning opportunities for all."*** The Education 2030 Framework for Action provides guidance for the implementation of this ambitious goal and commitments.

unesco
United Nations
Educational, Scientific
and Cultural Organization

Education 2030

This paper was commissioned by the UNESCO Section of Global Citizenship and Peace Education and the World Jewish Congress.
It was drafted by Mykola Makhortykh, and coordinated by Heather Mann under the supervision of Karel Fracapane and Cecilia Barbieri.
UNESCO is grateful to Claire Leibowicz for her review of the paper, and the feedback provided by Yfat-Barak Cheney.

Published in 2024 by the United Nations Educational, Scientific and Cultural Organization,
7, place de Fontenoy,
75352 Paris 07 SP,
France

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.
The ideas and opinions expressed in this publication are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.

Cover design: Studio Hortenzia
Inside icons: Anna Mortreux
Graphic design and layout: UNESCO/Anna Mortreux
Editing: Jay Sivell

Printed by UNESCO

Printed in France

Executive summary

AI and the Holocaust – rewriting history? Managing the impact of artificial intelligence on understanding of the Holocaust

The threats associated with AI on safeguarding the record of the Holocaust are manifold, including the potential for manipulation by malicious actors, the introduction of falsehoods or dissemination of biased information, and the gradual erosion of public trust in authentic records. This paper provides a warning of what is at stake for the preservation of historical truth in a digital era increasingly mediated by AI. This report highlights five major concerns:

1. AI automated content may invent facts about the Holocaust.

AI models have produced misleading or false narratives about the Holocaust. Data voids and biases have led to "hallucinations" in generative AI systems, producing incorrect or invented content that never occurred. Without AI literacy and research skills, users may not know how to verify AI-produced texts, or recognize the unreliability of the data.

2. Falsifying historical evidence: Deepfake Technology

Deepfake technology has potential to manipulate audio and video to fabricate Holocaust-related content. There is a need for mechanisms to prevent the misuse of AI in purposefully creating fake "evidence" that undermines the veracity of the established historical record of the Holocaust and spreads hate speech. Deepfakes of celebrities have been used to spread Nazi ideology, or to simulate conversations with Nazi leaders including Adolf Hitler.

3. AI models can be manipulated to spread hate speech

Targeted campaigns by violent extremist online groups can exploit AI flaws to promote hate speech and antisemitic content about the Holocaust. Chatbots and search engines have been hacked or manipulated by bad actors to spread Nazi ideology.

4. Algorithmic bias can spread Holocaust denial

Biased data sets have led to some search engines and AI chatbots downplaying Holocaust facts or promoting far-right content, including Holocaust denial.

5. Oversimplifying history

AI's tendency to focus on the most well-known aspects of the Holocaust, oversimplifies its complexity. The omission of lesser-known episodes and events in the history of the Holocaust reinforces stereotypical representations of the Holocaust and limits our understanding of a complex past which affected people in every country in Europe and in North Africa, and whose legacy continues to be felt worldwide.

While there are some benefits to be gained from AI technologies in education and research, to navigate these challenges and capitalize on the benefits, it is essential for AI designers, policymakers, educators, and researchers to collaborate closely. Only AI systems equipped with robust safeguards and human rights assessments, coupled with an increased focus on developing digital literacy skills, can uphold the integrity of historical truth and ensure the responsible use of artificial intelligence.

Table of contents

Introduction

Almost 80 years after the end of World War II, people are encountering inaccurate, false and misleading information about the Holocaust[1] and Nazi ideology across the digital sphere. Videos produced by artificial intelligence (AI 'deepfakes') depict Adolf Hitler or celebrities reading from Mein Kampf and be circulated on social media with few clues for the undiscerning viewer to distinguish fake from reality. The victims of the Holocaust are being re-victimized, as survivors are targeted by AI-generated hate speech with the malicious intent of casting doubt on the reality of their lived experiences and memories. Our understanding of this history, based on decades of research and the courage of thousands of Holocaust survivors to record their testimony, is threatened by people with an ideologically antisemitic agenda. Such threats are being accelerated and exacerbated by AI tools for generating, modifying and disseminating content. This paper explores the risks, challenges and opportunities of AI advancements for Holocaust knowledge and understanding, and asks: how should policymakers, online platforms (including social media and AI tool builders) and educators respond to this new reality? How can the facts of the Holocaust be safeguarded?

What is at risk?

Antisemitism, one of the most enduring forms of prejudice, has found new avenues to spread and amplify in the digital, and now the AI, era. AI-generated content transcends borders and is often spread further by AI on social media. In this interconnected information ecosystem, hateful content and disinformation about the Holocaust threaten to spread antisemitism, affecting Jewish communities around the globe and undermining the principles of equality and tolerance. This paper is concerned about the threat to peoples' ability to recognize truth about the history of a genocide, and the impact of fabrications.

In this era of digital revolution, we are witnessing profound transformations in how societies and individuals learn about the past. The ease of producing and accessing digital historical content brings new possibilities for interactive and creative ways to learn about history (Shandler, 2017). AI systems have the capacity to order and manage access to the large volume of information about the past available online, including information about the Holocaust. AI systems used by search engines and social media platforms direct access to content about the Holocaust in response to user queries. When optimized for historical accuracy, this can support the dissemination of knowledge and information about the Holocaust; however, the algorithms used by social media and search engines have been found to prioritize and promote content (including disinformation) that is attention- and engagement-focused, and prone to bias, potentially working against accuracy (UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 2018).

Generative AI is a subtype of AI used for creating content. Lack of awareness or understanding about how Generative AI systems function also creates new blind spots that must be addressed. Digital literacy and user education are valuable interventions, but policymakers and AI developers and implementers must also take actions and make design decisions that safeguard historical information. Users regularly accept AI-generated material without awareness of the data each system is trained on or how the systems function. Lack of transparency about training data and company moderation policies prevents users assessing how each AI product deals with sensitive subjects such as the history of the Holocaust, or manages prejudice in its data or user queries. Generative AI systems may produce inaccurate content; AI-driven content moderation systems may not filter out harmful, inaccurate Holocaust content.

[1] Also known as the Shoah, the systematic murder of 6 million Jews by Nazi Germany, its allies and collaborators.

These risks must be addressed to protect the memory of the Holocaust from abuse, denial and distortion,[2] and to prevent AI from magnifying the rise, reach and impact of antisemitic ideology in the digital sphere.

At the same time, AI can be used to enhance the work of educational and heritage institutions – by generating new ways of interacting with the past. Immersive mixed reality (XR) online experiences with Holocaust survivors and informed, fact-based chatbot discussions allow young people to ask questions about history that interest them in an interactive manner (Manca et al., 2023). AI also frees up new capacity for Holocaust research by allowing scholars to process large sets of archival materials. It can help identify patterns and generate new knowledge.

Governments, AI developers and educators must urgently grapple with the impact of AI on knowledge and understanding, with four in five (80%) young people using AI tools and services multiple times a day for education, entertainment and everyday activities, according to a UN study (UNOICT, 2022). As more young people turn to it as a digital tool to find quick answers to their questions about the Holocaust, educators and policymakers must be attuned to how this may impact our knowledge and understanding of the Holocaust, its consequences for Holocaust remembrance and research, and how to mitigate any erosion of human rights standards and values. Learners are increasingly using Generative AI to complete assignments and find information, meaning that companies and individuals designing these systems must take on the same kind of decision-making processes on content, narrative-shaping and reliability as Holocaust historians and educators. The role of those developing and deploying AI systems and tools must be central to considerations on how to govern the technology in the service of historical preservation and truth.

What is AI?

Artificial intelligence (AI) is a form of information processing technology that is distinguished by the 'capacity to process data and information in a way that resembles intelligent behavior, and typically includes aspects of reasoning, learning, perception, prediction, planning or control' (UNESCO, 2021, p. 10). While sometimes used interchangeably, machine learning (ML) is 'an application of AI in which computers use algorithms (rules) embodied in software to learn from data and adapt with experience' (Donahue, 2018).

Due to its increasing sophistication and capabilities, largely driven by the vast troves of data they are now built from, AI has been adopted in different sectors, from journalism to law enforcement to education to politics. AI tools can support the creative process, improve accessibility, save money and time, but the same tools can also support censorship, exacerbate prejudice, and contribute to the spread of inaccurate information. AI is often considered a dual-use technology: presenting both enormous challenges and opportunities that many seek to balance (Ueno, 2023).

What is Generative AI?

Generative AI (GenAI) is an artificial intelligence technology that automatically generates content in response to prompts written in natural-language conversational interfaces. Rather than simply curating existing webpages, by drawing on existing content, GenAI actually produces new content. The content can appear in formats that comprise all symbolic representations of human thinking: texts written in natural language, images (including photographs, digital paintings and cartoons), videos, music and software code. GenAI is trained using data collected from webpages, social media conversations and other online media. It generates its content by statistically analysing the distributions of words, pixels or other elements in the data that it has ingested and identifying and repeating common patterns (for example, which words typically follow which other words) (UNESCO, 2023d, 8).

[2] Holocaust denial refers specifically to any attempt to claim that the Holocaust/Shoah did not take place. Distortion of the Holocaust refers to intentional efforts to excuse or minimize the impact of the Holocaust or its principal elements, including collaborators and allies of Nazi Germany, for example minimizing the number of victims, obfuscating responsibility for the genocide or celebrating it. The International Holocaust Remembrance Alliance adopted a definition of denial and distortion in 2013. https://holocaustremembrance.com/resources/working-definition-holocaust-denial-distortion
The United Nations General Assembly condemned Holocaust denial and distortion in January 2022. https://documents.un.org/doc/undoc/ltd/n22/230/12/pdf/n2223012.pdf?token=PM8mzABOEzC58NyIH6&fe=true

Ethics and AI systems

Non-generative and generative AI systems

The scope of AI applications has increased hugely in recent decades, supported by innovations in computational power and data accessibility. Initially, the public encountered AI through non-generative systems – now known as traditional or narrow AI.[3] Internet users employ traditional AI every time they use web search engines (e.g. Google or Bing) or other recommender systems (e.g. YouTube recommendations or Facebook newsfeed). Non-generative AI is also behind some systems of facial recognition, fraud detection and spam filtering. It usually works by making separate computations based on a specific set of inputs, from which it can offer a response or decision.

Recent technological advances have produced forms of AI that can generate new content, such as those that do so based on 'natural language' prompts, including OpenAI's ChatGPT. This generative AI can produce content in various formats (e.g. text, image, or video) and is increasingly integrated with non-generative AI systems.

A growing number of AI systems now generate audio and visual content. For example, Midjourney, Adobe Firefly and OpenAI's Dall-E create images; Runway and OpenAI's Sora use AI to create video. Synthetic audio – artificially generated sound – has also emerged as a potentially robust and impactful type of generative content. These systems can be used for entertainment, to produce fun and creative social media content, but they have also been used to spread misleading content, including malicious deepfakes that may manipulate public opinion on high stakes facts.

Ethical challenges of using AI

The increased accessibility and sophistication of AI systems presents a number of ethical challenges. The list of possibilities is extensive, so, below, we summarize a few that are of particular relevance for Holocaust education and remembrance.

- GenAI outputs may appear authoritative and accurate, but AI systems lack human-level semantic understanding. An AI-generated text can look impressively human, as if the system understood the text it produced. However, generative AI does not comprehend information as a human does. It is a tool that identifies patterns in data and strings words together in ways that may not adequately or accurately capture the right answer to a particular question. AI models have a limited ability to understand the meaning of content they retrieve or generate. These limitations are of particular concern in the case of the Holocaust due to the complexity of the history, the need to distinguish between Holocaust testimony and Holocaust denial content, and the moral and ethical obligation to preserve an accurate historical record. Generative AI risks spreading distorted and offensive representations of the Holocaust that impede public understanding of the history, and make viewers and users overconfident about responses that appear more conversational and human-like than more typical search queries.

- AI systems can inherit human biases and amplify them. To learn what content to retrieve or generate, AI has to be trained using data. Often, these data come from the expanse of content on the internet, and they may include misleading or harmful content, which influences how AI systems interpret specific phenomena. AI systems therefore inherit human biases, potentially misrepresenting information about specific events or societal groups, reinforcing prejudices. In the case of the Holocaust, it can result in AI output amplifying antisemitic stereotypes[4] that may result in distorted accounts of history. The risk is illustrated by the case of Tay, a chatbot developed

[3] For more information on the differences between non-generative and generative forms of AI, see Marr (2023).
[4] For more information on addressing antisemitism, see https://www.unesco.org/en/education-addressing-antisemitism

by Microsoft. In 2016, Tay was deployed on Twitter and in a matter of hours had been exploited to spread antisemitic statements (Kraft, 2016).

- AI applications can generate and spread disinformation. AI systems can be used to generate false information, and AI may amplify the spread of such content. For example, OpenAI's Dall-E has introduced a filter to moderate responses to the prompt 'Holocaust'; however, writing 'a picture of the atrocities that happened to the Jews in the 20th century' in the prompt bar can override this moderation filter and produce misleading Holocaust imagery. In the case of the Holocaust, search engines may then prioritize content promoting Holocaust denial in response to general prompts (e.g. 'Holocaust"; Makhortykh at al., 2021). Moreover, Generative AI systems such as Google Bard or OpenAI's ChatGPT may respond to user prompts with historical inaccuracies. Google Bard has occasionally invented fake eyewitness testimonies when prompted for information about less commonly known Holocaust incidents. Similarly, generative AI systems may distort facts when asked for specific historical details, such as the number of victims who perished in a particular event during the Holocaust (Makhortykh et al., 2023c).

- AI systems can undermine privacy. AI limits individuals' ability to control what information about their past is retrieved (Esposito, 2017). Even before the AI boom, the families of Holocaust victims and survivors struggled to control what information about their relatives was and is publicly visible (e.g. via web search engines) and how it may be used online. Images of victims such as Anne Frank have been used to create offensive memes (González-Aguilar & Makhortykh, 2022) distributed across the internet by AI systems. Additionally, historical evidence containing information about individuals (e.g. photographs from concentration camps) may be used as training data for generative AI models. This can result in the images being manipulated and published without victims' or survivors' consent, including for commercial purposes without respect for copyright or ethics. AI may, for example, reproduce violent or intimate photographs of victims or corpses without awareness of the sensitivity of the content. Getting consent protocols for sensitive data incorporated into the training of generative AI models could help protect the privacy of victims, survivors and their families, and it must be explored – not just for individuals - but also for key archives and databases with high stakes content (Burgess & Rogers, 2024).

- AI systems often lack transparency. AI systems often lack transparency that would help understand aspects of their functionality, a key element required to hold AI companies accountable for possible errors in and harm from their systems. It is often unclear which data AI systems have been trained on or how algorithms decide what content to retrieve, generate, or prioritize. For instance, Google's search engine considers various features covering relevance, quality and usability (Google Search, 2023) to decide how to rank content in response to user queries, but how these concepts are put into operation is less clear. It is, therefore, difficult to determine why certain sources of information about the Holocaust reach greater visibility than others.

How does AI threaten knowledge and understanding about the Holocaust?

The rapid development of AI technologies, methods and tools, particularly in content generation, distribution and recommendation, without proper oversight and regulation, risks deeper and wider spread of hate speech, Holocaust denial and other forms of antisemitic content. Building on the previous section, this paper outlines several AI challenges and their implications for Holocaust disinformation and antisemitism:

1 Manipulating and leveraging AI models to produce and spread hate speech

AI is vulnerable to exploitation by those seeking to further Holocaust denial and hate. Data flaws in AI design and deployment can result in such systems referring to or recommending hateful and sometimes violent content, while other systems lack any guardrails against the spread of hate speech (Norocel & Lewandowski, 2023; Tufecki, 2018). Awareness of these flaws can help malicious actors promote hateful content. For instance, a user searching for 'Jewish baby stroller' on Google in 2020 was likely to encounter images of portable ovens. These search engine results, glorifying and mocking Holocaust history, were probably a result of coordinated efforts by antisemitic trolls using fringe websites to promote these images so that AI prioritized them (Keyser, 2020). Such manipulation is often facilitated by the active integration of targeted advertising as part of the business model of platforms using AI models, with a special emphasis on cultivating user engagement (UNESCO, 2023c). In response to the disturbing search engine results, Google said, 'When people search for images on Google, our systems largely rely on matching the words in your query to the words that appear next to images on the web page. For this query, which is for a product that does not actually exist, the closest matches are web pages that contain offensive and hateful content.'[5]

Generative AI systems have also been manipulated into generating Holocaust denial and distortion through a technique known as jailbreaking (Li et al., 2023). Jailbreaking aims to circumvent the programming restrictions of AI systems, including the ones preventing the generation of offensive content, or the retrieval of personal data used to train the model. In the case of generative AI systems, jailbreaking usually takes the form of users purposefully prompting AI systems to generate content in an unethical way (Li et al., 2023). Notably, many generative AI tools lack the filters and guardrails against harmful content that can be found in systems like OpenAI's ChatGPT or Google's Bard. This makes it possible for those looking to operate maliciously to generate Holocaust denial content easily, without having to get round protective software.

2 False statements and narratives in generative AI content

AI models can generate content that sometimes creates misleading or false narratives. If not properly supervised, guided, or moderated, AI systems can be especially prone to errors that result in factually incorrect statements about the Holocaust. These errors can sometimes be due to incorrect descriptions of Holocaust materials in data on which an AI model is trained or from which it draws information (e.g. Holocaust denial websites incorporated into training data for generative AI models). They can also be due to data voids, which occur when AI does not have enough information on a specific episode of the Holocaust, for example, because it cannot retrieve data in a particular language (Mulligan & Griffin, 2018).

Data voids may lead to the retrieval of content irrelevant to the Holocaust. In 2020, 36 per cent of image search results for the query 'Holocaust' on Baidu, a major Chinese-language search engine, were related to exploitation movies and death rock music rather than related to the genocide (Makhortykh et al., 2021). The proportion of irrelevant content on Baidu was even higher for Russian-language queries, with no images related to the Holocaust in the top 50 search results. The unavailability of relevant and reliable information in some search engines limits knowledge and understanding about the history from internet searches.

[5] https://www.jta.org/2020/09/25/united-states/a-google-search-leads-to-anti-semitic-images-it-may-be-an-extremist-campaign

Similar problems occur for novel forms of AI due to their inability to evaluate the accuracy of the content they generate. Generative AI models, therefore, are prone to hallucinations where they 'fill up information voids with generated content that is not factually supported' (Makhortykh et al., 2023). A common example of such hallucinations is the tendency of ChatGPT to invent references to non-existent information sources (Alkaissi & McFarlane, 2023).

AI has invented events, personalities and even whole phenomena relating to the Holocaust; one example of such hallucination on ChatGPT is the concept of the 'Holocaust by drowning,' which assumes that mass murder campaigns took place in which Nazis and their collaborators drowned Jews in rivers and lakes. Although historically no such campaigns took place, AI invented them based on the concept of the Holocaust by bullets – i.e. large-scale murder by shooting. Google's Bard has, on other occasions, hallucinated fake quotes from witnesses to support distorted narratives of Holocaust massacres (Makhortykh et al., 2023c).

3 Producing fake historical evidence

AI-generated content cannot always be distinguished from human-generated content, even by experts. For instance, the current versions of many generative AI text systems allow users to create content about the Holocaust without considering its potential misuse. This includes generating content that imitates historical evidence, such as survivor testimonies or the personal reflections of perpetrators. Models such as ChatGPT or Bard make it easy to produce inauthentic materials that look convincing to non-experts. Visual and audio generative AI content, including deepfakes, may be particularly concerning for their emotionally evocative misrepresentations of Holocaust history. AI might be used to modify or generate images, audio, or videos, to make it appear as if historical figures or survivors are saying or doing things they did not. This could be exploited to distort Holocaust-related content, creating fabricated testimonies, or altering historical records (Leibowicz, 2021). Under these circumstances, it is essential to introduce guardrails, limiting the potential abuse of AI to generate fake evidence of the Holocaust, or highlight for users that the evidence is generated by AI. Several ideas for enabling such content guardrails, and supporting the identification of AI-generated material, are recommended by the global non-profit Partnership on AI in their Responsible Practices for Synthetic Media Framework (Partnership on AI, 2023).

Image generation systems can contribute to Holocaust distortion by reproducing stereotypical representations of the past or publishing real images in new (misleading and false) contexts, causing confusion. The risks include image-generative AI making content seem more credible, emotive and compelling, and therefore harder to debunk than text-based content, increasing the effort needed to counter it. Multimodal content (i.e. visuals supporting text-based messages and vice versa) also reinforces credibility and allows more possibilities for interaction, particularly when it has been convincingly written by different 'voices'. A recent example was the use of generative AI to create images of Holocaust victims for the 2023 exhibition in Ashkelon's Palace of Culture. The exhibition attracted criticism for being inauthentic, and historicizing the victims of the Holocaust by modernizing and beautifying their images (Riba, 2023).

AI has also been manipulated to rehabilitate and glorify Nazi ideology. AI-powered conversational agents have, for instance, simulated conversations with Holocaust perpetrators. One example is the Historical Figures app, which allows users to chat with prominent Nazis such as Adolf Hitler or Joseph Goebbels using generative AI (Ingram, 2023). In the dialogues, the AI app falsely claimed that individuals such as Goebbels were not intentionally involved in the Holocaust and had tried to prevent violence against Jews (Horovitz, 2023). These impersonations of historical and famous figures concerningly normalize antisemitic and violent ideologies. Deepfake videos of the actor Emma Watson reading from Hitler's Mein Kampf circulated on 4chan and spread to other social media (Sellman, 2023). The videos used technology released by ElevenLabs, a research company, which allows users to use a text-to-audio tool to type in words and hear them repeated in the voice of an individual.

4 Jeopardizing belief in authentic historical evidence

Holocaust denial may, in fact, not require influential deepfakes of historical figures, nor survivors saying or doing things they did not, to cast doubt on the truth about the Holocaust.

Just the societal knowledge that content can be generated or manipulated with AI may increase disbelief in the real evidence of the Holocaust and survivor experiences. The liar's dividend – the potential to leverage knowledge about the existence of AI-generated content to suggest that real videos and content are AI-generated – may perhaps be more influential than the fake artifacts themselves (Citron & Chesney, 2019). One might not need fake survivor testimony or a fake of Hitler saying he loves the Jews to support Holocaust denial. It would only be necessary to suggest that real video evidence in archives of survivors or Nazis has been made with AI. This might possibly prompt broader rejection of evidence of the Holocaust (Leibowicz, 2021).

5 Oversimplifying Holocaust histories

The history of the Holocaust is immensely complex. While historians and educators are skilled in explaining this complex past to different audiences, including young people, research (Makhortykh et al., 2021; 2023) indicates that AI-driven systems, such as search engines, tend to focus on just a few aspects of the Holocaust. Paradoxically, the technology companies that intervene to remedy the bias and accuracy issues raised in points 1, 2 and 3 often do so by filtering queries to select authoritative sources; however, in doing so, they may consolidate Holocaust understanding to a few trusted sources.

For instance, AI may prioritize images of the liberated camps or encyclopedia-style summaries. Such selective representation of the Holocaust is not new. Yet in the case of AI systems, it is worsened by the lack of transparency as to how AI prioritizes information and its limited capacities for providing additional contextual information. Likewise, on social media platforms, information from AI recommender systems is ranked – causing some researchers to fear the creation of 'echo chambers' and 'bubble filters' (Leerseen, 2020).[6]

Such restrictions can limit users' ability to find historical information without interference and disrupt understanding of the past by oversimplifying the complex history of the Holocaust. Less well-known episodes of the Holocaust are often muted as a result. For instance, key 'Holocaust' into a basic internet search engine and 60 to 80 per cent of the top image results are of a single Holocaust site – i.e. Auschwitz-Birkenau (Makhortykh et al., 2021). Other histories, experiences, locations and testimonies of the Holocaust are excluded from search engines in a self-reinforcing pattern that risks oversimplifying our understanding of this complex past.

Similar problems are observed for more recent forms of generative AI. Text-generative AI systems (e.g. Google Bard) may generate false information about lesser-known events during the Holocaust, basing instead their data sets on better-known histories of places and events for which information is more widely available on the internet. For example, Bard's responses to prompts concerning the 1941 massacres in Liubar in Ukraine reiterate details from the more well-known narrative of Babyn Yar, resulting in historically inaccurate claims, both concerning how victims were killed and whether Ukrainian collaborators were involved in the killing.

6 Language bias: reinforcing gaps in global Holocaust understanding

AI systems are often programmed and designed to customize their performance for individual users depending on their location or the language they use. While this customization can often be useful (e.g. when searching for local services), it can reinforce gaps in understanding complex historical episodes among different social, cultural and geographic groups. The likelihood of being exposed to Holocaust denial varies depending on the words put into search engines for information about the Holocaust (Makhortykh et al., 2023b). In Cyrillic script (i.e. in Russian or Bulgarian languages), between 8 and 14 per cent of the top 10 search results for 'Holocaust' on Yandex and Bing promoted Holocaust denial in 2020 and 2021. By contrast, Holocaust denial was not present in the top 10 results for the same query in Latin script (i.e. in English and German languages) in 2021.

Queries in Russian-language search engines also retrieved more graphic images of the Holocaust than for those in English in 2020 (Makhortykh et al., 2021). For instance, more than 30 per cent of the top 50 image search outputs in Russian-language searches on Bing showed images of murdered victims, but it was less than 3 per cent for English-language searches.

[6] See UNESCO Guidelines for the governance of digital platforms: safeguarding freedom of expression and access to information through a multi-stakeholder approach (2023). https://unesdoc.unesco.org/ark:/48223/pf0000387339

There were also more historical photos of liberated camps in English-language than in Russian-language searches (40 per cent and 22 per cent on Google, respectively).

A similar pattern is observed for generative AI where the prompt language can influence the accuracy and reliability of the output. In the study examining how AI-powered chatbots respond to 74 prompts regarding the different aspects of the Holocaust in Ukraine (Makhortykh et al., 2023c), the researchers found the substantive differences in the chatbot responses depending on the language of the prompt. For instance, in the case of English-language prompts, Google Bard declined to respond only to 1% of prompts. However, when the same prompts were entered in Russian, the chatbot did not respond to more than 30% of prompts (Makhortykh et al., 2023c). Similarly, the factual correctness of outputs varied across languages and chatbots with the Ukrainian prompts for chatGPT returning the largest number of factually incorrect outputs (Makhortykh et al., 2023c). Under these circumstances, users of chatbots in particular languages are substantially more likely to receive incorrect information about the Holocaust.

Using AI to enhance understanding about the Holocaust

The scope of the Holocaust and the hundreds of thousands of testimonies from Holocaust survivors, perpetrators and witness, written or spoken in multiple languages across different countries and historical contexts, creates a vast historical record that would take over a lifetime for one individual to process. AI can organize, categorize and search the content of a testimony based on a series of parameters, providing new avenues for research and understanding, though still limited by the design of the database or algorithm. For instance, AI can be employed to reveal 'structures, patterns and trends that are not discernible when the focus remains on just a handful of close readings of individual texts' (Presner, 2016). AI systems are therefore increasingly used for supporting and advancing Holocaust research in different academic fields, by facilitating access to historical material. For instance, AI was used for extracting information about Holocaust victims from documents stored in the Arolsen Archives.[7] Specifically, it was used to help index documents, such as prisoner and transfer lists, which were particularly complex and time-consuming for humans to process (Lee, 2018; Arolsen Archives, 2022).

The advanced capacities of AI for recognizing patterns in data mean it can be used to generate new insights about existing historical materials. Blanke et al. (2020) used deep learning to analyze sentiment in Holocaust testimonies and achieve a better understanding of the context of family memories stored in archives. Similarly, AI can be used to help analyze materials that are created in digital format that address Holocaust memory on social media, for instance, to determine the role of the Holocaust in the context of how authoritarian regimes use the past (Makhortykh et al., 2021).

AI tools and Holocaust education

Alongside AI innovations in generative AI tools and algorithmic recommendation systems, in recent years Augmented Reality (AR) and Virtual Reality technologies have been increasingly applied to enable new ways of learning about the Holocaust. School and museum educators can benefit from the advances of AI technology to interact with historical figures, including Holocaust survivors.

[7] The Arolsen Archives are the international center on Nazi persecution with the world's most comprehensive archive on the victims and survivors of National Socialism. The collection has information on about 17.5 million people. It contains documents on the various victim groups targeted by the Nazi regime and is an important source of knowledge for society today. https://arolsen-archives.org/en/

These experiences may foster empathy, by encouraging students to learn more about an individual's experience of genocide and ask questions to create a personalized conversation.

The Dimensions in Testimony project by the USC Shoah Foundation[8] uses AI to enable students and heritage center visitors to talk to virtual embodiments of Holocaust survivors (Shur-Ofry & Pessach, 2019). Using simple AI, individuals can ask questions and get appropriate real-time responses, selected by the program from pre-recorded replies – thereby enabling new possibilities for inquiry-based education (USC Shoah Foundation, 2021). In 2023, 'Tell me, Inge' launched as an extended reality project by Meta and StoryFile in partnership with UNESCO, the World Jewish Congress and the Claims Conference (UNESCO, 2023b). It focuses on the story of the Holocaust survivor Inge Auerbacher and uses AI to analyze the user's voice-based prompts and select the most fitting pre-recorded responses together with 3D animations (Jones et al., 2023). This project aims to enable an immersive experience for the purposes of Holocaust education and remembrance. Meaningfully, the project involved a close collaboration between software and AI tool developers, the Holocaust survivor featured and her relatives, and oral historians of the Holocaust, supporting the project's factual accuracy and the respectfulness of its creative features.[9] The Israeli company D-ID, in collaboration with the March of the Living, similarly created AI-generated videos of those who fought in the Warsaw Uprising (D-ID, 2023).

AI educational projects have also provided opportunities for learners understand more about the different experiences of Holocaust victims. The Let Them Speak project (Naron & Toth, 2020) uses AI to facilitate the search for information within testimonies, as well as giving a voice to the victims who did not survive and could not share their experiences. The project extracts excerpts from the testimonies describing different forms of suffering and then uses machine learning to group them together in testimonial fragments, which serve as an embodiment of experiences not specific to any one individual but shared both by survivors and victims who were not able to testify.[10]

Recommendations for educators

Before using AI tools and approaches in the classroom, educators must consider the ethical and pedagogical issues. In line with the policy for education and research action specified in the UNESCO's Recommendation on the Ethics of AI, it is recommended that:

- **Educators should familiarize themselves with different AI technologies and select appropriate tools:** Educators should explore AI-powered educational platforms, digital archives and interactive experiences related to the Holocaust. Although the number of tools and resources is growing, educators should understand how these can support teaching and learning objectives before they use them. Not all tools enhance learning, and traditional methods may sometimes be more appropriate, particularly when dealing with a sensitive topic. Educators should consider factors such as the age and maturity of their students, available technology infrastructure, curriculum goals and the classroom context. Virtual reality (VR) and augmented reality (AR) are frequently combined with machine learning and other AI techniques to enhance the user experience (UNESCO, 2021b, p. 17). Educators can use these AI technologies to engage learners in the lived experiences of Holocaust victims and survivors. In general, there is limited evidence to support the effectiveness of the use of AR/VR to enhance the quality of learning, so these technologies should be used cautiously and always be guided by human educators. Educators should be careful to assess the reliability of AI technologies to ensure the experience is historically accurate, and that it shares the lived experiences of real victims and survivors. VR experiences that simulate the experience of a concentration camp, or other stages of genocide, are not suitable and should be avoided.[11]
- **Provide context to ensure meaningful engagement with AI technology:** Before introducing AI-powered materials or experiences, educators should provide students with appropriate historical information about the Holocaust. Without information about the historical context, including events, key figures and information about antisemitism, learning may be superficial or incomplete. The same must be done for the AI tools, too. Educate learners about how systems work, their limitations and how to navigate content authentication and detection tools (and their limitations) (USC Shoah Foundation, 2021).

[8] https://sfi.usc.edu/dit)

[9] The project is freely available via the dedicated website – https://inge.storyfile.com/ – available in English and German.

[10] The project is publicly available online: https://lts.fortunoff.library.yale.edu/anthology

[11] The International Holocaust Remembrance Alliance's Recommendations for Teaching and Learning about the Holocaust (2019) advise caution when using simulations and role play: 'Beware of simulation, creative writing or role play exercises that encourage learners to imagine they were directly involved in the Holocaust. Attempts "to relate" can lead to false equivalencies or trivialization as learners try to find comparisons with their own lives. Some young people may over-identify with the events of the Holocaust and become excited by the power and even the "glamour" of the Nazis. Some may demonstrate a morbid fascination with the suffering of the victims. Learners with traumatic life experiences or family histories can also experience intense stress as they reconnect with those episodes through the historical exploration.'

- **Promote critical thinking:** Educators should encourage students to critically evaluate AI-generated content, interpretations and perspectives related to the Holocaust. Education that uses AI should also develop AI literacy and help learners question AI outputs, understand evidence and sources, consider biases and analyze the reliability of information provided by AI tools. AI-powered experiences can be a catalyst for classroom discussions and reflective activities, including considering the ethics and risks of AI-generated representations of the Holocaust. Explain to students how to evaluate visual labels and technical signals attached to media (e.g., metadata) that convey whether or not the content is synthetic or manipulated (USC Shoah Foundation, 2021).
- **Reflect on ethical considerations:** Engaging learners in discussions about the ethical implications of using AI to learn about sensitive historical topics like the Holocaust enables young people to reflect on how and why they are using AI. In this way, educators can encourage learners to use AI responsibly in the wider world. Schools should encourage learners to reflect on issues such as privacy, consent, representation and the responsible use of technology in historical education. Moreover, learning about the impact of AI systems should include learning about, through and for human rights and fundamental freedoms, meaning that the approach and understanding of AI systems should be grounded in their impact on human rights and access to rights, as well as on the environment and ecosystem (UNESCO, 2021, p. 23).
- **Consider learner identity:** How do learners' identities, values, and backgrounds affect their interest in and preconceptions about AI and the Holocaust itself? Lessons that involve personal identity or cultural values may stimulate an audience's interest and motivation (UNESCO, 2022). Learners should understand that systems may be biased, and the different levels and ways in which bias and prejudices can be introduced into AI outputs. Educators need to be alert to culturally insensitive or contextually inaccurate language and the inadvertent perpetuation of stereotypes or cultural biases (UNESCO, 2023d).
- **Be vigilant about falsehoods and biased information:** Use extreme caution when using generative-AI to create lesson materials about the Holocaust, due to the risk of introducing bias, prejudice and erroneous information into teaching resources. Educators should look to trusted and reliable organizations to provide guidance and materials for teaching and learning about sensitive histories.
- **Protect data privacy and avoid plagiarism:** Educators must comply with the appropriate regulations on data privacy to protect learners' personal information online and ensure that learners are aware of the risks of using AI, including regarding intellectual property.[12]

Digital media and AI literacy in Holocaust education

The growing use of AI in Holocaust education and remembrance prompts the need for developing new forms of digital literacies for students, Holocaust educators and heritage practitioners. AI literacy brings together different conceptual expertise and practical skills, ranging from understanding how AI functions and can be used, to being able to evaluate AI performance and judging whether its implementations are ethical (Ng et al., 2021). In the context of the Holocaust, these literacies are crucial for identifying how AI systems can be used to acquire information about the past, and for detecting instances when AI may distort historical facts or contribute to Holocaust denial. The practical steps for developing AI literacies in relation to the Holocaust can take different forms:

- Critically investigating possible bias and distortion in how AI systems represent well-known historical facts. This might be achieved by close reading of textual and visual content generated by AI in response to a well-known episode of the Holocaust or the biography of the Holocaust victim. In the course of such close reading, students can identify what aspects of the event AI can understand properly, and for which aspects it may simplify or distort details;
- Experimenting with diverse ways of using AI to produce content about the past, to reflect on the ethical and historical implications of the technology. Such experiments might include the use of AI for visualizing part of a survivor testimony. It is crucial that such experimentation is accompanied by a critical reflection on the limits of the technology, the need to protect historical facts, and careful considerations of the impact of AI on Holocaust knowledge and understanding; and
- Understanding methods to identify and label AI-generated content, and their limitations, to more skillfully evaluate content as authentic or not (Partnership on AI, 2023).

[12] See UNESCO (2023) Guidance for generative AI in education and research.

How to counter Holocaust distortion in AI systems

AI systems rely on recognizing patterns in the data on which they were trained to retrieve or generate content. It is therefore difficult for AI systems to recognize nuances related to educating about and remembering complex historical events such as the Holocaust. Holocaust memory encompasses millions of experiences and stories of victims and survivors; it is shaped by the choices and actions of perpetrators, bystanders and rescuers. There are different narratives about the Holocaust, varying from global didactic accounts to more local individual-oriented stories, to claims aiming to deny or distort historical facts.

The complexity of the Holocaust makes it difficult for educators to decide how to organize information about it into a narrative that diverse groups of learners can sympathize and engage with. This task is further complicated by differences in narratives associated with different types of Holocaust remembrance and history writing (e.g. the more personal narratives of survivor groups and the more formal narratives of heritage institutions; different historical timelines, experiences and actors across countries; or differences in narratives about the Holocaust that developed in the West and in countries formerly under communist regimes). However, the growing impact of AI on how the Holocaust is remembered and taught about implies that companies and individuals designing AI systems must take on the same kind of responsibilities as Holocaust historians and educators. AI companies and developers must ensure that their products do not undermine the efforts of educators and historians by simplifying the past, or depicting it inaccurately. More urgently, they must endeavor to introduce monitoring and oversight, to avoid facilitating the promotion of Holocaust denial and antisemitism in all languages.

AI companies can look at how AI systems should be optimized for Holocaust-related information, and they should be transparent about their decisions and algorithms. Any attempt to generalize the different experiences of suffering will probably over simplify understanding of the complexity of the Holocaust. A more inclusive approach would be to ensure that AI recognizes the existence of diverse narratives about the Holocaust, without accommodating denial or misinformation, and will allow for the importance of freedom of expression and access to information.

This requires careful consideration of the training data that inform AI systems and how AI systems are leveraged for content moderation and filtering, and ensuring that AI-driven outputs reflect the variety of potential queries a diverse user base might put forward. Developing such mechanisms requires more intense collaboration between AI developers and other stakeholders in the field of Holocaust memory and education.

Monitoring and evaluating standards for AI systems' management of information about the Holocaust.

How should we evaluate the performance of AI systems in the management, distribution and production of Holocaust content? An answer is not easy, considering the lack of established industry standards on how AI systems should manage information about different socially relevant subjects. Another difficulty involves the fact that different AI systems and tools may be evaluated based on related but different processes – for example, AI systems that enable content to be recommended vs AI systems that allow content to be generated. Similarly, there is little clarity on what guardrails should be put in place to prevent the abuse of these systems to distort facts and manipulate public opinion. AI developers and countries must adhere to international human rights standards, including those embedded in the International Covenant on Civil and Political Rights (ICCPR) (UN, 1966). In line with the ICCPR, standards for information management by AI can focus on preventing the prioritized retrieval or generation of content that may be damaging to the rights and reputations of the individuals, threatening the national security or public order, advocate for hatred and propagate war (i.e. Articles 19 and 20 of ICCPR).

In the case of the Holocaust, this guidance would imply that the standards will focus on preventing AI from amplifying or generating antisemitic and denialist claims, and avoid regulating other aspects of Holocaust information management. Such a stance has already been adopted by some AI developers to avoid exploitation of system vulnerabilities by Holocaust deniers (as happened in the case of Microsoft Tay, described above), but it has yet to become the general standard.

At the same time, avoiding content denying and distorting the Holocaust may not be the only criterium that it is important to take into consideration when assessing how AI manages information about the Holocaust. Because of the complexity of the Holocaust as a historical phenomenon and the large volume of available information, developers of AI systems inevitably have to make choices about what types of lawful content about the Holocaust are to be prioritized or downgraded when retrieving or generating information. This may result in a simplified representation of a highly complex history that is at risk of being shaped by political or other agendas.

Integrating Holocaust-specific ethical values into AI design

Human rights impact assessment tools can be used to evaluate the performance of AI systems (UNESCO, 2023), to integrate ethical risk management in the system design. The core criteria of UNESCO's (2023) ethical impact assessment – i.e. fairness, non-discrimination and diversity – may need to be specified in relation to AI systems used for managing Holocaust-related information. For instance, are there additional requirements for ensuring the fairness of an AI system besides providing equal access to information about the Holocaust to all its users? Do AI developers and companies have an ethical duty to protect the historical record of a genocide, particularly one that is subject to antisemitic defamations?

Similarly, in the case of privacy and data protection, it is important to consider the particular circumstances under which much Holocaust evidence was produced. These circumstances include the lack of informed consent from victims, due to the evidence often coming from the perpetrators, or consent being granted for specific uses of materials or testimonies long before the rise of AI systems. There are also additional ethical questions concerning the relationship between individual rights and public interest (e.g. protection of historical truth). Shall privacy concerns be taken into consideration when deciding how AI can or cannot use information about Holocaust perpetrators and their relatives? Can the removal of such information be justified from the point of view of the right to erasure or the right to be forgotten? Are there specific practices for which the use of personal data by AI shall be supported or discouraged in the context of the Holocaust?

Consulting with different stakeholders, and community engagement

To counter the use of AI systems for Holocaust denial and distortion, it is essential to achieve a better understanding of how these systems can be used (and abused) in the context of Holocaust education and remembrance. Although a growing number of initiatives, such as the Digital Holocaust Memory project (https://reframe.sussex.ac.uk/digitalholocaustmemory/), aim to bring together stakeholders to discuss their experiences of using AI systems and to create best practices, there is still limited understanding of the uses of AI in this context. Just as important is achieving an understanding of how AI systems are used by the general public, young people and teachers, as well as 'bad actors' who manipulate AI technologies to spread Holocaust denial, distortion and antisemitic content.

AI developers can actively engage with these stakeholders to solicit feedback, share best practices, and co-create solutions that address the unique socio-cultural contexts and dynamics of online antisemitism. For example, the risk of AI systems producing antisemitic content or Holocaust denial and distortion would be reduced if developers engaged Jewish community groups in participatory design processes, focus groups and user testing sessions to gather insights, validate assumptions and refine prototypes. Content about the Holocaust benefits from oversight from experts and stakeholders, including historians, Holocaust survivors and their families, and educators, to ensure that AI technologies take into consideration their concerns, perspectives and contributions.

Historians, Holocaust survivors, educators and learners can also be integrated into efforts related to content authentication that would support the certification of real content in the AI age. Organizations such as the USC Shoah Foundation-funded Starling Lab for Data Integrity at Stanford University and the Coalition for Content Provenance and Authenticity (C2PA) are working on methods for cryptographically certifying that real content is indeed authentic, and that AI-generated content has indeed been generated by AI. Starling uses open-source tools, best practices and case studies to securely capture, store and verify digital content, with applications for journalism, history and law.[13]

[13] https://www.starlinglab.org/

Using AI technologies to counter Holocaust distortion

AI systems are increasingly used by a broad range of social media and AI platforms to identify content that may breach the terms of platform use and to remove, downgrade, or label such content (Gillespie, 2020; Saltz & Leibowicz, 2021). AI has been used to detect content supporting violent extremist ideologies (Aldera et al., 2021) and sexually explicit content (Cifuentes et al., 2022). Recently, an increasing amount of effort has been put in to applying AI systems to counter antisemitism online (Chapelan et al., 2023; Hoes, 2023). The Decoding Antisemitism project has brought together a transdisciplinary team of scholars from Germany, France and the UK, to detect and analyze antisemitic content online (https://decoding-antisemitism.eu/). The pilot project, by researchers at the Center for Research on Antisemitism (ZfA, TU Berlin) in collaboration with HTW Berlin, the University of Michigan School of Information, Cardiff University's HateLab and King's College London, analyzed explicit and implicit antisemitism on leading social media to create a code system and a guidebook to train AI models to identify antisemitic hate speech using a supervised machine learning approach.

A similar methodology may be applied to detecting Holocaust distortion. Based on the small set of experiments that followed the growing line of research investigating the capacities of generative AI to detect false information (Caramancion, 2023), ChatGPT and Bing AI are able to detect false claims related to the Holocaust (e.g. that gas chambers were used not for murder but disinfection), if given the definitions of different types of false claims (e.g. disinformation, misinformation and conspiracy theory). At the same time, they were not always able to correctly evaluate distorted information that was only partially false (e.g. that there were gas chambers in all Nazi concentration camps) or opinion (e.g. that gassing was the worst type of mass murder during the Holocaust) (Caramancion, 2023). The same semantic limitations that may prompt generative-AI systems to produce misleading and inaccurate content may also limit their capacity to detect misleading and inaccurate content.

Limitations

AI can enhance both knowledge and understanding of the Holocaust by bolstering research and education. It can shape what forms of future knowledge are developed about the Holocaust, as AI models and the datasets on which they are trained regulate the types of information about the Holocaust available to internet users. However, it is important to recognize the limitations of AI for countering Holocaust distortion and antisemitism, due to the constantly evolving nature of these phenomena and AI's limited semantic capacities. Added to AI's often opaque functionalities, it is a challenge for external actors to develop workable solutions. To address these limitations, it is crucial to develop more advanced AI systems, capable of better understanding the Holocaust as a unique and extraordinarily complex historical phenomenon, and the threats arising from distorting it.

Applying a value-sensitive design

The denial and distortion of the Holocaust is a human-created problem, fueled by prejudices, conspiracy theories, disinformation and hatred. While there is no technical fix that will ultimately 'solve' denial, there are several policy responses that can – with human oversight – reduce its proliferation through AI. To start with, AI models should be developed to be sensitive to the risk of reinforcing and reproducing prejudice when dealing with sensitive topics such as the history of the Holocaust, in line with value sensitive design. This approach advocates taking ethical values into consideration in the principles when designing new technologies, and it is gaining momentum in AI development (e.g. Umbrello & Van de Poel, 2021). Value-sensitive design acknowledges the importance of context-specific values that reflect stakeholders' interests and principles in the domain where the AI systems are applied. In the case of Holocaust content, AI system builders should uphold the following principles:

- Involve the Jewish community, Holocaust survivors and their descendants, educators, experts in antisemitism and Holocaust historians in the AI development cycle, to integrate their viewpoints and values in the system design.

- Ensure transparency and explainability concerning the design choices and technology underlying the AI system, and the sources of information about the Holocaust used to develop the system.
- Use a wide range of data sources to train AI systems, capturing various aspects of the history of the Holocaust across different European countries and globally, and in different languages, to prevent data voids.
- Consider the ethical and legal aspects of using AI to manage and generate information about the Holocaust, including matters of copyright and privacy of victims and survivors, but also the principles of fairness and diversity crucial for capturing the complex nature of the Holocaust and preventing its misuse and distortion.

Recommendations for countering AI-amplified Holocaust distortion

The recommendations for countering Holocaust distortion in the age of AI build on the core values and principles defined by UNESCO in its Recommendation on the Ethics of AI (UNESCO, 2021).

FOR POLICYMAKERS

- Integrate ethical AI principles into policymaking processes to guide the responsible development and deployment of AI technologies. Emphasize principles such as fairness, transparency, accountability and respect for human rights, to ensure that AI systems are designed and used in ways that uphold historical accuracy, dignity and integrity. Ensure that such values can be brought into operation in technological development and deployment.
- Develop and enforce regulatory frameworks that uphold human rights including freedom of expression, while specifically addressing the spread of prejudice and disinformation through AI-powered platforms and technologies. Encourage platforms to invest in transparent AI-powered content moderation tools capable of detecting and flagging Holocaust denial and distortion, and ensure that there is human oversight to navigate semantically complex decisions and inevitable errors.
- Establish mechanisms for monitoring and evaluating the effectiveness of policies and interventions aimed at countering AI-amplified Holocaust denial and distortion.
- Facilitate international cooperation through information sharing, exchange of best practices and joint initiatives among governments, intergovernmental organizations, and civil society actors to develop coordinated responses to AI-amplified Holocaust distortion.
- Encourage the establishment of national commissions on the ethics of AI, and the uptake by public administration, notably in the education sector, of UNESCO's Ethical Impact Assessment methodology for the procurement of education-related AI tools and services – to ensure there is a minimum standard that respects values, principles, and the standards of the UNESCO Recommendation on the Ethics of AI.
- Invest in interdisciplinary research projects involving historians, computer scientists, ethicists and social scientists, to develop AI systems and tools for monitoring and countering Holocaust denial and distortion online, including through education and AI literacy, and consider the appropriate balance of human-AI collaboration to optimize the reduction in Holocaust distortion.

FOR EDUCATION POLICYMAKERS

- Urgently invest in educational programs that develop digital and AI literacy skills with a special emphasis on learners' ability to navigate disinformation, prejudice and hate speech. Support on-going teaching and learning about the Holocaust through reviews of curricula, textbooks and teacher training, including the development of knowledge and skills to build resilience to Holocaust denial and distortion, and antisemitism.

- Develop educational materials and initiatives targeting students, educators and the public to promote media and information literacy (MIL), critical thinking and historical accuracy in online discourse. Curricula should include provision in ICT; computing and MIL for learners should develop an awareness of how AI tools work and their susceptibility to bias, so that they can assess and evaluate their outputs.

- Strengthen the capacity of education systems to develop understanding of prejudice, bias and antisemitism, and build resilience against disinformation and distorted representations of Holocaust history, both in their curricula and through teacher training. Educators can warn learners about AI generated disinformation and misinformation about the Holocaust, and help them identify and recognize it.

- Enable opportunities for computer science and ICT curricula to build critical thinking about the ways computing can affect society, including through the spread of disinformation and hate speech.

- Learners need sound foundational knowledge about the Holocaust and skills to analyze sources and evidence. Provide support for Holocaust educational organizations, museums and institutions to develop AI educational programs, and integrate digital literacy into Holocaust education curricula.

FOR ARCHIVES, MUSEUMS AND MEMORIALS

- Continue the digitization of Holocaust-related historical collections to expand the amount of data that can be used for training AI systems and improve their performance.

- Develop guidelines on what information should be used by AI systems for retrieving and generating outputs, and how these systems should deal with legal and privacy considerations, including ones related to copyright and the privacy of victims.

- Adopt AI systems to facilitate access to information about the Holocaust and enable new ways of learning about it.

- Archives, museums and memorials can offer training to researchers using their institutions on the AI systems they employ and its impact on search engine outputs. They may also support AI-system developers to understand the sensitivities, historical contexts and risks of using their databases in AI models.

FOR AI PLATFORMS AND DEVELOPERS

- Conduct human rights due diligence in relation to AI systems, evaluating the risks and impact of their policies and practices on human rights, and defining mitigation measures, including through the use of ethical impact assessment tools (UNESCO, 2023) to ensure that AI systems used for managing and generating information about the Holocaust follow the principles of AI ethics (e.g. fairness, transparency, explainability and accountability), and international human rights standards.

- Adhere to international human rights standards, including in AI system design and moderation, and curation of content that is retrieved and produced. AI systems should follow relevant international human rights standards, including the UN Guiding Principles on Business and Human Rights. Design should ensure that harm is prevented; content moderation and curation policies and practices should be consistent with human rights standards, whether implemented algorithmically or by individuals, with knowledge of local languages and linguistic context, and adequate protection and support for human moderators.

- Ensure that AI systems are transparent, including openness about how they operate, with understandable and auditable policies as well as multistakeholder-agreed metrics for evaluating performance. This includes transparency about the tools, systems and processes used to moderate and curate content on their platforms, including their automated processes and the results they produce.

- Embed authenticity and transparency signals such as cryptographically signed metadata (e.g. C2PA, Content Credentials) in generative

AI media to support identification of AI-generated content by those distributing it and audiences encountering it.[14]

- Develop procedures that ensure a continual evaluation of the quality of training data for AI systems, including the adequacy of the data collection and selection processes, which would be diverse and inclusive of Holocaust remembrance.

- Foster partnerships with Holocaust survivors, descendants and Jewish communities to ensure their voices are heard in policymaking processes. Consult with experts, community groups and organizations to better understand their concerns, experiences and needs in fighting Holocaust denial and distortion online. Empower young people and social media communities to actively participate in countering misinformation and promoting accurate historical narratives on digital platforms.

- Ensure that AI system developers make information and tools available for users to understand and make informed decisions about the digital services they use, helping them assess the information on the platform as well as understanding how to raise complaints and get redress. This should include targeted media and information literacy programs available in several languages.

- Make AI systems accountable to relevant stakeholders – to users, the public, advertisers and the regulatory system – in implementing their terms of service and content policies; and give users the ability to challenge content-related decisions, whether they be users whose content was taken down or users complaining about content violating international human rights law.

FOR RESEARCHERS

- Conduct more empirical research into how AI systems deal with information about the Holocaust and how these systems are used by different stakeholders to critically assess the possibilities offered and risks posed by AI systems.

- Facilitate the development of monitoring and evaluation standards for assessing how different forms of AI deal with information about the Holocaust, and design criteria for improving the performance of AI systems in the context of Holocaust education and remembrance.

- Develop forensic expertise and technical solutions, such as icons, interstitial warnings and watermarking, compliant with different regulatory approaches to help easily identify fake Holocaust content for the broader public and in the media.

[14] https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/

References

Aldera, S., Emam, A., Al-Qurishi, M., Alrubaian, M., & Alothaim, A. (2021). Online extremism detection in textual content: a systematic literature review. IEEE Access, 9, 42384–96.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus, 15(2).

Arolsen Archives. (2022). #everynamecounts uses AI to uncover information on victims of Nazi persecution. https://arolsen-archives.org/en/news/everynamecounts-uses-ai-to-uncover-information-on-victims-of-nazi-persecutionormation-on-victims-of-nazi-persecution/

Blanke, T., Bryant, M., & Hedges, M. (2020). Understanding memories of the holocaust – A new approach to neural networks in the digital humanities. Digital Scholarship in the Humanities, 35(1), 17–33.

Bradshaw, S. (2019). Disinformation optimised: Gaming search engine algorithms to amplify junk news. Internet policy review, 8(4), 1–24.

Burgess, M., & Rogers, R. (2024, April 10). How to Stop Your Data From Being Used to Train AI. Wired. Retrieved April 20, 2024, from https://www.wired.com/story/how-to-stop-your-data-from-being-used-to-train-ai/

Caramancion, K. M. (2023, June). Harnessing the Power of ChatGPT to Decimate Mis/Disinformation: Using ChatGPT for Fake News Detection. In 2023 IEEE World AI IoT Congress (AIIoT) (pp. 0042–46). IEEE.

Chapelan, A., Ascone, L., Becker, M. J., Bolton, M., Haupeltshofer, P., Krasni, J., & Vincent, C. (2023). Decoding Antisemitism: An AI-driven Study on Hate Speech and Imagery Online. Fifth Discourse Report.

Cifuentes, J., Sandoval Orozco, A. L., & Garcia Villalba, L. J. (2022). A survey of artificial intelligence strategies for automatic detection of sexually explicit videos. Multimedia Tools and Applications, 1–18.

Delegation of the European Union to Ukraine. (2023). Statement by High Representative on Russia's misuse of the Holocaust in its current aggression. https://www.eeas.europa.eu/delegations/ukraine/statement-high-representative-russia%E2%80%99s-misuse-holocaust-its-current-aggression_en?s=232

D-ID. (2023, April 17). The Heroes Speak: Marking 80 Years for the Warsaw Ghetto Uprising. D-ID Blog. https://www.d-id.com/news/the-heroes-speak-marking-80-years-for-the-warsaw-ghetto-uprising/

Donahue, Lauri. (2018, January 3). A Primer on Using Artificial Intelligence in the Legal Profession. Harvard Journal of Law & Technology. https://jolt.law.harvard.edu/digest/a-primer-on-using-artificial-intelligence-in-the-legal-profession

Eskens, S., Helberger, N., & Moeller, J. (2017). Challenged by news personalisation: five perspectives on the right to receive information. Journal of Media Law, 9(2), 259–284.

Esposito, E. (2017). Algorithmic memory and the right to be forgotten on the web. Big Data & Society, 4(1), 2053951717703996.

Gibson, P. L., & Jones, S. (2012). Remediation and Remembrance: 'Dancing Auschwitz' Collective Memory and New Media. Journal for Communication Studies, 5(10).

Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2), 2053951720943234.

González-Aguilar, J. M., & Makhortykh, M. (2022). Laughing to forget or to remember? Anne Frank memes and mediatization of Holocaust memory. Media, Culture & Society, 44(7), 1307–29.

Hoes, E., Altay, S., & Bermeo, J. (2023). Using ChatGPT to Fight Misinformation: ChatGPT Nails 72% of 12,000 Verified Claims.

Horovitz, M. (2023). Chatbot denounced for generating remorseful responses from top Nazi figures. https://www.timesofisrael.com/chatbot-denounced-for-generating-remorseful-responses-from-top-nazi-figures/

Hurst, L. (2023). AI deepfakes are being weaponised in the race for US president – and Trump is the latest target. https://www.euronews.com/next/2023/06/09/ai-deepfakes-are-being-weaponised-in-the-race-for-us-president-and-trump-is-the-latest-tar

Ingram, D. (2023). A chatbot that lets you talk with Jesus and Hitler is the latest controversy in the AI gold rush. NBC News. https://www.nbcnews.com/tech/tech-news/chatgpt-gpt-chat-bot-ai-hitler-historical-figures-open-rcna66531

Jones, V., Gaylord, W., Palitz, A., Auerbacher, I., Csabai, A., Franke, D., & Smith, S. (2023). Tell Me, Inge … An Interactive Interview with a Holocaust Survivor. In ACM SIGGRAPH 2023 Real-Time Live! (pp. 1–2).

Keyser, Z. (2020). Google responds after search term yields antisemitic allusion to Holocaust. https://www.jpost.com/diaspora/antisemitism/google-responds-after-search-term-yields-antisemitic-allusion-to-holocaust-643811

Kuznetsova, E., Makhortykh, M., Vziatysheva, V., Stolze, M., Baghumyan, A., & Urman, A. (2023). In Generative AI we Trust: Can Chatbots Effectively Verify Political Information?. arXiv. https://doi.org/10.48550/arXiv.2312.13096

Kraft, A. (2016). Microsoft shuts down AI chatbot after it turned into a Nazi. https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/

Lee, B. C. G. (2019). Machine learning, template matching, and the International Tracing Service digital archive: Automating the retrieval of death certificate reference cards from 40 million document scans. Digital Scholarship in the Humanities, 34(3), 513–535. https://doi.org/10.1093/llc/fqy063

Leerssen, P. J. (2020). The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems. European Journal of Law and Technology, 11(2). https://ejlt.org/index.php/ejlt/article/view/786

Leibowicz, C. (2021, May 4). Preparing for a World of Holocaust Deepfakes. Tablet Magazine.

Li, H., Guo, D., Fan, W., Xu, M., & Song, Y. (2023). Multi-step jailbreaking privacy attacks on ChatGPT. arXiv preprint arXiv:2304.05197.

Makhortykh, M., Urman, A., & Ulloa, R. (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).

Makhortykh, M., Lyebyedyev, Y., & Kravtsov, D. (2021b). Past Is another resource: remembering the 70th anniversary of the victory day on livejournal. Nationalities Papers, 49(2), 375-388.

Makhortykh, M., Zucker, E. M., Simon, D. J., Bultmann, D., & Ulloa, R. (2023). Shall androids dream of genocides? How generative AI can change the future of memorialization of mass atrocities. Discover Artificial Intelligence, 3(1), 28.

Makhortykh, M., Urman, A., Ulloa, R., & Kulshrestha, J. (2023b). Can an algorithm remember the Holocaust? Comparative algorithm audit of Holocaust-related information on search engines. Digital Memory (pp. 79–93). Wallstein Verlag.

Makhortykh, M., Vziatysheva, V., & Sydorova, M. (2023c). Generative AI and Contestation and Instrumentalization of Memory about the Holocaust in Ukraine. Eastern European Holocaust Studies, 1(2), 349-355

Manca, S., Raffaghelli, J.E., & Sangrà, A. (2023). A learning ecology-based approach for enhancing Digital Holocaust Memory in European cultural heritage education, Heliyon, 9(9). Marr, B. (2023). The Difference Between Generative AI And Traditional AI: An Easy Explanation For Anyone. Forbes. https://www.forbes.com/sites/bernardmarr/2023/07/24/the-difference-between-generative-ai-and-traditional-ai-an-easy-explanation-for-anyone/

Mulligan, D., & Griffin, D.S. (2018). Rescripting Search to Respect the Right to Truth. Georgetown Law Technology Review, 557–584. National Academies of Sciences, Engineering, and Medicine. (2022). Fostering Responsible Computing Research: Foundations and Practices [Consensus Report]. Washington, DC: The National Academies Press. https://doi.org/10.17226/26507

Naron, S., & Toth, G. M. (2020). Let Them Speak: An Effort to Reconnect Communities of Survivors in a Digital Archive. Mass Violence and Memory in the Digital Age: Memorialization Unmoored, 71-94.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041.

Norocel, O. C., & Lewandowski, D. (2023). Google, data voids, and the dynamics of the politics of exclusion. Big Data & Society, 10(1), 20539517221149099.

Partnership on AI. (2023, December 19). Building a Glossary for Synthetic Media Transparency Methods, Part 1: Indirect Disclosure. PAI Blog. https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/

Partnership on AI. (2023, February 27). PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action. PAI Blog. https://syntheticmedia.partnershiponai.org/

PBS. (2023). Antisemitic incidents on rise across the US, report finds. https://www.pbs.org/newshour/politics/antisemitic-incidents-on-rise-across-the-u-s-report-finds

Presner, T. (2016). 'The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive,' in: Probing the Ethics of Holocaust Culture, edited by Claudio Fogu, Wulf Kansteiner, and Todd Presner (Cambridge: Harvard University Press), 175–202.

Protect the Facts. (2023). Debunking Inappropriate Holocaust Comparisons: The COVID-19 Yellow Star. https://www.againstholocaustdistortion.org/news/debunking-inappropriate-holocaust-comparisons-the-covid-19-yellow-star

Ramgopal, K. (2020). Survey finds 'shocking' lack of Holocaust knowledge among millennials and Gen Z. NBC News. https://www.nbcnews.com/news/world/survey-finds-shocking-lack-holocaust-knowledge-among-millennials-gen-z-n1240031

Riba, N. (2023). These Holocaust AI Generated Images Distort History. Haaretz. https://www.haaretz.com/israel-news/2023-02-01/ty-article-magazine/.premium/new-israeli-exhibit-showcases-ai-generated-holocaust-images/00000186-0c8d-dc4a-a3be-efbf01e20000

Saltz, E., & Leibowicz, C. (2021, June 14). Fact-Checks, Info Hubs, and Shadow Bans: A Landscape Review of Misinformation Interventions. PAI Blog.

Sellman, M. (2023). Emma Watson reads Mein Kampf on 4Chan in deepfake audio trick. The Times. https://www.thetimes.co.uk/article/ai-4chan-emma-watson-mein-kampf-elevenlabs-9wghsmt9c

Shandler, J. (2017). Holocaust Memory in the Digital Age: Survivors' Stories and New Media Practices. Stanford, CA: Stanford University Press.

Simchon, A., Edwards, M., & Lewandowsky, S. (2023). The persuasive effects of political microtargeting in the age of generative AI. OSF (Open Society Foundations). 10.31234/osf.io/62kxq

Shur-Ofry, M., & Pessach, G. (2019). Robotic Collective Memory. Wash. UL Rev., 97, 975.

Sullivan, D. (2016). Google's top results for 'Did the Holocaust happen' now expunged of denial sites. https://searchengineland.com/google-holocaust-denial-site-gone-266353

Tiebel, A. & Eddy, M. (2010). Holocaust survivor's dance sparks controversy. NBC News. https://www.nbcnews.com/id/wbna38270574

Tufekci, Z. (2018). YouTube, the great radicalizer. New York Times. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. AI and Ethics, 1(3), 283-296.

Ueno, H. (2023). Artificial Intelligence as Dual-Use Technology. In: Hatzilygeroudis, I.K., Tsihrintzis, G.A., Jain, L.C. (eds) Fusion of Machine Learning Paradigms. Intelligent Systems Reference Library, vol 236. Springer, Cham. https://doi.org/10.1007/978-3-031-22371-6_2

UN Office of Information and Communications Technology (OICT). (2022). A Future with AI: Voices of Global Youth report. https://unite.un.org/sites/unite.un.org/files/a_future_with_ai-final_report.pdf

UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. (2018a). Report of the Special Rapporteur to the General Assembly on AI and its impact on freedom of opinion and expression. A/73/348. https://www.ohchr.org/EN/Issues/Freedom Opinion /Pages/ReportGA73.aspx

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNESCO. (2021b). AI and education: guidance for policymakers. https://unesdoc.unesco.org/ark:/48223/pf0000376709

UNESCO. (2022). K-12 AI curricula: a mapping of government-endorsed AI curricula. https://unesdoc.unesco.org/ark:/48223/pf0000380602?posInSet=2&queryId=4c5c5e7b-90f1-405b-be32-af5a23ed1268

UNESCO (2023). Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000386276

UNESCO. (2023b). Launch of 'Tell me, Inge', new extended reality experience about the Holocaust. https://www.unesco.org/en/articles/launch-tell-me-inge-new-extended-reality-experience-about-holocaust

UNESCO. (2023c). Platform problems and regulatory solutions: findings from a comprehensive review of existing studies and investigations

UNESCO. (2023d.) Guidance for generative AI in education and research. https://unesdoc.unesco.org/ark:/48223/pf0000386693/

Urman, A., & Makhortykh, M. (2023). The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat. 10.31219/osf.io/q9v8f

USC Shoah Foundation. (2021). USC Shoah Foundation Launches Web-Based Interactive Biography of Holocaust Survivor and Educator Pinchas Gutter on IWitness. https://sfi.usc.edu/news/2021/04/31101-usc-shoah-foundation-launches-web-based-interactive-biography-holocaust-survivor

USC Shoah Foundation (Director). (2021, October 7). Deepfakes and Holocaust Testimony | USC Shoah Foundation. https://www.youtube.com/watch?v=xqUDFAAPjsM

Van Hoboken, J. (2012). Search engine freedom: on the implications of the right to freedom of expression for the legal governance of Web search engines (Vol. 27). Kluwer Law International BV.

Walden, V. (2023). AI Literacies and Media Education. The Media Education Association. https://www.themea.org.uk/post/ai-literacies-and-media-education?fbclid=IwAR1SjbauUvD96VgPCSOk6O9-QkIHsQoWaEC53Cnn3PFQ24Vhj7s3VYJVaW4

Ward, B. (2021). Europe's Worrying Surge of Antisemitism. Human Rights Watch. https://www.hrw.org/news/2021/05/17/europes-worrying-surge-antisemitism

unesco
United Nations
Educational, Scientific
and Cultural Organization

In partnership with

WORLD
JEWISH
CONGRESS

AI and the Holocaust: rewriting history?

The impact of artificial intelligence on understanding the Holocaust

The threats associated with AI on safeguarding the record of the Holocaust are manifold, including the potential for manipulation by malicious actors, the introduction of falsehoods or dissemination of biased information, and the gradual erosion of public trust in authentic records. This paper provides a warning of what is at stake for the preservation of historical truth in a digital era increasingly mediated by AI. While there are some benefits to be gained, such as enhanced engagement and interaction opportunities for learners, as well as more efficient data processing capabilities for researchers, to navigate these challenges and capitalize on the benefits, it's essential for AI designers, policymakers, educators, and researchers to collaborate closely. Only AI systems equipped with robust safeguards and human rights assessments, coupled with an increased focus on developing digital literacy skills, can uphold the integrity of historical truth and ensure the responsible use of artificial intelligence.

Stay in touch

https://www.unesco.org/en/teaching-holocaust-genocide

Follow @UNESCO on social media