Journal of Risk Research

ISSN: 1366-9877 (Print) 1466-4461 (Online) Journal homepage: www.tandfonline.com/journals/rjrr20

Generative AI and digital resilience: a research agenda

Christine Große & Leif Sundberg

To cite this article: Christine Große & Leif Sundberg (07 Aug 2025): Generative AI and digital resilience: a research agenda, Journal of Risk Research, DOI: 10.1080/13669877.2025.2539105

To link to this article: https://doi.org/10.1080/13669877.2025.2539105

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 07 Aug 2025.

Submit your article to this journal

Article views: 114

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rjrr20

JOURNAL OF RISK RESEARCH
https://doi.org/10.1080/13669877.2025.2539105

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Generative AI and digital resilience: a research agenda

Christine Große[a,b] and Leif Sundberg[b,c]

[a]Division of Computer Science, Electrical and Space Engineering, Department of Digital Services and Systems, Luleå University of Technology, Luleå, Sweden; [b]Risk and Crisis Research Centre, Mid Sweden University, Sundsvall, Sweden; [c]Department of Informatics, Umeå University, Umeå, Sweden

ABSTRACT

Generative artificial intelligence (GAI) systems, such as ChatGPT and DALL-E, enable the automatic creation of content that closely resembles human-produced work, allowing for both constructive and potentially harmful applications. This study investigates the impact of GAI on digital resilience in society, focusing on the opportunities and challenges these technologies present. Digital resilience is defined as the capacity to maintain societal functions despite disruptions to or by digital systems. This study draws on a systematic literature review of 164 research articles, with findings categorized into three key perspectives: technology, organization, and governance. Within each perspective, themes for further research are identified. The results highlight the importance of recognizing GAI's dual nature to both support and threaten digital resilience. To promote secure digitalization, our study calls for more empirical research on the risks associated with GAI applications, transparent artificial intelligence usage, and the establishment of ethical guidelines. The study also advocates for increased interdisciplinary research to ensure that new technologies and resilience can coexist as foundational elements of a resilient, democratic society.

ARTICLE HISTORY
Received 22 November 2024
Accepted 25 June 2025

KEYWORDS
Generative AI; digital resilience; large language models; governance

1. Introduction

According to the World Economic Forum's 2023 Global Risks Report (World Economic Forum 2023), artificial intelligence (AI)-driven mis- and disinformation is currently one of the most pressing global threats. Emerging disruptive technologies—specifically generative artificial intelligence (GAI) tools that create various types of content, including text, images, and videos—have rapidly become accessible to the public. It is easy to envision a near future where anyone with access to a digital photograph of a person's face could seamlessly insert it into videos across any context. While these tools offer entertainment opportunities (e.g. placing actor Arnold Schwarzenegger into the musical *Sound of Music*), they can also be used for malicious purposes, such as disinformation (e.g. fabricated videos depicting political leaders). The adoption of AI-related technologies, such as machine learning (ML), deep learning (DL), and GAI (e.g. ChatGPT), has surged over the past decade (Dwivedi et al. 2023). This growth has been enabled by increased access to affordable computing power, large datasets, and more advanced algorithms (Duan, Edwards, and Dwivedi 2019). While these advancements can bring societal,

CONTACT Christine Große christine.grosse@ltu.se Division of Computer Science, Electrical and Space Engineering, Department of Digital Services and Systems, Luleå University of Technology, Luleå, Sweden; Risk and Crisis Research Centre, Mid Sweden University, Sundsvall, Sweden.
© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

organizational, and individual benefits, they also present serious challenges, including threats of unemployment (Ponce Del Castillo 2023) and the risk of "intelligent" systems discriminating against specific groups (Sadowski and Andrejevic 2020). A particularly troubling issue linked to AI systems is disinformation. New GAI systems are capable of generating highly realistic texts, images, and videos (e.g. deepfakes) that are difficult to distinguish from authentic human-made content. Recently, deepfakes have spread widely on social media (Karnouskos 2020), raising concerns about their potential to influence public opinion. Consequently, AI technologies are increasingly seen as tools that could be exploited by foreign powers to manipulate public sentiment, as evidenced by their use in the conflict between Russia and Ukraine and during elections in the United States and Israel (Große 2023). As such, in addition to the potential benefits GAI technologies bring to society, their extensive use also entails many challenges. In this paper, we investigate GAI in relation to digital resilience, defined here as the capacity to maintain societal functions despite disruptions to or by digital systems (Shandilya et al. 2024).

Against this backdrop, we identify a need to review literature that examines how GAI can be deployed safely, with relevant governance mechanisms in place to ensure responsible AI development and build resilience against its adverse effects (Schintler and McNeely 2022). The purpose of this study is to synthesize knowledge on the relationship between GAI and digital resilience by investigating the benefits and risks associated with GAI and identify needs for further research.

We pose the following research question (RQ): *How can generative AI contribute to or counteract digital resilience?*

Our study builds on a systematic literature review on GAI technologies, their applications, and their implications for digital resilience. The study contributes with providing a foundation for further investigations in this area through ten directions for future research.

The remainder of this paper is organized as follows. Section 2 provides background on GAI and digital resilience; Section 3 describes the methods for data collection and analysis; Section 4 presents the findings; Section 5 discusses the results; and Section 6 concludes the study.

2. Background: generative AI and digital resilience

2.1. Generative AI

The development of AI has accelerated rapidly in recent years, with modern AI systems characterized by varying degrees of autonomy and complexity (Berente et al. 2021). Advances in ML (Jordan and Mitchell 2015) and DL (LeCun, Bengio, and Hinton 2015), along with access to vast amounts of digital data, have enabled the creation of increasingly sophisticated systems. The evolution of neural networks has allowed machines to detect and interpret complex phenomena previously beyond their capacity, for example recognizing handwritten texts (McKim 2022) or tumors in X-ray images.

While the recent increase in practical AI applications has brought these systems to the forefront, their roots trace back to regression methods of the late nineteenth century (Galton 1889; Cramer 2002) and algorithmic advancements in the second half of the twentieth century (Kelley 1960; Rumelhart, Hinton, and Williams 1986; LeCun et al. 1989). However, commercially viable AI systems were rare, aside from some expert systems in the 1980s, which soon fell out of favor due to perceived inefficiency (Gill 1995), difficulties handling uncertainty, and limited learning capabilities (Russell and Norvig 2022). A key distinction between these earlier systems and today's AI is that the former were typically rule-based, whereas contemporary systems are designed to "learn" from data.

The recent advancement of AI has been driven by powerful computational capabilities and access to vast datasets (McKim 2022), contributing to the emergence of terms like "big data" (Lycett 2013; Günther et al. 2017; Jones 2019). In 2012, Alex Krizhevsky and colleagues marked

an important milestone by winning the ImageNet Large Scale Visual Recognition Challenge with their DL, AlexNet (Krizhevsky, Sutskever, and Hinton 2017). This achievement is widely regarded as the start of a "golden era" for large, data-driven models across fields such as image recognition, as well as domains like music (MusicNet: Thickstun, Harchaoui, and Kakade 2016) and 3D object recognition (ShapeNet: Chang et al. 2015). This breakthrough also spurred heightened interest in natural language processing (NLP) models designed to interpret and generate language, exemplified by bidirectional encoder representations from transformers and generative pretrained transformer, as described by Devlin et al. (2018) and Brown et al. (2020), respectively.

The next generation of AI systems promises capabilities beyond simple categorization and object recognition. Advances in neural networks, particularly through generative adversarial networks (Goodfellow et al. 2020), have enabled AI to not only recognize but also generate content. Central to these capabilities is the transformer architecture, an efficient framework that allows ML models to process text data by leveraging different types of weighting. This architecture incorporates both an encoder and decoder in a process known as "self-attention," which enables more nuanced and context-aware learning (Vaswani et al. 2017).

Generative AI is a collective term for these technologies, which have gained attention in recent years. The "new" aspect of GAI lies in its ability to generate content—such as text and images—based on user instructions, called "prompts." Popular platforms include ChatGPT, CodePilot, DALL-E 2, Midjourney, and Stable Diffusion. An early milestone for GAI came when an AI-generated painting won an art competition in 2022 (Dataconomy 2022).

Generative AI's adoption surged after OpenAI launched ChatGPT in November 2022. Within a few months, ChatGPT amassed over 100 million users—by comparison, Instagram took two and a half years to reach this milestone, and TikTok nine months (Reuters 2023). In early 2023, OpenAI released a more advanced version, ChatGPT-4, although the company has not disclosed the specific data used to train its models (OpenAI 2023). These large language models (LLMs), sometimes referred to as "foundation models" (Schneider, Meske, and Kuss 2024), are trained on vast datasets, allowing them to achieve a more nuanced "understanding" of context and content than earlier AI systems (Sundberg and Holmström 2024). This advancement has enabled GAI to perform a broad range of tasks, from classifying and summarizing text to programming, creating smart assistants, and supporting various types of data analysis. Consequently, GAI is expected to drive substantial value and innovation (Dwivedi et al. 2023). At the same time, critical voices have raised concerns about how these models are trained, the societal impacts of their widespread adoption, and the potential risks should they fall into the "wrong" hands (Feuerriegel et al. 2023). These concerns form the motivation for this study.

2.2. Digital resilience

This study addresses the need for research on how GAI can be securely implemented with appropriate governance mechanisms to ensure responsible AI development and strengthen societal resilience against adverse impacts (Schintler and McNeely 2022).

The term resilience originates from Latin and is often defined as a system's ability to return to its original state after a temporary disruption (Holling 1973). However, resilience following a disruption does not necessarily imply a return to the previous state (Handmer and Dovers 1996). It may also signify a pathway toward transformation (Miles and Petridou 2016), opening new opportunities and fostering actions that reflect an evolved outlook. Resilience can therefore be viewed as the adaptive behavior of a system, capable of evolving by coping with vulnerabilities through recovery, transformation, and adaptation to new conditions. It emerges as a characteristic of complex, interconnected systems, meaning it is often easier to identify resilience once it manifests rather than to engineer it directly. A common definition of resilience is the ability to withstand singular or extraordinary events (Kendra and Wachtendorf 2003). Resilience also

involves the capacity of a society or system (e.g. critical infrastructure) to maintain essential functions when exposed to a shock (Rose 2007).

Consequently, we argue that resilience should be approached from a broad perspective (Große 2022; Shandilya et al. 2024), particularly in studies of emerging technologies. Digital resilience, therefore, should encompass both technical and social dimensions: beyond technical attributes, it includes human behaviors and cultural factors. It may involve adapting to rapid changes and displaying organizational flexibility. Resilience also requires the capacity to manage and cope with the stresses induced by unexpected events (Salehi, Veitch, and Musharraf 2020). Effective adaptability is important for crisis management, particularly in situations where multiple actors must collaborate across organizational boundaries. The resilience of society depends largely on how well organizations and individuals can adapt to evolving circumstances, which calls for a measured approach to adjustment rather than reactive changes. This aspect of resilience is especially important in the context of digital resilience, as it fortifies societal resistance against disinformation and influence campaigns from foreign actors (see Große 2023).

The risks associated with AI and digitalization are inherently systemic. As digital systems increasingly shape the fabric of daily life and organizational processes, these risks become embedded within our routines and structures (Sundberg 2024). Regulating these systems poses challenges due to the absence of standardized definitions and the rapid pace of technological advancement. The shift toward a more digital society also complicates risk and crisis communication (Ngwenyama, Henriksen, and Hardt 2023), as people encounter vast amounts of information, making it harder to differentiate between verified public information and, for instance, social media disinformation (Karnouskos 2020). Previous research highlights various challenges in (inter-)organizational communication, particularly when dealing with classified information (Andersson 2023), with important knowledge and relationships often lost through staff turnover. Managing uncertainty from multiple sources, especially in planning and decision-making that involves many actors, remains a critical issue in risk and crisis management (Große 2019), where the role of GAI is now a consideration.

Research has demonstrated the important role of digital technologies in navigating societal disruptions, such as the COVID-19 pandemic, spurring increased interest in digital resilience (Cuel, Ponte, and Virili 2022; Boh et al. 2023; Tim, Leidner, and University of New South Wales 2023), with AI emerging as a key technology. Although AI holds promise for addressing societal challenges, concerns persist about its irresponsible use. Here, researchers have emphasized the need for an "AI-resilient" society, characterized by openness, awareness, and agreements (e.g. through regulations) where humans can navigate the uncertainties AI technologies bring (Bartz-Beielstein 2019). Building such a society requires cross-disciplinary engagement (Buzzanell 2023) to form a comprehensive understanding of these transformative technologies' implications for societal preparedness and resilient digitalization. We contend that this understanding must embrace a holistic view, moving beyond a narrow focus on technical risks alone. A sociotechnical approach is essential to build resilience and insight into managing the complexities that accompany the rise of GAI technologies. Therefore, our literature review, further described in the following section, incorporates both technical and social dimensions and their interactions, along with ethical and legal considerations, to build an integrated understanding of GAI's role in strengthening resilient democratic societies. Accordingly, this study contributes a foundation for advancing digital resilience.

3. Materials and methods

3.1. Sampling

The search and selection of publications for this systematic literature review followed a four-step process, involving the identification of relevant scientific articles and a thematic analysis of selected articles (Vom Brocke et al. 2009; Paré et al. 2015). The steps are as follows: 1) Defining

Table 1. Search strategy, including keywords, hits in scopus, and level of analysis.

Search Terms	Hits	Level of Analysis
TITLE-ABS-KEY ("generative AI" OR "generative artificial intelligence" OR "large language model*" OR chatgpt OR llm)	10,397	—
AND TITLE-ABS-KEY (preparedness OR resilience OR risk)	773	Title
AND TITLE-ABS-KEY (preparedness OR resilience OR risk OR threat OR hybrid OR secur* OR public OR safe*) AND NOT (student* OR teach* OR in-situ OR patient*)	1,008	Title
TITLE-ABS-KEY ("generative AI" OR "generative artificial intelligence" OR "large language model*" OR chatgpt OR llm) AND TITLE-ABS-KEY (preparedness OR resilience OR risk OR threat OR hybrid OR secur* OR public OR safe* OR defen*) AND NOT (student* OR teach* OR in-situ OR patient*)	1,030	Title, keywords
AND (LIMIT-TO (LANGUAGE,"English") OR LIMIT-TO (LANGUAGE,"German"))	919	Title, keywords
LIMITED TO articles, conference, editorial, review	773	Abstract
After first review of the selection: Relevant for systematic literature review	344	Full text
After removal of conference proceedings (books) and inaccessible and incorrect texts	323	Full text

Table 2. Most frequent sources of publications in the dataset.

Number of Publications	Source
11	Findings of the Association for Computational Linguistics. Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023
10	ACM International Conference Proceeding Series
10	CEUR Workshop Proceedings
8	IEEE Access
8	Lecture Notes in Computer Science Including Subseries LN Artificial Intelligence and Bioinformatics
4	Proceedings of the Annual Meeting of the Association for Computational Linguistics
3	AI & Society

the search strategy, 2) Reviewing titles and keywords, 3) Initial analysis of abstracts, and 4) Identifying each publication's focus and application area.

Due to Scopus's extensive coverage of scientific publications, we selected this database as the primary source. Various keywords were combined to refine the search focus. Table 1 summarizes the search and selection process.

The review of titles and keywords guided the refinement of search terms. Selection criteria were limited by language, English or German, and by document type, including journal articles, conference papers, editorials, and reviews. This process yielded an initial set of 773 publications for further analysis. Based on reviews of the abstracts the review of the abstracts, 344 publications were deemed relevant to the literature review, though 21 were subsequently excluded for various reasons, including lack of full-text availability, incorrect full texts, or identification as introductory material for conference proceedings.

After identifying each publication's focus and application area, 132 articles were assessed as irrelevant for this review. Additionally, 27 publications were excluded due to their focus on narrow cases of AI application without broader scientific reflection on the implications in a wider context. The final selection for in-depth analysis comprised 164 scientific contributions, published across 103 journals and conferences. Table 2 displays the most frequent sources of the reviewed publications.

The search for this review reveals a vast increase in publication rate, with a modest number of papers published before 2023, followed by over 700 publications throughout 2023 and a similar number in the first half of 2024. This dramatic increase is closely linked to the growing relevance of GAI tools, such as ChatGPT, across various fields. It also reflects the rapid evolution of the research landscape, highlighting that this study serves as a snapshot, providing a foundation for proposals for future research. Approximately 50% of the publications are conference

papers, while around 40% are journal articles, with the remaining contributions comprising reviews and editorials.

3.2. Analytical procedure

In this paper, we adopt a broad perspective to investigate the relationship between GAI and digital resilience. We do not view the phenomenon under study as solely "technical," even though technical artifacts and algorithms are integral to the ways GAI manifests. To provide an analytical framework for understanding the broader consequences of these technologies, we engage with three analytical levels: technology, organization, and governance (adapted from, Johansson and Eriksson 2024; see also, Tornatzky and Fleischer 1990):

- **Technology.** Modern technologies such as GAI are manifested through artifacts, hardware, and software (algorithms) in our surroundings.
- **Organizations.** Organizations represent domains where technology is developed and used and are also critical arenas for ensuring the responsible development of AI systems.
- **Governance.** Institutions embody formal (e.g., legislation and rules) and informal (e.g., norms, values, routines) methods of governing technology. They also include cognitive dimensions that shape how we interpret and learn social interactions (Scott 2013).

The papers were initially categorized into three groups based on their focus (technology, organization, governance). We then conducted a thematic analysis within each group. To ensure transparency and rigor in the coding and analysis of the the selected papers, we employed the Gioia methodology (Gioia, Corley, and Hamilton 2013). This methodology was chosen because it allows for the generation of theoretical abstractions from initial observations and has been effectively used in previous literature reviews. In line with this approach, we first extracted first-order codes from the papers, then identified second-order themes based on common features within these codes. Finally, these themes were aggregated into higher-level constructs, aiming to encapsulate the overall insights from the literature. This process required several iterations, and some papers were challenging to categorize into a single theme. While the papers under the "technology" group were rather straight-forward to categorize, as their scope was relatively distinct, many papers addressed both organizational and governance issues. Thus, as expected from our qualitative approach, there were inevitable overlaps between some groups/themes/papers. To solve ambiguities and achieve intercoder agreement we engaged in a "negotiated agreement" where "researchers code a transcript, compare codings, and then discuss their disagreements in an effort to reconcile them and arrive at a final version in which as many discrepancies as possible have been resolved." (Campbell et al. 2013, p. 305). Here, both authors took turns categorizing the papers, and iteratively generating themes. The insights into the characteristics of GAI technology and its implications for society were organized into the three perspectives: technology (86 papers), organization (25 papers), and governance (53 papers), each with the associated themes, highlighting both potential threats and benefits.

4. Results

4.1. Technology

This section summarizes key findings regarding how GAI technologies (e.g. LLMs) can present both threats and opportunities in relation to digital resilience (see Table 3). The articles categorized under this theme primarily focus on the characteristics of the technology itself. Most of these publications are from the fields of computer science and engineering.

Table 3. Technology.

Theme	Description
Generation and detection of adversarial content	Exploration of methods for generating and detecting AI-generated "deepfakes," disinformation, and toxic content.
Cybersecurity	Examination of language models as both potential threats and defenses against cyberattacks.
Inherent vulnerabilities	Discussion of vulnerabilities arising from model training data and the inherent ambivalence in the content generated by AI systems (e.g. "hallucinations").

4.1.1. Theme 1: adversarial content generation

Using LLM-based chatbots in sensitive domains presents new challenges due to the unpredictable nature of these models (Başar et al. 2022). Although many LLMs are equipped with safety measures to mitigate harmful outputs, there are ways to manipulate them, through training and prompting, to generate "adversarial texts" (Crothers, Japkowicz, and Viktor 2023; Dong et al. 2023; Wan et al. 2023; You, Hammoudeh, and Lowd 2023; Mumuni and Mumuni 2024). Scholars have proposed several strategies to prevent harmful outputs from LLMs (Kandpal, Wallace, and Raffel 2022; Deng et al. 2023; Ha, Tran, and Kim 2023; Prabhumoye et al. 2023; Vashishtha et al. 2023; Yu 2023; Chen, Paliwal, and Yan 2023a; Chen et al. 2023b; Chen, He, and Wu 2024).

As noted by Akhtar (2023), there is an "arms race" between the generation and detection of deepfakes and disinformation (see also Mirsky and Lee 2021; Naitali et al. 2023; Wang et al. 2023). Researchers have developed tools to detect various forms of AI-generated content, including fake news (Bruno et al. 2023; Aman 2024; Patil et al. 2024), fake social media profiles (Ayoobi, Shahriar, and Mukherjee 2023), hate speech (Gupta 2022; Luo et al. 2023; Roy et al. 2023), AI-generated images (Das et al. 2023; Gupta, Raja, et al. 2023;; Khan and Malik 2023), videos (Gupta, Raja, et al. 2023), audio (Khan and Malik 2023; Yu, Zhai, and Zhang 2023), text (Liyanage and Buscaldi 2023; Pan 2023), "internet trolls" (Ezzeddine et al. 2023), and disinformation (Faye et al. 2023; Pan et al. 2023; Paraschiv and Dascalu 2023; Hamed, Zachara-Szymanska, and Wu 2024).

4.1.2. Theme 2: cybersecurity

In their paper "From ChatGPT to ThreatGPT," Gupta, Akiri, et al. (2023) highlight how LLMs are vulnerable to jailbreaking, reverse psychology, and prompt injection attacks. In addition to being susceptible to manipulation *via* prompts, LLMs can be used in cyberattacks to conduct social engineering, phishing, hacking, and the generation of payloads and malware. However, these models can also be used to protect against such attacks through automation of cyberattack detection, malware identification, system defense measures, incident reporting, and threat intelligence. They can also support the development of secure code, ethical guidelines, and incident response plans. As such, LLMs are double-edged swords, capable of both defending against cybersecurity threats and introducing new risks (Hu and Chen 2023; Taveekitworachai et al. 2023; Bezzi 2024; Ozkan-Okay et al. 2024), such as enabling "phishing" attacks (Begou et al. 2023).

Alwahedi et al. (2024) suggest that GAI and LLMs can enhance cyber threat detection due to their ability to understand language and process large volumes of unstructured data. They can be employed to analyze network traffic and logs, reinforce encryption, improve authentication and access control, support blockchain and IoT-based security systems, and assist with penetration testing (see also Ferrag, Debbah, et al. 2023:1; 2023:2). The context awareness of such models is particularly beneficial in protecting critical infrastructure, including so-called safety-critical systems (Diemert and Weber 2023).

Further studies have demonstrated how LLMs can be used to automatically generate malware (Botacin 2023; Chatzoglou et al. 2023; Pa Pa et al. 2023), as well as to strengthen countermeasures like malicious webpage detection (Li and Gong 2023), email scam decoying

(Cambiaso and Caviglione 2023), penetration testing (Happe and Cito 2023), and the repair or generation of secure code (He and Vechev 2023; Nair, Sadhukhan, and Mukhopadhyay 2023; Said Ozturk et al. 2023; Pearce et al. 2023), including automatic "self-healing" code (Khlaisamniang et al. 2023; see also Zhang et al. 2023). Moreover, GAI can generate synthetic network traffic for cybersecurity training (Khosh Kholgh and Kostakos 2023), detect vulnerabilities in programming code (Bakhshandeh et al. 2023; Jain et al. 2023; Liu, Deng, et al. 2023; Purba et al. 2023; Tihanyi et al. 2023; Chen et al. 2023c), test vulnerabilities on websites (Gallus et al. 2023), and create "honeypots" to engage and detect attackers (Ragsdale and Boppana 2023).

4.1.3. Theme 3: inherent vulnerabilities

Large language models are vulnerable to piracy and redistribution. Scholars have proposed "watermarking" as a method to address these issues by "labeling" the content generated by a model to protect intellectual property (Dai et al. 2024). These models have faced criticism for scraping vast amounts of data from the internet without proper consideration for privacy or copyright concerns (Kshetri 2023). Additionally, LLMs are susceptible to "training data extraction attacks" (Carlini et al. 2021; Li, Guo, et al. 2023; Subramani et al. 2023), which pose serious threats to privacy. Recent research has introduced countermeasures to prevent the extraction of sensitive data used to train these models (Behnia et al. 2022; Safa Ozdayi et al. 2023). On a more positive note, Agarwal, Chan, and Pattabiraman (2023) highlight that LLMs possess an advantage in terms of their high fault tolerance, making them less sensitive to transient hardware failures, such as those caused by solar storms or power disruptions.

Like other AI systems, GAI systems are sensitive to biases in their training data and lack transparency. These systems are also unaware of the limits of their own knowledge. As a result, in addition to exhibiting prejudice and discriminatory behaviors, they can generate *hallucinations*, "which manifest themselves in confidently generated results that seem plausible but are unreasonable with respect to the source of information" (Banh and Strobel 2023, p. 63; see also Fayyad 2023; Pinter and Elhadad 2023). Acerbi and Stubbersfield (2023) note that LLMs display patterns similar to those of humans in their outputs, including biases related to gender, social groups, and the type of messages conveyed, particularly when used for summarizing content (see also Fields, Chovanec, and Madiraju 2024). However, LLMs can also be used for content moderation to mitigate such biases (Franco, Gaggi, and Palazzi 2023).

The literature presents various methods for preventing hallucinations in LLMs (Ramakrishna et al. 2023), including iterative prompting (Jha et al. 2023), aligning training data (Semnani et al. 2023), and involving humans to assess the outputs (Ji et al. 2023). Hammond and Leake (2023) argue that LLMs need further development, particularly with the integration of higher levels of "symbolic AI," to enhance their reasoning capabilities.

4.2. Organization

This section focuses on the use and implications of GAI in organizational contexts. Table 4 provides an overview of the themes covered in this section.

Table 4. Organization.

Theme	Description
Knowledge management and innovation	Considerations necessary for using GAI in organizations to create value and facilitate innovation.
Organizational policy	Policies to ensure proper use of GAI.
Human alignment	Efforts to align AI systems with important ethical values (e.g. transparency, equality, accountability, justice).

4.2.1. Theme 4: knowledge management and innovation

Both private and public organizations are increasingly adopting GAI. The literature highlights strategies for addressing the challenges associated with these systems (see Section 4.1) and for creating organizational structures that enable innovation and learning. As noted by Alavi, Leidner, and Mousavi (2024), GAI introduces several organizational implications for knowledge management. While the technology can enhance information processing and improve organizational learning, it also presents risks such as biases, reduced social interactions and cognitive efforts, and the marginalization of certain groups. Iqbal (2023) discusses a shift from traditional knowledge management to intelligence engineering, where LLMs are fine-tuned on organizational data to create organizational value. These authors suggest that using open-source LLMs can mitigate privacy and security concerns (see also Culotta and Mattei 2024). Moreover, LLMs can help generate value from open government data (Loukis et al. 2023), and with proper prompt engineering, GAI tools can support regulatory compliance (Ioannidis et al. 2023).

Thus, it is important to consider the dual nature of GAI, acting both as a threat and as a facilitator of innovation (Cook, Hagiu, and Wright 2024). Managing GAI-driven innovation requires not only educating employees but also preparing for shifts in organizational power structures that accompany the implementation of new technologies (Holm et al. 2023).

Ebert and Louridas (2023) argue that tools such as ChatGPT, CoPilot, Tagnine, and transformers available on platforms like HuggingFace can greatly benefit the software development industry. Gupta, Sharma, et al. (2024) emphasize the need to address concerns related to the precision, security, and ethical implications of these technologies (see also Ozkaya 2023). Singla et al. (2023) demonstrate how LLMs can be used to analyze errors within the software development supply chain.

4.2.2. Theme 5: organizational policy

The use of LLMs and chatbots presents several challenges for organizations due to the unpredictable behavior of these systems (Başar et al. 2022), requiring new competencies among employees, such as the ability to mitigate potential negative effects. This may involve training LLM operators to ensure data privacy and helping LLM users develop effective prompting skills. The organizational changes introduced by GAI, therefore, require effective policymaking within organizations (Milic-Frayling 2023; Pasupuleti, Vadapalli, and Mader 2023). Meanwhile, GAI systems can be leveraged to tailor corporate cybersecurity policies (McIntosh et al. 2023). However, implementing LLMs in organizations raises concerns about information and data security, as well as uncertainties regarding their outputs. These issues must be addressed through proper governance (O'Leary 2023). Managers and decision-makers must understand the properties, capabilities, limitations, and underlying mechanisms of LLMs to use them effectively for decision-making purposes, particularly in areas that impact people, such as recruitment, organizational strategy development, and operational process support (Chen et al. 2023d).

4.2.3. Theme 6: human alignment

While tools such as ChatGPT can perform a variety of functions—including generating insights into financial markets, improving customer service, generating financial summaries, forecasting, personalizing investment recommendations, and preventing fraud—these systems are also susceptible to biases, disinformation, privacy and legal issues, and negative effects on the workforce (Oppenlaender et al. 2023). Research in this area proposes several solutions, such as enhanced training of LLMs, human supervision, policy alignment, and collaboration with domain experts (Khan and Umer 2024), as well as the use of synthetic data (Tkachenko 2023) to align these systems with ethical principles. Numerous principles related to the development of "responsible AI" are also presented (Lu et al. 2023), aiming to balance technological progress with human involvement (Patel, Pandya, and Sojitra 2023). Edenberg and Wood (2023) argue that it is

insufficient for GAI and algorithms to be designed as "neutral" agents; instead, they should actively promote justice, equity, and accountability. Similarly, Kieseberg et al. (2023) propose various measures to achieve "Controllable AI," including detecting and managing control loss, ensuring transparency in AI decision-making, and addressing biases and backdoors (Akbar, Khan, and Liang 2023).

4.3. Governance

Generative AI is becoming an increasingly integral component of many complex, global systems, where its implementation does not neatly align with state borders. Researchers have raised concerns that GAI could undermine fundamental rights and existing laws (Favela and Amon 2023; Feldstein 2023). There are also indications of a growing imbalance between technological advancement and the institutional capacity to govern it, leading to increasing power asymmetries in society (Luitse and Denkena 2021; Rosengrün 2023), along with a lack of accountability in the use of AI in public administration (Transatlantic Reflection Group on Democracy and the Rule of Law in the Age of "Artificial Intelligence" 2023). The literature also highlights various issues related to intellectual property, privacy and data protection, market monopolies, cybercrime, the societal status of AI subjects, legal regulation, and ethical considerations (e.g. Li, Guo, et al. 2023; Qu, Bai, and Zhang 2023). Table 5 summarizes the themes covered in the following sections.

4.3.1. Theme 7: legitimacy, democracy, and humanity

The legitimacy of AI technology evolves in relation to the competence and trust society places in it. Literature on the social status of GAI spans a spectrum, from AI being seen as a mere, albeit sophisticated, tool, to AI as a competent and creative partner, and even as a quasi-magician (Bran et al. 2023; Fisher 2023). Studies have shown that media portrayals of AI are often sensationalized, oscillating between depictions of AI as a solution to major societal challenges and warnings of imminent, systemic dangers. This highlights the role of media discourse in shaping public attitudes toward AI technology (Roe and Perkins 2023). Furthermore, the conflation of science with ideological hegemony may impede global technological progress. For instance, in China, the convergence of "scientainment," patriotic education, and the perceived association of AI technology with political entities has fostered negative attitudes toward foreign tools and spurred conspiracy theories and nationalist sentiments (Zou and Liu, Deng, et al. 2023).

Venturini (2023) rejects the notion that machines like GAI are detached from their creators, arguing that these technologies only exist through the support of human assistants—developers, sellers, and users. In this light, the risks associated with GAI do not stem from an unlikely singularity, but rather from their embeddedness "in the dull and exploitative industry of the digital attention economy" (Venturini 2023, p. 101).

The analysis of GAI's impact on democracy has become an interdisciplinary research area, focusing on the intersections between AI and democratic values, such as self-rule, equality, elections, and competition between state systems (Jungherr 2023). Studies caution that the

Table 5. Governance.

Theme	Description
Legitimacy, democracy, and humanity	The impact of AI on democratic processes and legitimacy, addressing issues of ethics, equality, justice, and representativeness.
Governance, legislation, and regulation	Strategies to manage AI-related risks, regulate AI development and use, and ensure accountability.
Trust and power	Maintaining public trust in government and public administration and addressing concerns related to (inappropriate) power projection.

growing monopolization in the technology sector, with power concentrated in a few large companies, poses risks of societal disruption. For instance, the deployment of LLMs may reinforce hegemonic worldviews, marginalize vulnerable communities (Luitse and Denkena 2021), or even lead to an epistemic tipping point in the public infosphere (Nannini 2022). These shifts have been linked to asymmetric power structures driven by commercial interests related to GAI, which may undermine democracy by promoting the spread of disinformation and conspiracy theories (Rosengrün 2023). This misalignment, where essential societal values are overlooked or contradicted during the development of AI systems, has been recognized as a serious issue. It could lead to the permanent disempowerment of humanity, as such misalignments increase with the development of more capable AI systems (Dung 2023; Editorial Nature 2023). To mitigate these risks, some propose training LLMs in cultural norms and cognitive tasks, aiming to develop language agents that can are more aligned with human values (Goldstein & Kirk-Giannini 2023; Ramezani and Xu 2023).

The recognized potential of GAI to facilitate human rights violations has led researchers to call for intensified efforts to prioritize ethical discussions and promote cross-disciplinary collaboration in AI development and its responsible use (Favela and Amon 2023; Gromova, Ferreira, and Begishev 2023; Head et al. 2023; Kaffee et al. 2023; Ortega, Tran, and Bandeen 2023; Piñeiro-Martín et al. 2023; Horzyk 2024).

4.3.2. Theme 8: AI governance, legislation, and regulation

The governance of GAI is an expanding area of research as legislation surrounding these systems evolves, such as the European AI Act (EU 2024/16891689 2024). The AI Act is a legislative proposal from the European Union aimed at regulating AI use. It categorizes AI systems into four risk levels: unacceptable risk, high risk, limited risk, and minimal risk. AI systems deemed of unacceptable risk, such as state-run social scoring, are entirely banned. High-risk systems, used in critical areas like health and education, require stricter regulation and monitoring. Limited-risk systems must meet specific transparency requirements, while minimal-risk systems face few or no restrictions. The goal is to ensure that AI is used safely and ethically within the EU. However, the European AI Act has faced criticism for its inability to support comprehensive AI risk management and its failure to adequately address GAI-related risks (Elgesem 2023; Helberger and Diakopoulos 2023; Novelli et al. 2023; Rubinic et al. 2024;; Vahabava 2023; Wörsdörfer 2024).

There is a growing consensus in the literature that more comprehensive approaches are necessary to manage the expanding risk landscape associated with the use and misuse of AI technologies in society. Issues include privacy, copyrights, and the responsible use of publicly available information (Dermawan 2024;; Fenwick and Jurcys 2023). Additionally, research has explored the area between deception and negligence regarding consumer information provided by AI tools acting as corporate agents, to address issues of responsibility (Herbosch 2024). A subset of this literature discusses GAI in relation to open data. One empirical study argues that open data "will not be sufficient to mitigate the harmful consequences of thoughtless deployment of LLMs, nor the questionable copyright implications of scraping all publicly available data from the internet" (Liesenfeld, Lopez, and Dingemanse 2023, p. 4). However, the availability of open data can also facilitate original research and promote accountability for data use. Meanwhile, the scholarly discussion of the ethical, moral, and legal implications of AI and other technologies raises concerns about the responsibilities of developers and the IT industry for the abuse of their products in supporting financial fraud, identity theft, chaos, discord, disinformation, and hybrid warfare (Gradon 2023).

The literature has also explored various means of governing AI systems and their development. For instance, three regulatory mechanisms to govern GAI systems have been proposed: industrial observability, public inspectability, and technical modifiability (Ferrari, van Dijck, and van den Bosch 2023). Here, research advocates for global approaches, such as regulatory

sandboxes, to implement responsible oversight (Gonzalez Torres, Kajava, and Sawhney 2023). Other suggestions include that AI governance should aim to balance security and development, adhere to human-centered principles, and promote a comprehensive legal framework supplemented by specialized management approaches for GAI (Li, Guo, et al. 2023). A study examining European AI Liability Directives advocates for a more comprehensive approach to strengthen coordination and harmonize regulation, while also addressing enhancements related to non-discrimination, explainability, and sustainability issues, both within Europe and globally (Hacker 2023).

However, the regulations for GAI are fragmented worldwide. Only a few countries have issued AI-related regulations (Feldstein 2023; Gonzalez Torres, Kajava, and Sawhney 2023; Katrenčík, Mucha, and Zatrochová 2023). The literature also expresses concerns about competitive imbalances or the relocation of corporate operations to regions without regulations—effects that often follow enforcement actions by regulatory bodies and can undermine the effectiveness of national and supranational regulation (Park et al. 2023). Moreover, the broader use of GAI tools in international law, particularly in treaty interpretation, is expected to impact the dominant narratives in this context (Nelson 2023).

4.3.3. Theme 9: trust and power

Generative AI faces the same challenge as other AI systems in terms of explainability: it is difficult to understand why the models make certain decisions. This limitation hinders its use in public administration (Cantens 2024). However, the literature envisions a transformative impact on the intellectual production within public institutions, including the potential replacement or "enslavement" of officials by machines (Cantens 2024), as well as automation of tasks such as disease case examination, documentation, and decision-making (Sonntagbauer, Haar, and Kluge 2023).

Consequently, researchers emphasize the need for greater focus on training public officials to understand and use GAI in a nuanced manner, and to engage both public and private actors in ethical discussions (Nannini 2022; Favela and Amon 2023; Linegar, Kocielnik, and Alvarez 2023; Masotina, Musi, and Spagnolli 2023; Cantens 2024). For instance, research on official pro-vaccination COVID-19 messaging revealed that while AI (with human supervision) can generate effective public health messages, individuals prefer to receive such messages from human institutions rather than GAI sources (Karinshak et al. 2023). This suggests that public administration must be particularly careful and transparent when implementing GAI technology to maintain public trust.

Moreover, studies have shown how GAI can support public policy processes related to climate change by monitoring technological innovation in society. This research illustrates the potential for enhanced information while also highlighting challenges such as data scarcity, algorithmic interpretability, and the need for cross-disciplinary collaboration (Toetzke, Probst, and Feuerriegel 2023).

The advent of deepfakes to spread disinformation, is a significant threat to the trust placed in public organizations and government information. Studies have investigated how deceptive AI can be understood (Sarkadi 2023) and have emphasized the need for a holistic approach that integrates technical solutions with public awareness, and legislative action. This approach is essential to mitigate the potential for deepfakes to severely distort political discourse, destabilize markets, and escalate disinformation in online discussions (Mubarak et al. 2023). Here, Barbaro and Skumanich (2023) highlights the impact of disinformation, cyberbullying, and hate speech, all amplified by GAI and spread across unmoderated information-sharing platforms, on societal stability. As AI technology evolves, it can enable an enhanced form of information dissemination that is adaptive, interactive in real-time, and can be personalized for any purpose and targeted user. This evolution can be exploited to exert undue influence and manipulate

public sentiment (Rosenberg 2023). Conversely, GAI is also envisioned to support public crisis preparedness by enabling the simulation and training of potential crisis scenarios (Karinshak 2024).

At the state level, GAI has been recognized as a tool that can support authoritarian governments by advancing propaganda and surveillance efforts, further strengthening the spyware and cybercrime market (Feldstein 2023). Therefore, governments must establish common standards for the ethical and responsible use of AI tools, including in military operations and defense (O'Connell 2023; Rubinic et al. 2024).

5. Discussion

In this chapter, we discuss potential areas for future research in the intersection between digital resilience and AI risk management, based on the findings. Specifically, our coding of the materials identified three aggregated themes, as illustrated in Figure 1, which are further elaborated upon in the following sections.

5.1. Generative AI as a cat-and-mouse game

The findings from the literature within the "technology" perspective illustrate a cat-and-mouse dynamic, where the technology driving AI-generated content continuously evolves, while researchers simultaneously develop methods to detect and mitigate this content, such as deepfakes. This dual-use nature of GAI has led to research on how it can both facilitate and help combat disinformation and cyberattacks.

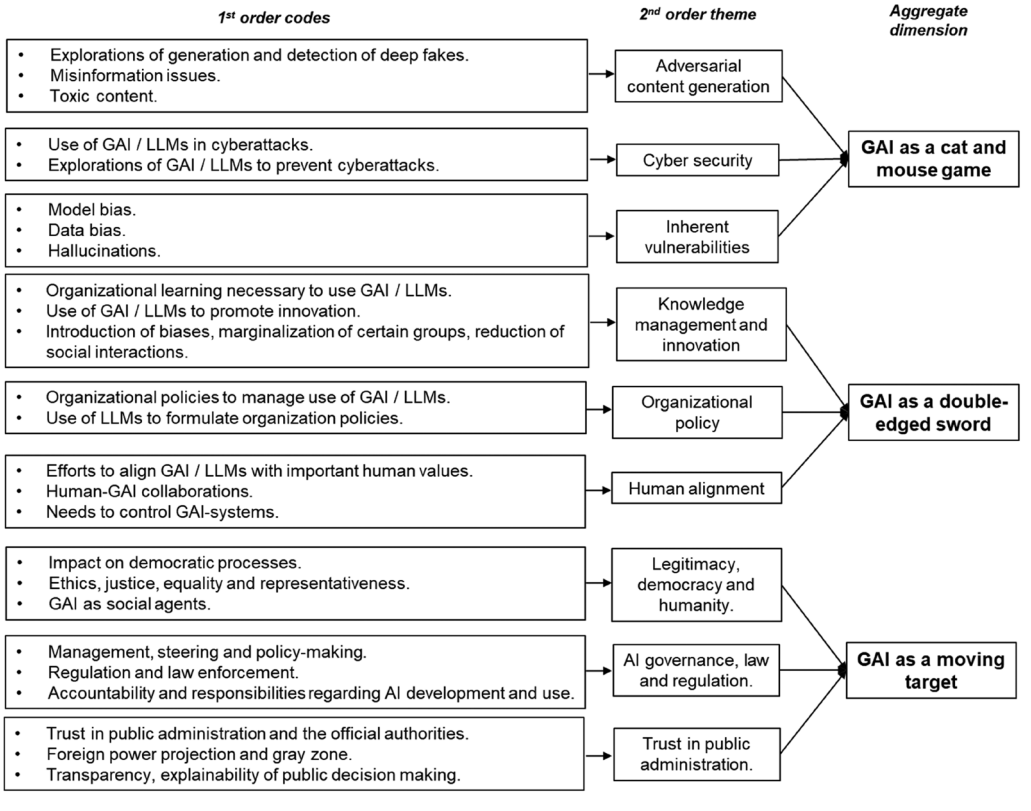

Figure 1. Coding and themes.

Two key problem areas emerged in the literature: first, the inherent vulnerabilities in LLMs that make them susceptible to various types of (manual and automated) attacks; and second, the intentional use of GAI to produce harmful content, such as malware, toxic material, and disinformation.

It is important to note that tools like ChatGPT are designed with built-in guardrails to prevent the generation of harmful or inappropriate content, whereas uncensored language models lack these safeguards. Importantly, GAI systems designed to detect harmful content are trained using such content. One example of this is the use of uncensored models to generate "synthetic" training data, such as for training models to identify phishing attempts. In other words: it takes one to know one.

Meanwhile, new risks introduced by these tools, particularly in relation to digital resilience and cybersecurity, must be acknowledged. These models are built on existing data, and while LLMs may make harmful tools more accessible to a broader range of users, cybersecurity should be viewed as part of a larger, interconnected ecosystem.

In sum, there is an ongoing "cat-and-mouse" game between attackers and defenders, where GAI is used both to generate and identify harmful content and cyber-attacks. Beyond technical considerations, it is essential to address how human behavior evolves as these technologies proliferate, highlighting the growing importance of interdisciplinary approaches to advance digital resilience.

5.2. Generative AI as a double-edged sword

Generative AI is recognized for its potential to enhance competitiveness and drive innovation, but researchers also caution about the risks associated with its misuse. Organizations are integrating GAI into their operations, and these environments are where people will both benefit from and be exposed to these tools. For instance, organizations are employing techniques like retrieval-augmented generation ("RAG"), which enhances the accuracy of language models by cross-referencing them with external sources (see, Lewis et al. 2020), to mitigate risks like hallucinations. Looking at the literature as a whole, GAI is framed as a double-edged sword, where organizations can drive innovation but must also develop new competencies and implement robust measures to ensure these models align with ethical standards and guidelines.

One edge of the sword points at the perceived need to accelerate the adoption of GAI to secure potential business benefits, while the other edge of the sword relates to the various costs of such innovation, including technology, employee training, and considerations of appropriate use and for what tasks and purposes. Hence, it is important to understand the organizational value creation processes and if and how GAI can outperform the benefits compared to the costs.

In addition, the digital ecosystem in organisations is difficult to oversee and manage, as AI technologies are increasingly incorporated in widely used business software. These preconditions also reinforce the double-edged nature of GAI: while it becomes easier to use, it is not always clear to the user that AI is being used, which in turn also challenges the development of meaningful policies and guideline for GAI use in organizations.

In sum, GAI constitutes a double-edged sword for organizations, as these technologies are presented as facilitators for innovation and growth, while also introducing several disruptions, threats and uncertainties. An important avenue for future research on digital resilience is to explore the conditions necessary to establish principles and best practices that guide the ethical development and deployment of GAI systems in organizations.

5.3. Generative AI as a moving target

Researchers highlight issues like discrimination, manipulation of democratic processes, and the destabilization of public opinion in the wake of GAI use. The rapid evolution of GAI has sparked

a polarized debate, with AI being perceived as both a utopian advancement and a dystopian threat.

A critical point in the literature challenges the perception of GAI as "neutral," "objective," or "independent." Instead, it emphasizes that GAI is a technology that needs to play a role in promoting justice, equality, and accountability in society. Moreover, efforts toward global governance of these systems, based on ethical principles, must consider the diverse norms and values present across different regions of the world.

As indicated, governments are increasingly recognising the risks associated with GAI, leading to the development of new policies and regulations. However, monitoring compliance with national and international regulations is a complex endeavour and requires adapted methodologies and a better understanding of how GAI is and should be embedded in digital services used on a regular basis.

The findings suggest that policy makers must develop a broad understanding of AI and respond to challenges relating to upholding public trust. As the growing monopolization in the technology sector leads to power concentrated in the hand of a few large companies, the risks of societal disruption need enhanced attention. The findings also highlight emerging risks associated with the use of GAI for the purpose of inappropriate foreign power projection. Officials involved in civil preparedness and digital resilience need therefore to consider inter-organizational and societal perspectives, as the impact of AI transcends traditional administrative boundaries.

In sum, the rapid evolution of GAI constitutes a moving target for governments, which makes it challenging to adapt countermeasures and regulations to address the potential negative effect of these technologies. However, the literature addressing the ethical, organizational, and accountability challenges governments face with GAI remains limited, highlighting a critical gap in digital resilience research that warrants further exploration.

5.4. Proposals for future research

From the initial literature search in early January 2024 to the completion of the analysis in June 2024 and this article in November 2024, the number of publications identified through the search string grew from 773 to 1,364 and 3,580, respectively, highlighting the rapid growth of this field and the need for continued attention. Based on the findings presented in the previous sections, this study proposes ten future research avenues. Table 6 summarizes the focus areas, with a brief motivation and a suggested RQ for each. These questions are broadly framed to offer an umbrella for diverse studies, employing various perspectives, research methods, and theoretical frameworks, and exploring interconnected issues.

6. Conclusions

The objective of this study was to consolidate knowledge on GAI and its implications for society's digital resilience by examining both the benefits and risks associated with GAI. Additionally, the study aimed to identify key knowledge gaps, providing a foundation for future research. The literature review was guided by the following RQ:

> How can generative AI contribute to or counteract digital resilience?

This study sought to enhance understanding of digital resilience in an era of increasing GAI use. The swift pace of technological development and research presents important challenges for individuals, organizations, governance, and policymakers. This systematic literature review provides an overview of several challenges that require further investigation. Based on the findings, nine key themes were identified where GAI either introduces new risks or

Table 6. Proposals for future research.

Future Research Proposal	Motivation and Example Research Questions
Moving from the lab to reality	Much of the research on GAI is conducted in controlled, experimental settings. Empirical studies of its practical applications, especially in areas like cybersecurity, are needed. *RQ: What are the lessons learned from real-world applications of GAI?*
Outlining new risks posed by GAI	While the literature outlines potential benefits and threats, case studies on the tangible impacts of GAI on digital resilience are sparse. *RQ: What are the risks accompanying GAI, and how can they be understood, evaluated, and mitigated?*
Studying a society where anyone can generate anything	Research has largely focused on the technology itself. There is a need for interdisciplinary studies examining the societal implications of living in an AI-mediated world. *RQ: What are the societal impacts of increased GAI use, particularly on social media, and how can they be analyzed and understood?*
Understanding organizations' role in resilient GAI use	Organizations are the arena where GAI will be implemented and use. These are also settings where important principles of digital resilience are to be enacted. *RQ: How can organizational GAI be used to uphold principles such as justice, equality, and responsibility?*
Improving leadership and decision-making	Organizations are key to the implementation of GAI, and these settings are where important principles of digital resilience must be enacted. *RQ: What competencies do leaders and decision-makers need to ensure the safe and responsible use of GAI*?
Intra- and interorganizational AI governance	The effects of AI usage extend beyond organizational and national borders, necessitating new forms of cooperation across private and public sectors. *RQ: How can collaboration and coordination be designed to address the systemic risks posed by AI use?*
Trust in the public sector	Public organizations must not only develop a broad understanding of AI but also address both technical and organizational challenges to uphold public trust. *RQ: How can the management and implementation of GAI in public administration be organized responsibly to preserve trust in institutions?*
Enactment of legal requirements	New laws and regulations, such as the EU's AI Act, will have a profound impact globally. Research is needed to examine how these laws and regulations are enacted within organizations. *RQ: How can providers and users be supported in applying pan-national laws, such as the EU's AI Act?*
Studying the relationship between disinformation and trust	The public sector faces challenges in maintaining public trust, especially during activities such as crisis communication, in an era of widespread disinformation. *RQ: How can the public sector manage risk and crisis communication in an era where discerning truth from falsehood is increasingly difficult?*
Preventing inappropriate exertion of influence and power.	There is a broad "hybrid" zone between peace and war where disinformation and misinformation are used to destabilize and manipulate. *RQ: How can hidden power projections be made visible, focusing on actors' goals, the credibility of information sources, and ethical norms?*

amplifies existing challenges, warranting continued research. As indicated by the results, GAI accelerates a growing arms race between attackers and defenders of digital systems, with both attacks and defenses increasingly automated. A particularly challenging aspect is the sheer volume of data, information, and disinformation that must be filtered, processed, and understood. The surge in information, driven by automated content creation and distribution, parallels a rise in automated content monitoring, leading to a "cat-and-mouse" game in which algorithms are created to detect content generated by other algorithms. It is thus essential to recognize the dual-use nature of GAI. While GAI can serve multiple purposes, the benefits for one stakeholder may not necessarily align with the interests of others. For instance, research indicates that LLMs, trained on existing data, often disadvantage minorities and reinforce existing social norms. Beyond issues of bias, there is the risk of deliberate misuse, such as using GAI to exert power and destabilize societies, particularly in the context of gray-zone or hybrid warfare. Therefore, further research is needed to understand GAI as a

moving target, exploring how its development can be managed to protect democratic values and global sustainability goals. This includes examining governance and collaboration at the organizational level (within companies and government agencies), and at the societal level, considering the interplay of individual, local, national, and global interests. Meanwhile, it is essential to address ethical, regulatory, delegation, and accountability issues, along with risk management and conflict resolution strategies.

Much of the existing research has primarily focused on technological development, AI model training, and applied examples. However, many of the challenges identified in this study are sociotechnical in nature. Therefore, we advocate for more interdisciplinary studies that address these challenges and examine the underlying issues where GAI will play an amplifying role. This study highlights the potential implications of GAI and related technologies for society's digital resilience and ultimately proposes ten directions for future research. The development of GAI may necessitate a renewed emphasis on human dialogue and interaction to preserve public trust, countering the temptation for government agencies to rely on GAI for information generation. Future initiatives should strive to help society cultivate a more nuanced understanding of AI, moving beyond utopian and dystopian perspectives, and encouraging a more thoughtful and refined discourse around GAI.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Myndigheten för Samhällsskydd och Beredskap.

References

Andersson, S. 2023. "Problems in Information Classification: Insights from Practice." *Information & Computer Security* 31 (4): 449–462. https://doi.org/10.1108/ICS-10-2022-0163.

Bartz-Beielstein, T. 2019. "Why we Need an AI-Resilient Society." https://doi.org/10.48550/arXiv.1912.08786, https://arxiv.org/pdf/1912.08786

Berente, N., B. Gu, J. Recker, and R. Santhanam. 2021. "Managing Artificial Intelligence." *MIS Quarterly* 45 (3): 1433–1450.

Boh, W., P. Constantinides, B. Padmanabhan, and S. Viswanathan. 2023. "Building Digital Resilience against Major Shocks." *MIS Quarterly* 47 (1): 343–360.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, … D. Amodei. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901.

Buzzanell, P. M. 2023. "Risk, Resilience, and Ethical Considerations in Artificial Intelligence." *Emerging Media*. 1(1), 30-39. https://doi.org/10.1177/27523543231188274

Campbell, J. L., C. Quincy, J. Osserman, and O. K. Pedersen. 2013. "Coding in-Depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement." *Sociological Methods & Research* 42 (3): 294–320. https://doi.org/10.1177/0049124113500475.

Chang, A. X., T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, … F. Yu. 2015. "Shapenet: An Information-Rich 3D Model Repository." https://doi.org/10.48550/arXiv.1512.03012, https://arxiv.org/pdf/1512.03012

Cramer, J. S. 2002. The Origins of Logistic Regression. Tinbergen Institute Working Paper No. 2002-119/4, Accessed Juni 2024, from SSRN: https://ssrn.com/abstract=360300 or https://doi.org/10.2139/ssrn.360300.

Cuel, R., D. Ponte, and F. Virili. 2022. *Exploring Digital Resilience*. Berlin/Heidelberg, Germany: Springer.

Dataconomy. 2022. Accessed Juni 2024. https://dataconomy.com/2022/09/26/ai-artwork-wins-art-competition/.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." https://doi.org/10.48550/arXiv.1810.04805, https://arxiv.org/pdf/

Duan, Y., J. S. Edwards, and Y. K. Dwivedi. 2019. "Artificial Intelligence for Decision Making in the Era of Big Data–Evolution, Challenges and Research Agenda." *International Journal of Information Management* 48: 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021.

Dwivedi, Y. K., N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, … R. Wright. 2023. "So What If ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy." *International Journal of Information Management* 71: 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.

EU 2024/1689. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Accessed May 31, 2024. Accessible online: Tillgänglig online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Feuerriegel, S., R. DiResta, J. A. Goldstein, S. Kumar, P. Lorenz-Spreen, M. Tomz, and N. Pröllochs. 2023. "Research Can Help to Tackle AI-Generated Disinformation." *Nature Human Behaviour* 7 (11): 1818–1821. https://doi.org/10.1038/s41562-023-01726-2.

Galton, F. 1889. *Natural Inheritance*. Vol. 42. New York:Macmillan.

Gill, T. G. 1995. "Early Expert Systems: Where Are They Now?" *MIS Quarterly* 19 (1): 51–81. https://doi.org/10.2307/249711.

Gioia, D. A., K. G. Corley, and A. L. Hamilton. 2013. "Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology." *Organizational Research Methods* 16 (1): 15–31. https://doi.org/10.1177/1094428112452151.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, … Y. Bengio. 2020. "Generative Adversarial Networks." *Communications of the ACM* 63 (11): 139–144. https://doi.org/10.1145/3422622.

Große, C. 2019. "Sources of Uncertainty in Swedish Emergency Response Planning." *Journal of Risk Research* 22 (6): 758–772. https://doi.org/10.1080/13669877.2018.1459796.

Große, C. 2022. "A Glimpse of Sustainability Culture? Reflecting the Concept of Sustainability in Society's Resilience." In *Proceedings of the 32nd European Safety and Reliability Conference (ESREL)*, 165–172. Singapore: Research Publishing.

Große, C. 2023. "Hybrid and Information Warfare: Challenging Topics for Risk Communication." In *Proceedings of the 33rd European Safety and Reliability Conference (ESREL)*. Singapore: Research Publishing, 1999–2006.

Günther, W. A., M. H. R. Mehrizi, M. Huysman, and F. Feldberg. 2017. "Debating Big Data: A Literature Review on Realizing Value from Big Data." *The Journal of Strategic Information Systems* 26 (3): 191–209. https://doi.org/10.1016/j.jsis.2017.07.003.

Handmer, J. W., and S. R. Dovers. 1996. "A Typology of Resilience: Rethinking Institutions for Sustainable Development." *Industrial & Environmental Crisis Quarterly* 9 (4): 482–511. https://doi.org/10.1177/108602669600900403.

Holling, C. S. 1973. "Resilience and Stability of Ecological Systems." *Annual Review of Ecology and Systematics* 4 (1): 1–23. https://doi.org/10.1146/annurev.es.04.110173.000245.

Johansson, B. J., and P. Eriksson. 2024. "A Maturity Model to Guide Inter-Organisational Crisis Management and Response Exercises." *International Journal of Disaster Risk Reduction* 106: 104413. https://doi.org/10.1016/j.ijdrr.2024.104413.

Jones, M. 2019. "What we Talk about When we Talk about (Big) Data." *The Journal of Strategic Information Systems* 28 (1): 3–16. https://doi.org/10.1016/j.jsis.2018.10.005.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine Learning: Trends, Perspectives, and Prospects." *Science (New York, N.Y.)* 349 (6245): 255–260. https://doi.org/10.1126/science.aaa8415.

Karnouskos, Stamatis. 2020. "Artificial Intelligence in Digital Media: The Era of Deepfakes." *IEEE Transactions on Technology and Society* 1 (3): 138–147. https://doi.org/10.1109/TTS.2020.3001312.

Kelley, H. J. 1960. "Gradient Theory of Optimal Flight Paths." *ARS Journal* 30 (10): 947–954. https://doi.org/10.2514/8.5282.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2017. "Imagenet Classification with Deep Convolutional Neural Networks." *Communications of the ACM* 60 (6): 84–90. https://doi.org/10.1145/3065386.

Liu, Y., G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, … Y. Liu. 2023. "Jailbreaking Chatgpt via Prompt Engineering: An Empirical Study." https://doi.org/10.48550/arXiv, https://arxiv.org/pdf/2305.13860

LeCun, Y., B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. 1989. "Handwritten Digit Recognition with a Back-Propagation Network." *Advances in Neural Information Processing Systems* 2: 398–404.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, … D. Kiela. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks." *Advances in Neural Information Processing Systems* 33: 9459–9474.

Lycett, M. 2013. "Datafication': making Sense of (Big) Data in a Complex World." *European Journal of Information Systems* 22 (4): 381–386. https://doi.org/10.1057/ejis.2013.10.

McKim, J. 2022. "2. Deep Learning the City: The Spatial Imaginaries of AI." In *Seeing the City Digitally*, 35–56. Amsterdam: Amsterdam University Press.

Miles, L., and E. Petridou. 2016. "Entrepreneurial Resilience: Role of Policy Entrepreneurship in the Political Perspective of Crisis Management." In *Organisational Resilience: Concepts, Integration and Practice*, 67–81. Boca Raton, FL: CRC Press.

Kendra, J. M., and T. Wachtendorf. 2003. "Elements of Resilience after the World Trade Center Disaster: Reconstituting New York City's Emergency Operations Centre." *Disasters* 27 (1): 37–53. https://doi.org/10.1111/1467-7717.00218.
Ngwenyama, O., H. Z. Henriksen, and D. Hardt. 2023. "Public Management Challenges in the Digital Risk Society: A Critical Analysis of the Public Debate on Implementation of the Danish NemID." *European Journal of Information Systems* 32 (2): 108–126. https://doi.org/10.1080/0960085X.2021.1907234.
OpenAI. 2023. GPT-4 Technical Report. accessed September 2024. Arxiv.org. https://arxiv.org/abs/2303.08774.
Paré, G., M. Trudel, M. Jaana, and S. Kitsiou. 2015. "Synthesizing Information Systems Knowledge. A Typology of Literature Reviews." *Information & Management* 52 (2): 183–199. https://doi.org/10.1016/j.im.2014.08.008.
Paul, J., and J. Menzies. 2023. "Developing Classic Systematic Literature Reviews to Advance Knowledge: Dos and Don'ts." *European Management Journal* 41 (6): 815–820. https://doi.org/10.1016/j.emj.2023.11.006.
Ponce Del Castillo, A. 2023. "Generative AI, Generating Precariousness for Workers? *AI & Society* 39: 1–2.
Reuters. 2023. ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note. Accessed May 2024. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.
Rose, A. 2007. "Economic Resilience to Natural and Man-Made Disasters: Multidisciplinary Origins and Contextual Dimensions." *Environmental Hazards* 7 (4): 383–398. https://doi.org/10.1016/j.envhaz.2007.10.001.
Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533–536. https://doi.org/10.1038/323533a0.
Russell, S. J., and P. Norvig. 2022. *Artificial Intelligence: A Modern Approach*. Harlow: Pearson Education, Inc.
Sadowski, J., and M. Andrejevic. 2020. "More than a Few Bad Apps." *Nature Machine Intelligence* 2 (11): 655–657. https://doi.org/10.1038/s42256-020-00246-2.
Salehi, V., B. Veitch, and M. Musharraf. 2020. "Measuring and Improving Adaptive capacity in resilient Systems by Means of an Integrated DEAMachine Learning Approach." *Applied Ergonomics* 82: 102975. https://doi.org/10.1016/j.apergo.2019.102975.
Schintler, L. A., and C. L. McNeely. 2022. "Artificial Intelligence, Institutions, and Resilience: Prospects and Provocations for Cities." *Journal of Urban Management* 11 (2): 256–268. https://doi.org/10.1016/j.jum.2022.05.004.
Schneider, J., C. Meske, and P. Kuss. 2024. "Foundation Models: A New Paradigm for Artificial Intelligence." *Business & Information Systems Engineering*: 66: 221–231. https://doi.org/10.1007/s12599-024-00851-0.
Scott, W. R. 2013. *Institutions and Organizations: Ideas, Interests, and Identities*. Stanford: Sage Publications.
Shandilya, S. K., A. Datta, Y. Kartik, and A. Nagar. 2024. "What Is Digital Resilience?" In *Digital Resilience: Navigating Disruption and Safeguarding Data Privacy. EAI/Springer Innovations in Communication and Computing*. Cham: Springer.
Sundberg, L. 2024. "Towards the Digital Risk Society: A Review." *Human Affairs* 34 (1): 151–164. https://doi.org/10.1515/humaff-2023-0057.
Sundberg, L., and J. Holmström. 2024. "Innovating by Prompting: How to Facilitate Innovation in the Age of Generative AI." *Business Horizons* 67 (5): 561–570. https://doi.org/10.1016/j.bushor.2024.04.014.
Thickstun, J., Z. Harchaoui, and S. Kakade. 2016. "Learning Features of Music from Scratch." https://doi.org/10.48550/arXiv., https://arxiv.org/pdf/1611.09827
Tim, Y., and D. E. Leidner. 2023. "Digital Resilience: A Conceptual Framework for Information Systems Research." *Journal of the Association for Information Systems* 24 (5): 1184–1198. https://doi.org/10.17705/1jais.00842.
Tornatzky, L. G., and M. Fleischer. 1990. *The Processes of Technological Innovation*. Lexington, MA: Lexington Books.
Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, … I. Polosukhin. 2017. "Attention is All You Need." *Advances in Neural Information Processing Systems* 30: 6000–6010.
Vom Brocke, J., A. Simons, B. Niehaves, K. Reimer, R. Plattfaut, and A. Cleven. 2009. Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process." European Conference on Information Systems (ECIS) Proceedings, Paper 161.17th European Conference on Information Systems (ECIS 2009), Verona, Italy, June 8–10.
World Economic Forum. 2023. Global Risks Report 2023. accessed Juni 2024. https://www.weforum.org/publications/global-risks-report-2023/.

Appendix A. Papers from the literature review

Acerbi, A., and J. M. Stubbersfield. 2023. "Large Language Models Show Human-like Content Biases in Transmission Chain Experiments." *Proceedings of the National Academy of Sciences of the United States of America* 120 (44): e2313790120. (https://doi.org/10.1073/pnas.2313790120.
Agarwal, U. K., A. Chan, and K. Pattabiraman. 2023. "Resilience Assessment of Large Language Models under Transient Hardware Faults." Proceedings - International Symposium on Software Reliability Engineering, ISSRE.
Akbar, M. A., A. A. Khan, and P. Liang. 2023. "Ethical Aspects of ChatGPT in Software Engineering Research." In IEEE Transactions on Artificial Intelligence, S. 1–14.
Akhtar, Z. 2023. "Deepfakes Generation and Detection: A Short Survey." *Journal of Imaging* 9 (1): 18. https://doi.org/10.3390/jimaging9010018.
Alavi, M., D. E. Leidner, and R. Mousavi. 2024. "Knowledge Management Perspective of Generative Artificial Intelligence." *Journal of the Association for Information Systems* 25 (1): 1–12. https://doi.org/10.17705/1jais.00859.

Alwahedi, F., A. Aldhaheri, M. A. Ferrag, A. Battah, and N. Tihanyi. 2024. "Machine Learning Techniques for IoT Security. Current Research and Future Vision with Generative AI and Large Language Models." *Internet of Things and Cyber-Physical Systems* 4: 167–185. https://doi.org/10.1016/j.iotcps.2023.12.003.

Aman, M. 2024. "Large Language Model Based Fake News Detection." *Procedia Computer Science* 231: 740–745.). https://doi.org/10.1016/j.procs.2023.12.144.

Ayoobi, N., S. Shahriar, and A. Mukherjee. 2023. "The Looming Threat of Fake and LLM-Generated LinkedIn Profiles." Challenges and Opportunities for Detection and Prevention. HT 2023 - The 34th ACM Conference on Hypertext and Social Media.

Bakhshandeh, A., A. Keramatfar, A. Norouzi, and M. M. Chekidehkhoun. 2023. "Using ChatGPT as a Static Application Security Testing Tool." *ISeCure* 15 (3):51–58.

Banh, L., and G. Strobel. 2023. "Generative Artificial Intelligence." *Electronic Markets* 33 (1): 63. https://doi.org/10.1007/s12525-023-00680-1.

Barbaro, F., and A. Skumanich. 2023. "On the Importance to Study Fringe Social Networks and Their Impending Use of GenAI to Promote Mal-Info: Gab as a Case Study." *CEUR Workshop Proceedings* 3593: 22–26.

Başar, E., I. Hendrickx, E. Krahmer, G.-J. Bruijn, and T. De Bosse. 2022. Hints of Independence in a Pre-scripted World. On Controlled Usage of Open-domain Language Models for Chatbots in Highly Sensitive Domains. In Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - V Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 1, 401-407. DOI: 10.5220/0010914300003116

Begou, N., J. Vinoy, A. Duda, and M. Korczynski. 2023. "Exploring the Dark Side of AI." Advanced Phishing Attack Design and Deployment Using ChatGPT. 2023 IEEE Conference on Communications and Network Security (CNS), Orlando, FL, USA, 2-5 Oct. 2023.

Behnia, R., M. R. Ebrahimi, J. Pacheco, and B. Padmanabhan. 2022. "EW-Tune. A Framework for Privately Fine-Tuning Large Language Models with Differential Privacy." IEEE International Conference on Data Mining Workshops, ICDMW. 28 November 2022 - 01 December 2022, Orlando, FL, USA

Bezzi, M. 2024. *Large Language Models and Security*. IEEE Security and Privacy, S. 2–10.

Botacin, M. 2023. GPThreats-3. Is Automatic Malware Generation a Threat? Proceeding - 44th IEEE Symposium on Security and Privacy Workshops, SPW 2023. May 22-25, 2023, San Francisco, CA, USA.

Bran, E., C. Rughinis, G. Nadoleanu, and M. G. Flaherty. 2023. "The Emerging Social Status of Generative AI." Vocabularies of AI Competence in Public Discourse. Proceedings - 2023 24th International Conference on Control Systems and Computer Science, CSCS 2023. 24-26 May 2023, Bucharest, Romania

Bruno, A., P. L. Mazzeo, A. Chetouani, M. Tliba, and M. A. Kerkouri. 2023. "Insights into Classifying and Mitigating LLMs' Hallucinations." *CEUR Workshop Proceedings* 3563, S: 50–63.

Cambiaso, E., and L. Caviglione. 2023. "Scamming the Scammers: Using ChatGPT to Reply Mails for Wasting Time and Resources." CEUR Workshop Proceedings 3488. Proceedings of the Italian Conference on Cybersecurity (ITASEC 2023), Bari, Italy, May 2-5, 2023.

Cantens, T. 2024. How Will the State Think with ChatGPT? The Challenges of Generative Artificial Intelligence for Public Administrations. *AI and Society*. 40, 133–144. https://doi.org/10.1007/s00146-023-01840-9

Carlini, N., F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, et al. 2021. "Extracting Training Data from Large Language Models." In Proceedings of the 30th USENIX Security Symposium, S. 2633–2650.

Chatzoglou, E., G. Karopoulos, G. Kambourakis, and Z. Tsiatsikas. 2023. "Bypassing Antivirus Detection. Old-School Malware, New Tricks." In Proceedings of the 18th International Conference on Availability, Reliability and Security. August 29 to September 1, 2023, in Benevento, Italy.

Chen, B., A. Paliwal, and Q. Yan. 2023a. "Jailbreaker in Jail." Moving Target Defense for Large Language Models. MTD 2023 - Proceedings of the 10th ACM Workshop on Moving Target Defense. MTD '23, November 26, 2023, Copenhagen, Denmark. DOI: https://doi.org/10.1145/3605760.3623764

Chen, B., G. Wang, H. Guo, Y. Wang, and Q. Yan. 2023b. "Understanding Multi-Turn Toxic Behaviors in Open-Domain Chatbots." ACM International Conference Proceeding Series. RAID '23, October 16–18, 2023, Hong Kong, China; https://doi.org/10.1145/3607199.3607237

Chen, Y., Z. Ding, L. Alowain, X. Chen, and D. Wagner. 2023c. "DiverseVul. A New Vulnerable Source Code Dataset for Deep Learning Based Vulnerability Detection." ACM International Conference Proceeding Series. RAID '23, October 16–18, 2023, Hong Kong, China; https://doi.org/10.1145/3607199.3607242

Chen, Y., T. X. Liu, Y. Shan, and S. Zhong. 2023d. "The Emergence of Economic Rationality of GPT." *Proceedings of the National Academy of Sciences of the United States of America* 120 (51): e2316205120. (https://doi.org/10.1073/pnas.2316205120.

Chen, M., G. He, and J. Wu. 2024. "ZDDR: A Zero-Shot Defender for Adversarial Samples Detection and Restoration," in *IEEE Access*, vol. 12, pp. 39081-39094, 2024, doi: 10.1109/ACCESS.2024.3356568

Cook, S., A. Hagiu, and J. Wright. 2024. "Turn Generative AI from an Existential Threat into a Competitive Advantage." *Harvard Business Review*. 102 (1), 118–125.

Crothers, E. N., N. Japkowicz, and H. L. Viktor. 2023. "Machine-Generated Text. A Comprehensive Survey of Threat Models and Detection Methods." *IEEE Access* 11: 70977–71002.

Culotta, A., and N. Mattei. 2024. "Use Open Source for Safer Generative AI Experiments." *In: MIT Sloan Management Review* 65 (2): 11–12.

Dai, L., J. Mao, L. Xu, X. Fan, and X. Zhou. 2024. "SecNLP. An NLP Classification Model Watermarking Framework Based on Multi-Task Learning." *Computer Speech & Language* 86: 101606. https://doi.org/10.1016/j.csl.2023.101606.

Das, N., A. Kotal, D. Roseberry, and A. Joshi. 2023. "Change Management Using Generative Modeling on Digital Twins." Proceedings - 2023 IEEE International Conference on Intelligence and Security Informatics, ISI 2023.

Deng, B., W. Wang, F. Feng, Y. Deng, Q. Wang, and X. He. 2023. "Attack Prompt Generation for Red Teaming and Defending Large Language Models." In Findings of the Association for Computational Linguistics: EMNLP 2023, S. 2176–2189. https://doi.org/10.18653/v1/2023.findings-emnlp.143.

Dermawan, A. 2024. "Text and Data Mining Exceptions in the Development of Generative AI Models. What the EU Member States Could Learn from the Japanese "Nonenjoyment" Purposes?" *The Journal of World Intellectual Property* 27 (1): 44–68. https://doi.org/10.1111/jwip.12285.

Diemert, S., and J. H. Weber. 2023. Can Large Language Models Assist in Hazard Analysis? Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14182 LNCS, 410–422.

Dong, H., J. Dong, S. Wan, S. Yuan, and Z. Guan. 2023. "Transferable Adversarial Distribution Learning. Query-Efficient Adversarial Attack against Large Language Models." *Computers & Security* 135: 103482. https://doi.org/10.1016/j.cose.2023.103482.

Dung, L. 2023. "Current Cases of AI Misalignment and Their Implications for Future Risks." *Synthese* 202 (5): 138. https://doi.org/10.1007/s11229-023-04367-0.

Ebert, C., and P. Louridas. 2023. "Generative AI for Software Practitioners." *IEEE Software* 40 (4): 30–38. https://doi.org/10.1109/MS.2023.3265877.

Edenberg, E., and A. Wood. 2023. "Disambiguating Algorithmic Bias. From Neutrality to Justice." AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI Ethics, and Society. AIES 2023, Montréal, QC, Canada, August 8-10, 2023.

Editorial Nature. 2023. "Stop Talking about Tomorrow's AI Doomsday When AI Poses Risks Today." *Nature* 618: 885–886.

Elgesem, D. 2023. "The AI Act and the Risks Posed by Generative AI Models." CEUR Workshop Proceedings 3431. Proceedings of the 5th Symposium of the Norwegian AI Society (NAIS 2023), Bergen, Norway, June 14-15, 2023.

Ezzeddine, F., O. Ayoub, S. Giordano, G. Nogara, I. Sbeity, E. Ferrara, and L. Luceri. 2023. "Exposing Influence Campaigns in the Age of LLMs: A Behavioral-Based AI Approach to Detecting State-Sponsored Trolls." *EPJ Data Science* 12 (1): 46. https://doi.org/10.1140/epjds/s13688-023-00423-4.

Favela, L. H., and M. J. Amon. 2023. "The Ethics of Human Digital Twins. Counterfeit People, Personhood, and the Right to Privacy." 2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence. DTPI 2023. October 16th, 2023 to November 9th, 2023, Macau, China.

Faye, G., W. Ouerdane, G. Gadek, S. Gahbiche, and S. Gatepaille. 2023. "A Novel Hybrid Approach for Text Encoding. Cognitive Attention To Syntax Model to Detect Online Misinformation." *Data & Knowledge Engineering* 148: 102230. https://doi.org/10.1016/j.datak.2023.102230.

Fayyad, U. M. 2023. "From Stochastic Parrots to Intelligent Assistants - The Secrets of Data and Human Interventions." *IEEE Intelligent Systems* 38 (3): 63–67. https://doi.org/10.1109/MIS.2023.3268723.

Feldstein, S. 2023. "The Consequences of Generative AI for Democracy, Governance and War." *Survival* 65 (5): 117–142. https://doi.org/10.1080/00396338.2023.2261260.

Fenwick, M., and P. Jurcys. 2023. "Originality and the Future of Copyright in an Age of Generative AI." *Computer Law & Security Review* 51: 105892. https://doi.org/10.1016/j.clsr.2023.105892.

Ferrag, M. A., M. Debbah, and M. Al-Hawawreh. 2023. "Generative AI for Cyber Threat-Hunting in 6G-Enabled IoT Networks." Proceedings - 23rd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing Workshops, CCGridW 2023. Bangalore, India, May 1-4, 2023 and Ferrag, van Dijk et al 2023: Coral Bay, Pafos, Cyprus June 19-21, 2023

Ferrag, M. A., D. Hamouda, M. Debbah, L. Maglaras, and A. Lakas. 2023. "Generative Adversarial Networks-Driven Cyber Threat Intelligence Detection Framework for Securing Internet of Things." Proceedings - 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things, DCOSS-IoT 2023.

Ferrari, F., J. van Dijck, and A. van den Bosch. 2023. "Observe, Inspect, Modify. Three Conditions for Generative AI Governance." *New Media and Society*. 27(5), 2788-2806. https://doi.org/10.1177/14614448231214811.

Fisher, J. A. 2023. Centering the Human: Digital Humanism and the Practice of Using Generative AI in the Authoring of Interactive Digital Narratives. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). LNCS, 73–88.

Fields, J., K. Chovanec, and P. Madiraju. 2024. "A Survey of Text Classification with Transformers. How Wide? How Large? How Long? How Accurate? How Expensive? How Safe?" *IEEE Access*: 1: 6518–6531.

Franco, M., O. Gaggi, and C. E. Palazzi. 2023. "Analyzing the Use of Large Language Models for Content Moderation with ChatGPT Examples." Proceedings of the 2023 Workshop on Open Challenges in Online Social Networks, OASIS 2023, Held in Conjunction with the 34th ACM Conference on Hypertext and Social Media, HT 2023. Rome Italy September 4 - 8, 2023

Gallus, P., M. Stepanek, T. Racil, and P. Frantis. 2023. "Generative Neural Networks as a Tool for Web Applications Penetration Testing." 2023 Communication and Information Technologies, KIT 2023 - 12th International Scientific Conference, Proceedings. 11-13 Oct. 2023, Tatranské Zruby, Slovakia

Goldstein, S., and C. D. Kirk-Giannini. 2023. "Language Agents Reduce the Risk of Existential Catastrophe." *AI and Society*. 40, 959–969. https://doi.org/10.1007/s00146-023-01748-4

Gonzalez Torres, A. P., K. Kajava, and N. Sawhney. 2023. Emerging AI Discourses and Policies in the EU: Implications for Evolving AI Governance. In: Pillay, A., Jembere, E., J. Gerber, A. (eds) Artificial Intelligence Research. SACAIR 2023. Communications in Computer and Information Science, vol 1976. Springer, Cham. https://doi.org/10.1007/978-3-031-49002-6_1. Muldersdrift, South Africa, December 4–8

Gradon, K. T. 2023. "Electric Sheep on the Pastures of Disinformation and Targeted Phishing Campaigns. The Security Implications of ChatGPT." *IEEE Security & Privacy* 21 (3): 58–61. https://doi.org/10.1109/MSEC.2023.3255039.

Gromova, E. A., D. B. Ferreira, and I. R. Begishev. 2023. "ChatGPT and Other Intelligent Chatbots. Legal, Ethical and Dispute Resolution Concerns." *Revista Brasileira de Alternative Dispute Resolution* 5 (10), S): 153–175.

Gupta, Sachin. 2022. "Hate Speech Detection Using OpenAI and GPT-3." *International Journal of Emerging Technology and Advanced Engineering* 12 (5): 132–138. https://doi.org/10.46338/ijetae0522_15.

Gupta, G., K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad. 2023. "A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods." *Electronics* 13 (1): 95. https://doi.org/10.3390/electronics13010095.

Gupta, M., C. Akiri, K. Aryal, E. Parker, and L. Praharaj. 2023. "From ChatGPT to ThreatGPT. Impact of Generative AI in Cybersecurity and Privacy." *IEEE Access* 11: 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381.

Gupta, S., P. Sharma, S. Chaudhary, V. Kumar, S. P. Singh, M. Lourens, and N. Beri. 2024. "Study on the Beneficial Impacts and Ethical Dimensions of Generative AI in Software Product Management." *International Journal of Intelligent Systems and Applications in Engineering* 12 (8s): 251–264.

Ha, H., D. Tran, and D. Kim. 2023. "Black-Box Adversarial Attacks Against Language Model Detector." ACM International Conference Proceeding Series. 12th International Symposium on Information and Communication Technology (SOICT 2023), Ho Chi Minh City, Vietnam on 7-8 December 2023.

Hacker, P. 2023. "The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future." *Computer Law & Security Review* 51: 105871. https://doi.org/10.1016/j.clsr.2023.105871.

Hamed, A. A., M. Zachara-Szymanska, and X. Wu. 2024. "Safeguarding Authenticity for Mitigating the Harms of Generative AI. Issues, Research Agenda, and Policies for Detection, Fact-Checking, and Ethical AI." *iScience* 27 (2): 108782.

Hammond, K., and D. Leake. 2023. "Large Language Models Need Symbolic AI." *CEUR Workshop Proceedings* 3432, S: 204–209.

Happe, A., and J. Cito. 2023. "Getting Pwn'd by AI. Penetration Testing with Large Language Models." ESEC/FSE 2023 - Proceedings of the 31st ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 5th - 7th December 2023, an Francisco, California, United States

He, J., and M. Vechev. 2023. "Large Language Models for Code. Security Hardening and Adversarial Testing." CCS 2023 - Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. 12th International Symposium on Information and Communication Technology, SOICT 2023, Ho Chi Minh, Vietnam, December 7-8, 2023

Head, C. B., P. Jasper, M. McConnachie, L. Raftree, and G. Higdon. 2023. "Large Language Model Applications for Evaluation. Opportunities and Ethical Implications." In *New Directions for Evaluation 2023* (178-179), S. 33–46.

Helberger, N., and N. Diakopoulos. 2023.ChatGPT and the AI Act. Internet Policy Review, 12(1). https://doi.org/10.14763/2023.1.1682, https://policyreview.info/pdf/policyreview-2023-1-1682.pdf

Herbosch, M. 2024. "Fraud by Generative AI Chatbots. On the Thin Line between Deception and Negligence." *Computer Law & Security Review* 52: 105941. https://doi.org/10.1016/j.clsr.2024.105941.

Holm, J. R., D. S. Hain, R. Jurowetzki, and E. Lorenz. 2023. "Innovation Dynamics in the Age of Artificial Intelligence. Introduction to the Special Issue." *Industry and Innovation* 30 (9): 1141–1155. https://doi.org/10.1080/13662716.2023.2272724.

Horzyk, A. 2024. Data Protection and Privacy: Risks and Solutions in the Contentious Era of AI-Driven Ad Tech. In: Luo, B., Cheng, L., Wu, ZG., Li, H., Li, C. (eds) Neural Information Processing. ICONIP 2023. Communications in Computer and Information Science, vol 1968. Springer, Singapore. https://doi.org/10.1007/978-981-99-8181-6_27. ICONIP 2023, Changsha, China, November 20–23, 2023

Hu, C., and J. Chen. 2023. "A Dimensional Perspective Analysis on the Cybersecurity Risks and Opportunities of ChatGPT-Like Information Systems." Proceedings - 2023 International Conference on Networking and Network Applications, NaNA 2023.

Ioannidis, J., J. Harper, M. S. Quah, and D. Hunter. 2023. "Gracenote.ai: Legal Generative AI for Regulatory Compliance." *In: CEUR Workshop Proceedings* 3423: 20–31.

Iqbal, U. 2023. "From Knowledge Management to Intelligence Engineering - A Practical Approach to Building AI inside the Law-Firm Using Open-Source Large Language Models." *CEUR Workshop Proceedings* 3423, S: 49–57.

Jain, R., N. Gervasoni, M. Ndhlovu, and S. Rawat. 2023. "A Code Centric Evaluation of C/C++ Vulnerability Datasets for Deep Learning Based Vulnerability Detection Techniques." ACM International Conference Proceeding Series. ISEC '23: Proceedings of the 16th Innovations in Software Engineering Conference Article No.: 6, Pages 1 - 10 https://doi.org/10.1145/3578527.3578530. Allahabad India February 23 - 25, 2023

Jha, S., S. K. Jha, P. Lincoln, N. D. Bastian, A. Velasquez, and S. Neema. 2023. "Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting." Proceedings - 2023 IEEE International Conference on Assured Autonomy, ICAA 2023. June 6-8, 2023, Baltimore, MD, USA

Ji, Z., T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung. 2023. "Towards Mitigating Hallucination in Large Language Models via Self-Reflection." Findings of the Association for Computational Linguistics: EMNLP 2023, 1827–1843. Singapore, dec 6-10, 2023

Jungherr, A. 2023. "Artificial Intelligence and Democracy. A Conceptual Framework." *Social Media and Society* 9 (3): 1–14.

Kaffee, L.-A., A. Arora, Z. Talat, and I. Augenstein. 2023. "Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing." Findings of the Association for Computational Linguistics: EMNLP 2023, 13977–13998. https://doi.org/10.18653/v1/2023.findings-emnlp.932. Singapore, dec 6-10, 2023

Kandpal, N., E. Wallace, and C. Raffel. 2022. "Deduplicating Training Data Mitigates Privacy Risks in Language Models." *Proceedings of Machine Learning Research* 162: 10697–10707.

Karinshak, E. 2024. "Simulations in Human-AI Crisis Management Systems." *Journal of Contingencies and Crisis Management* 32 (1): 1 - 29, Article No.: 116. https://doi.org/10.1111/1468-5973.12525.

Karinshak, E., S. X. Liu, J. S. Park, and J. T. Hancock. 2023. "Working With AI to Persuade. Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages." *Proceedings of the ACM on Human-Computer Interaction* 7: e12525 (CSCW1).

Katrenčík, I., B. Mucha, and M. Zatrochová. 2023. "Ethical Implications of Artificial Intelligence Data Usage. A Case Study of Slovakia and Global Perspectives." IDIMT 2023: New Challenges for ICT and Management - 31st Interdisciplinary Information Management Talks. Sept. 6–8, 2023 Hradec Králové, Czech Republic

Khan, A., and K. M. Malik. 2023. "Securing Voice Biometrics. One-Shot Learning Approach for Audio Deepfake Detection." WIFS 2023 - IEEE Workshop on Information Forensics and Security. December 4 to 7, 2023, Nuremberg, Germany.

Khan, M. S., and H. Umer. 2024. "ChatGPT in Finance. Applications, Challenges, and Solutions." *Heliyon* 10 (2): e24890. https://doi.org/10.1016/j.heliyon.2024.e24890.

Khlaisamniang, P., P. Khomduean, K. Saetan, and S. Wonglapsuwan. 2023. "Generative AI for Self-Healing Systems." 18th International Conference on Artificial Intelligence and Natural Language Processing and International Conference on Artificial Intelligence and Internet of Things, iSAI-NLP 2023. 27-29 Nov. 2023, Bangkok, Thailand

Kholgh, D. K., and P. Kostakos. 2023. "PAC-GPT. A Novel Approach to Generating Synthetic Network Traffic With GPT-3." *IEEE Access* 11: 114936–114951. https://doi.org/10.1109/ACCESS.2023.3325727.

Kieseberg, P., E. Weippl, A. M. Tjoa, F. Cabitza, A. Campagner, and A. Holzinger. 2023. Controllable AI - An Alternative to Trustworthiness in Complex AI Systems? Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14065 LNCS, 1–12.

Kshetri, N. 2023. "Cybercrime and Privacy Threats of Large Language Models." *IT Professional* 25 (3): 9–13. https://doi.org/10.1109/MITP.2023.3275489.

Li, H., D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song. 2023. "Multi-Step Jailbreaking Privacy Attacks on ChatGPT." Findings of the Association for Computational Linguistics: EMNLP 2023, 4138–4153. Singapore, dec 6-10, 2023 https://doi.org/10.18653/v1/2023.findings-emnlp.272.

Li, J., X. Cai, and L. Cheng. 2023. "Legal Regulation of Generative AI. A Multidimensional Construction." *International Journal of Legal Discourse* 8 (2): 365–388. https://doi.org/10.1515/ijld-2023-2017.

Li, L., and B. Gong. 2023. "Prompting Large Language Models for Malicious Webpage Detection." 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning, PRML 2023.

Liesenfeld, A., A. Lopez, and M. Dingemanse. 2023. "Opening up ChatGPT. Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators. Proceedings of the 5th International Conference on Conversational User Interfaces." CUI 2023. Eindhoven, The Netherlands, from 19–21st July 2023.

Linegar, M., R. Kocielnik, and R. M. Alvarez. 2023. "Large Language Models and Political Science." *Frontiers in Political Science* 5: 1257092. https://doi.org/10.3389/fpos.2023.1257092.

Liu, X., Y. Tan, Z. Xiao, J. Zhuge, and R. Zhou. 2023. "Not The End of Story: An Evaluation of ChatGPT-Driven Vulnerability Description Mappings." Proceedings of the Annual Meeting of the Association for Computational Linguistics, S. 3724–3731.

Liyanage, V., and D. Buscaldi. 2023. Detecting Artificially Generated Academic Text: The Importance of Mimicking Human Utilization of Large Language Models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13913 LNCS, 558–565.

Loukis, E., S. Saxena, N. Rizun, M. I. Maratsi, M. Ali, and C. Alexopoulos. 2023. ChatGPT Application vis-a-vis Open Government Data (OGD): Capabilities, Public Values, Issues and a Research Agenda. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14130 LNCS, 95–110.

Lu, Q., Y. Luo, L. Zhu, M. Tang, X. Xu, and J. Whittle. 2023. "Developing Responsible Chatbots for Financial Services. A Pattern-Oriented Responsible Artificial Intelligence Engineering Approach. In." *IEEE Intelligent Systems* 38 (6): 42–51. https://doi.org/10.1109/MIS.2023.3320437.

Luitse, D., and W. Denkena. 2021. "The Great Transformer. Examining the Role of Large Language Models in the Political Economy of AI." *Big Data & Society* 8 (2). https://doi.org/10.1177/20539517211047734.

Luo, C. F., R. Bhambhoria, S. Dahan, and X. Zhu. 2023. "Legally Enforceable Hate Speech Detection for Public Forums." Findings of the Association for Computational Linguistics: EMNLP 2023, S. 10948–10963. https://doi.org/10.18653/v1/2023.findings-emnlp.730.

Masotina, M., E. Musi, and A. Spagnolli. 2023. "Transparency is Crucial for User-Centered AI, or is It? How This Notion Manifests in the UK Press Coverage of GPT." ACM International Conference Proceeding Series. In 15th Biannual Conference of the Italian SIGCHI Chapter CHItaly 2023), September 20-22, 2023, Torino, Italy. ACM, New York, NY, USA, 11 Pages. https://doi.org/10.1145/3605390.3605413

McIntosh, T., T. Liu, T. Susnjak, H. Alavizadeh, A. Ng, R. Nowrozy, and P. Watters. 2023. "Harnessing GPT-4 for Generation of Cybersecurity GRC Policies. A Focus on Ransomware Attack Mitigation." *Computers & Security* 134: 103424. https://doi.org/10.1016/j.cose.2023.103424.

Milic-Frayling, N. 2023. "On the Cusp. Computing Thrills and Perils and Professional Awakening." Proceedings of the VLDB Endowment 16 (12): 4152–4159.

Mirsky, Y., and W. Lee. 2021. "The Creation and Detection of Deepfakes." *In: ACM Computing Surveys* 54 (1):1 - 4, Article No.: 7.

Mubarak, R., T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson. 2023. "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats." *IEEE Access* 11: 144497–144529. https://doi.org/10.1109/ACCESS.2023.3344653.

Mumuni, F., and A. Mumuni. 2024. "Improving Deep Learning with Prior Knowledge and Cognitive Models. A Survey on Enhancing Explainability, Adversarial Robustness and Zero-Shot Learning." *Cognitive Systems Research* 84:101188.

Nair, M., R. Sadhukhan, and D. Mukhopadhyay. 2023. How Hardened is Your Hardware? Guiding ChatGPT to Generate Secure Hardware Resistant to CWEs. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13914 LNCS, 320–336.

Naitali, A., M. Ridouani, F. Salahdine, and N. Kaabouch. 2023. "Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions." *Computers* 12 (10): 216. https://doi.org/10.3390/computers12100216.

Nannini, L. 2022. "Voluminous yet Vacuous? Semantic Capital in an Age of Large Language Models." CEUR Workshop Proceedings 3547. Proceedings of the Workshop on Ethics and Trust in Human-AI Collaboration: Socio-Technical Approaches (ETHAICS 2023), Macao, August 21, 2023.

Nelson, J. W. 2023. "Large Language Models and the Treaty Interpretation Game." *Cambridge International Law Journal* 12 (2): 305–327. https://doi.org/10.4337/cilj.2023.02.08.

Novelli, C., F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi. 2023. "Taking AI Risks Seriously: A New Assessment Model for the AI Act." *AI and Society*. 39, 2493–2497. https://doi.org/10.1007/s00146-023-01723-z

O'Connell, M. E. 2023. "Banning Autonomous Weapons: A Legal and Ethical Mandate." *Ethics & International Affairs* 37 (3): 287–298. https://doi.org/10.1017/S0892679423000357.

O'Leary, D. E. 2023. "Enterprise Large Language Models. Knowledge Characteristics, Risks, and Organizational Activities." *Intelligent Systems in Accounting, Finance and Management* 30 (3): 113–119. https://doi.org/10.1002/isaf.1541.

Oppenlaender, J., J. Silvennoinen, V. Paananen, and A. Visuri. 2023. Perceptions and Realities of Text-to-Image Generation. In 26th International Academic Mindtrek Conference (Mindtrek '23), October 03--06, 2023, Tampere, Finland. ACM, New York, NY, USA 10 Pages. https://doi.org/10.1145/3616961.3616978

Ortega, E., M. Tran, and G. Bandeen. 2023. "AI Digital Tool Product Lifecycle Governance Framework through Ethics and Compliance by Design†," 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, 5-6 June 2023, pp. 353-356, doi: 10.1109/CAI54212.2023.00155.

Ozdayi, M. S., C. Peris, J. Fitzgerald, C. Dupuy, J. Majmudar, H. Khan, et al. 2023. "Controlling the Extraction of Memorized Data from Large Language Models via Prompt-Tuning." In Proceedings of the Annual Meeting of the Association for Computational Linguistics 2, S. 1512–1521. Toronto, Canada, July 9th to July 14th, 2023

Ozkan-Okay, M., E. Akin, O. Aslan, S. Kosunalp, T. Iliev, I. Stoyanov, and I. Beloev. 2024. "A Comprehensive Survey. Evaluating the Efficiency of Artificial Intelligence and Machine Learning Techniques on Cyber Security Solutions." *IEEE Access* 12: 12229–12256.

Ozkaya, I. 2023. "Application of Large Language Models to Software Engineering Tasks. Opportunities, Risks, and Implications." *IEEE Software* 40 (3): 4–8. https://doi.org/10.1109/MS.2023.3248401.

Ozturk, O. S., E. Ekmekcioglu, O. Cetin, B. Arief, and J. Hernandez-Castro. 2023. "New Tricks to Old Codes. Can AI Chatbots Replace Static Code Analysis Tools?" ACM International Conference Proceeding.

Pa Pa, Y. M., S. Tanizaki, T. Kou, M. van Eeten, K. Yoshioka, and T. Matsumoto. 2023. "An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware." ACM International Conference Proceeding Series. 16th Cyber Security Experimentation and Test Workshop, CSET 2023 - Hybrid, Marina Del Rey, United States, 7 Aug 2023

Pan, J. 2023. "A Foundation Model Approach to Detect Machine Generated Text." IEEE Region 10 Annual International Conference, Proceedings/TENCON. Singapore, dec 6-10, 2023

Pan, Y., L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Y. Wang. 2023. "On the Risk of Misinformation Pollution with Large Language Models." Findings of the Association for Computational Linguistics: EMNLP 2023, 1389–1403. https://doi.org/10.18653/v1/2023.findings-emnlp.97.

Paraschiv, A., and M. Dascalu. 2023. "UPB at ACTI: Detecting Conspiracies Using Fine Tuned Sentence Transformers." CEUR Workshop Proceedings 3473. Proceedings of the Eighth Evaluation Campaign of Natural Language processing and Speech Tools for Italian (EVALITA 2023), Parma, italy, September 7-8, 2023.

Park, A., J. Kietzmann, J. Killoran, Y. Cui, P. V. Esch, and A. Dabirian. 2023. "Nothing Is Harder to Resist Than the Temptation of AI." *IT Professional* 25 (6): 13–20. https://doi.org/10.1109/MITP.2023.3340529.

Pasupuleti, R., R. Vadapalli, and C. Mader. 2023. "Cyber Security Issues and Challenges Related to Generative AI and ChatGPT." Proceedings - 2023 10th International Conference on Social Networks Analysis, Management and Security, SNAMS 2023. Abu Dhabi, UAE. November 21-24, 2023.

Patel, C. R., S. K. Pandya, and B. M. Sojitra. 2023. "Perspectives of ChatGPT in Pharmacology Education, and Research in Health Care. A Narrative Review." *Journal of Pharmacology and Pharmacotherapeutics* 14 (3): 171–177. https://doi.org/10.1177/0976500X231210427.

Patil, M., H. Yadav, M. Gawali, J. Suryawanshi, J. Patil, A. Yeole, et al. 2024. "A Novel Approach to Fake News Detection Using Generative AI." *International Journal of Intelligent Systems and Applications in Engineering* 12 (4s): 343–354.

Pearce, H., B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt. 2023. "Examining Zero-Shot Vulnerability Repair with Large Language Models." Proceedings - IEEE Symposium on Security and Privacy. May 22-25, 2023, San Francisco, CA, USA

Piñeiro-Martín, A., C. García-Mateo, L. Docío-Fernández, and M. D. C. López-Pérez. 2023. "Ethical Challenges in the Development of Virtual Assistants Powered by Large Language Models." *Electronics* 12 (14): 3170. https://doi.org/10.3390/electronics12143170.

Pinter, Y., and M. Elhadad. 2023. "Emptying the Ocean with a Spoon: Should We Edit Models?" Findings of the Association for Computational Linguistics: EMNLP 2023, S. 15164–15172. Singapore, dec 6-10, 2023

Prabhumoye, S., M. Patwary, M. Shoeybi, and B. Catanzaro. 2023. "Adding Instructions during Pretraining: Effective Way of Controlling Toxicity in Language Models." EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 2628–2643.

Purba, M. D., A. Ghosh, B. J. Radford, and B. Chu. 2023. "Software Vulnerability Detection Using Large Language Models." Proceedings - 2023 IEEE 34th International Symposium on Software Reliability Engineering Workshop, ISSREW 2023. Florence, Italy, 9th to 12th October 2023

Qu, Y., B. Bai, and Z. Zhang. 2023. "The New Generation of Artificial Intelligence Technology ChatGPT Causes. Potential Legal Risks and Regulatory Countermeasures." 2023 8th International Conference on Computer and Communication Systems, ICCCS 2023. Guangzhou, China on April 21-24, 2023

Ragsdale, J., and R. V. Boppana. 2023. "On Designing Low-Risk Honeypots Using Generative Pre-Trained Transformer Models with Curated Inputs." *IEEE Access* 11: 117528–117545. https://doi.org/10.1109/ACCESS.2023.3326104.

Ramakrishna, A., R. Gupta, J. Lehmann, and M. Ziyadi. 2023. "INVITE: A Testbed of Automatically Generated Invalid Questions to Evaluate Large Language Models for Hallucinations." Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, dec 6-10, 2023, S. 5422–5429. https://doi.org/10.18653/v1/2023.findings-emnlp.360.

Ramezani, A., and Y. Xu. 2023. "Knowledge of Cultural Moral Norms in Large Language Models." Proceedings of the Annual Meeting of the Association for Computational Linguistics 1, ACL 2023, Toronto, Canada, July 9-14, 2023, S. 428–446.

Roe, J., and M. Perkins. 2023. "'What They're Not Telling You about ChatGPT': exploring the Discourse of AI in UK News Media Headlines." *Humanities and Social Sciences Communications* 10 (1): 753. https://doi.org/10.1057/s41599-023-02282-w.

Rosenberg, L. 2023. "Generative AI as a Dangerous New Form of Media." Proceedings of the International Multi-Conference on Society, Cybernetics and Informatics, IMSCI, Virtual Conference September 12-15, 2023.165–170. https://doi.org/10.54808/IMSCI2023.01.165.

Rosengrün, S. 2023. "Everything but the Truth. On the Relevance of Algorithms." *Analecta Hermeneutica* 15 (1), S): 16–26.

Roy, S., A. Harshavardhan, A. Mukherjee, and P. Saha. 2023. "Probing LLMs for Hate Speech Detection: Strengths and Vulnerabilities." In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, dec 6-10, 2023, S. 6116–6128. https://doi.org/10.18653/v1/2023.findings-emnlp.407.

Rubinic, Igor, Marija Kurtov, Ivan Rubinic, Robert Likic, Paul I. Dargan, and David M. Wood. 2024. "Artificial Intelligence in Clinical Pharmacology. A Case Study and Scoping Review of Large Language Models and Bioweapon Potential." *British Journal of Clinical Pharmacology* 90 (3): 620–628. https://doi.org/10.1111/bcp.15899.

Sarkadi, S. 2023. "Deceptive AI and Society." *IEEE Technology and Society Magazine* 42 (4): 77–86. Shttps://doi.org/10.1109/MTS.2023.3340232.

Semnani, S. J., V. Z. Yao, H. C. Zhang, and M. S. Lam. 2023. "WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia." Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, dec 6-10, 2023. S. 2387–2413. https://doi.org/10.18653/v1/2023.findings-emnlp.157.

Singla, T., D. Anandayuvaraj, K. G. Kalu, T. R. Schorlemmer, and J. C. Davis. 2023. "An Empirical Study on Using Large Language Models to Analyze Software Supply Chain Security Failures." SCORED 2023 - Proceedings of the 2023 Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses. SCORED 2023, Copenhagen, Denmark, 30 November 2023.

Sonntagbauer, M., M. Haar, and S. Kluge. 2023. "Artificial Intelligence: How Will ChatGPT and Other AI Applications Change Our Everyday Medical Practice? [Künstliche Intelligenz: Wie Werden ChatGPT Und Andere KI-Anwendungen Unseren Ärztlichen Alltag Verändern?]" *Medizinische Klinik, Intensivmedizin Und Notfallmedizin* 118 (5): 366–371. https://doi.org/10.1007/s00063-023-01019-6.

Subramani, N., A. S. Luccioni, J. Dodge, and M. Mitchell. 2023. "Detecting Personal Information in Training Corpora: An Analysis." Proceedings of the Annual Meeting of the Association for Computational Linguistics, Singapore, dec 6-10, 2023. S. 208–220.

Taveekitworachai, P., F. Abdullah, M. C. Gursesli, M. F. Dewantoro, S. Chen, A. Lanata, A. Guazzini, and R. Thawonmas. 2023. Breaking Bad: Unraveling Influences and Risks of User Inputs to ChatGPT for Game Story Generation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14384 LNCS, 285–296.

Tihanyi, N., T. Bisztray, R. Jain, M. A. Ferrag, L. C. Cordeiro, and V. Mavroeidis. 2023. "The FormAI Dataset." Generative AI in Software Security through the Lens of Formal Verification. ACM International Conference Proceeding Series. PROMISE 2023: Proceedings of the 19th International Conference on Predictive Models and Data Analytics in Software Engineering Pages 33–43, https://doi.org/10.1145/3617555.36178, San Francisco, CA, USA, 8 December 2023

Tkachenko, N. 2023. "Opportunities for Synthetic Data in Nature and Climate Finance." *Frontiers in Artificial Intelligence* 6: 1168749. https://doi.org/10.3389/frai.2023.1168749.

Toetzke, M., B. Probst, and S. Feuerriegel. 2023. "Leveraging Large Language Models to Monitor Climate Technology Innovation." *Environmental Research Letters* 18 (9): 091004. https://doi.org/10.1088/1748-9326/acf233.

Transatlantic Reflection Group on Democracy and the Rule of Law in the Age of "Artificial Intelligence". 2023. "A Manifesto on Enforcing Law in the Age of 'Artificial Intelligence' (2023)." *European Law Journal* 29 (1–2): 249–255.

Vahabava, M. 2023. "The Risks Associated with Generative AI Apps in the European Artificial Intelligence Act (AIA)." *CEUR Workshop Proceedings* 3456, S: 7–18.

Vashishtha, A., S. S. K. Prasad, P. Bajaj, V. Chaudhary, K. Cook, S. Dandapat, et al. 2023. Performance and Risk Trade-offs for Multi-word Text Prediction at Scale. EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023, May 2-6, 2023, Dubrovnik, Croatia. S. 2181–2197.

Venturini, T. 2023. "Bruno Latour and Artificial Intelligence." *Tecnoscienza* 14 (2): 101–114.

Wan, Y., J. Zhao, A. Chadha, N. Peng, and K.-W. Chang. 2023. "Are Personalized Stochastic Parrots More Dangerous? Evaluating Persona Biases in Dialogue Systems." In Findings of the Association for Computational Linguistics: EMNLP 2023, S. 9677–9705. https://doi.org/10.18653/v1/2023.findings-emnlp.648.

Wang, C., H. Sharifzadeh, S. Varastehpour, and I. Ardekani. 2023. "Analysis and Comparison of Deepfakes Detection Methods for Cross-Library Generalisation." 2023 20th Annual International Conference on Privacy, Security and Trust, PST 2023.

Wörsdörfer, M. 2024. "Mitigating the Adverse Effects of AI with the European Union's Artificial Intelligence Act. Hype or Hope? In." *Global Business and Organizational Excellence* 43 (3): 106–126. https://doi.org/10.1002/joe.22238.

You, W., Z. Hammoudeh, and D. Lowd. 2023. "Large Language Models Are Better Adversaries: Exploring Generative Clean-Label Backdoor Attacks Against Text Classifiers." In Findings of the Association for Computational Linguistics: EMNLP 2023, S. 12499–12527. https://doi.org/10.18653/v1/2023.findings-emnlp.833.

Yu, Z. 2023. "A Multi-Dimensional Generic Evaluation Framework for the Security of Large Language Models." 2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2023.

Yu, Z., S. Zhai, and N. Zhang. 2023. "AntiFake. Using Adversarial Audio to Prevent Unauthorized Speech Synthesis." CCS 2023 - Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security.

Zhang, Q., C. Fang, B. Yu, W. Sun, T. Zhang, and Z. Chen. 2023. "Pre-Trained Model-Based Automated Software Vulnerability Repair. How Far Are We?." IEEE Transactions on Dependable and Secure Computing, S. 1–18.

Zou, W., and Z. Liu. 2023. "Unraveling Public Conspiracy Theories Toward ChatGPT in China. A Critical Discourse Analysis of Weibo Posts." *Journal of Broadcasting and Electronic Me*. 68(1), 1–20. https://doi.org/10.1080/08838151.2023.2275603