Journal of Computational Social Science
https://doi.org/10.1007/s42001-024-00250-1

SURVEY ARTICLE

GenAI against humanity: nefarious applications of generative artificial intelligence and large language models

Emilio Ferrara[1]

Received: 10 November 2023 / Accepted: 21 January 2024

Abstract

Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) are marvels of technology; celebrated for their prowess in natural language processing and multimodal content generation, they promise a transformative future. But as with all powerful tools, they come with their shadows. Picture living in a world where deepfakes are indistinguishable from reality, where synthetic identities orchestrate malicious campaigns, and where targeted misinformation or scams are crafted with unparalleled precision. Welcome to the darker side of GenAI applications. This article is not just a journey through the meanders of potential misuse of GenAI and LLMs, but also a call to recognize the urgency of the challenges ahead. As we navigate the seas of misinformation campaigns, malicious content generation, and the eerie creation of sophisticated malware, we'll uncover the societal implications that ripple through the GenAI revolution we are witnessing. From AI-powered botnets on social media platforms to the unnerving potential of AI to generate fabricated identities, or alibis made of synthetic realities, the stakes have never been higher. The lines between the virtual and the real worlds are blurring, and the consequences of potential GenAI's nefarious applications impact us all. This article serves both as a synthesis of rigorous research presented on the risks of GenAI and misuse of LLMs and as a thought-provoking vision of the different types of harmful GenAI applications we might encounter in the near future, and some ways we can prepare for them.

Keywords AI · Generative AI · Large Language Models · Risks · Social media

✉ Emilio Ferrara
emiliofe@usc.edu
https://scholar.google.com/citations?user=0r7Syh0AAAAJ

[1] Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA 90007, USA

Published online: 22 February 2024

Introduction

In March 2019, a UK-based energy firm's CEO was duped out of $243,000. The culprit? Not a seasoned con artist, but an AI-generated synthetic voice so convincingly mimicking the company's German parent firm's CEO that it led to a costly misstep (see Table 1A). This incident, while startling, is just the tip of the iceberg when it comes to the nefarious potential applications of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs).

GenAI and LLMs are transforming the landscape of natural language processing, content generation, and understanding. Their potential seems endless, promising innovations that could redefine the way humans and machines interact with each other or partner to work together [10, 26]. However, lurking in the shadows of these advancements are challenges that threaten the very fabric of our cybersecurity, ethics, and societal structures [15].

This paper ventures into the darker alleys of GenAI, with a focus on LLMs. From their potential role in scaling up misinformation campaigns, to the creation of targeted scams or custom-tailor-made alibies, the risks are profound [5]. From the subtle perpetuation of biases to the blatant reinforcement of stereotypes [2, 6, 21], GenAI can become mirrors reflecting and amplifying the imperfections of our society [1, 8].

Imagine a world where AI-powered botnets dominate social media [7, 29], where harmful or radicalizing content is churned out by algorithms [23], and where the lines between reality and AI-generated content blur [3]. A world where the same technology that can be used to restore lost pieces of art or ancient documents [4], can also be used to fabricate evidence, craft alibis, and conceive the "perfect crime" (see Table 1B). Many of these scenarios that until recently we would have ascribed to futuristic science fiction are already enabled by GenAI and LLMs.

As we navigate the complexities of these issues, this paper highlights the urgency of robust mitigation strategies, ethical guidelines, and continuous monitoring for GenAI and LLM systems [9, 14]. This exploration not only aims to summarize rigorous research on GenAI abuse, but also to ignite a discourse on the dual nature of these technologies.

Definition and mechanisms of generative AI and LLMs

Generative AI refers to artificial intelligence systems that can generate new content, including text, images, and audio, based on existing data [3]. Unlike traditional AI, which focuses on recognizing patterns or making predictions, GenAI actively creates novel outputs. This involves complex algorithms and models that learn from large datasets, recognize underlying structures, and emulate them in unique ways [4]. Large Language Models (LLMs), a subset of GenAI systems, specifically deal with textual data [10]. They are trained on extensive corpora of text, learning language patterns, syntax, and context. LLMs like GPT (Generative Pretrained Transformer) are capable of producing coherent, contextually relevant text, resembling human writing. Their mechanisms involve understanding input queries, accessing their extensive training data, and generating appropriate textual responses, which can range from answering questions to creating content.

Technological advancements and democratization

The democratization of Generative AI represents a pivotal change in AI technology. The early 2020s period witnessed significant advancements in the technical capabilities of GenAI, marked by improvements in machine learning algorithms, particularly in neural networks. These advancements led to the creation of more sophisticated and efficient models that are capable of understanding and generating complex data patterns.

At the same time, there was a marked decrease in the cost of developing and deploying GenAI systems. This was due to both the falling prices of computing power and the increased availability of open-source tools and platforms, making GenAI accessible to a wider range of users and developers. Furthermore, the proliferation of user-friendly interfaces and cloud-based services has made GenAI technologies more accessible to non-specialists. This broader access has catalyzed a wave of innovation and creativity across various sectors, allowing individuals and smaller organizations to take advantage of GenAI for various applications, thus democratizing the field of artificial intelligence.

Together, these factors have differentiated the current landscape of GenAI from previous generations of AI technologies, both in terms of technological sophistication and societal impact.

Regulatory landscape

The European Union (EU) and China have actively engaged in discussions to regulate Generative AI (GenAI). In the EU, the focus has been on establishing frameworks that ensure the ethical use of AI, focusing on data privacy, transparency, and accountability. The proposed regulations aim to categorize AI systems according to their risk levels and apply the corresponding oversight measures. China, on the other hand, has focused on harnessing the potential of GenAI while safeguarding national security and social stability. The Chinese approach includes stringent data control measures and guidelines to prevent the misuse of AI technologies, especially in areas like surveillance and censorship.

These divergent approaches reflect the complexities and varying priorities in GenAI governance, illustrating the challenges in creating a universally accepted regulatory framework. The policies in these regions are likely to influence global standards and practices in the GenAI domain [14].

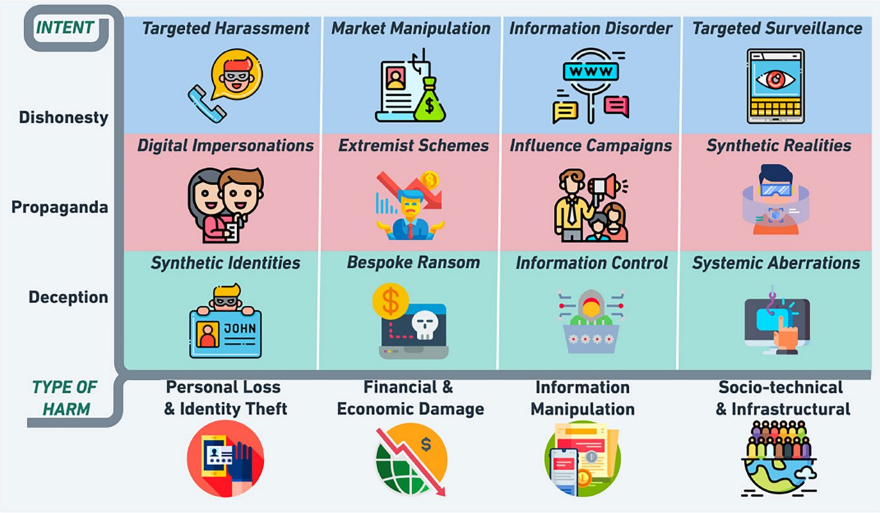
Understanding GenAI abuse: a taxonomy

Figure 1 offers an overview of the potential dangers associated with the misuse of generative AI models by charting the intersection between the type of harm that can be inflicted and the underlying intentions of malicious actors.

The types of harm encompass threats to an individual's personal identity, such as identity theft, privacy breaches, or personal defamation, which we term as "Harm to the Person." Then, we have the potential for financial loss, fraud, market manipulation, and other economic harms, which fall under "Financial and Economic Damage." The distortion of the information ecosystem, including the spread of misinformation, fake news, and other forms of deceptive content [28], is categorized as "Information Manipulation." Lastly, broader harms that can impact communities, societal structures, and critical infrastructures, including threats to democratic processes, social cohesion, and technological systems, are captured under "Societal, Socio-technical, and Infrastructural Damage."

On the other side of the matrix, we have the goals (i.e., intent) of malicious actors. "Deception" involves misleading individuals or entities for various purposes, such as scams, impersonation, or other fraudulent activities [16]. "Propaganda" is the intent to promote a particular political, ideological, or commercial agenda, often by distorting facts or manipulating emotions. And "Dishonesty" covers a range of activities where the truth is concealed or misrepresented for personal gain, competitive advantage, or other ulterior motives. Naturally, this dimension does not fully encompass the goals or motivations behind all possible types of misuse of GenAI, but it serves as a guide to frame nefarious applications with respect to their intent to harm.

In this 3 × 4 matrix, each cell represents a unique combination of harm and malicious intent, illustrating the multifaceted forms of abuse possible with generative AI. For instance, AI-generated impersonation for identity theft might be found at the intersection of "Harm to the Person" and "Deception." Similarly, AI-driven fake news campaigns to influence public opinion could be represented at the crossroads of "Information Manipulation" and "Propaganda."



INTENT TYPE OF HARM	Personal Loss & Identity Theft	Financial & Economic Damage	Information Manipulation	Socio-technical & Infrastructural
Dishonesty	Targeted Harassment	Market Manipulation	Information Disorder	Targeted Surveillance
Propaganda	Digital Impersonations	Extremist Schemes	Influence Campaigns	Synthetic Realities
Deception	Synthetic Identities	Bespoke Ransom	Information Control	Systemic Aberrations

Fig. 1 Charting the landscape of nefarious applications of generative artificial intelligence and large language models

Table 1 News articles about nefarious GenAI and LLM applications

Ref.	News title	Media outlet	URL
A	Fraudsters used AI to mimic CEO's voice in unusual cybercrime case	Wall Street Journal	https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402
B	People are creating records of fake historical events using AI	Vice	https://www.vice.com/en/article/k7zqdw/people-are-creating-records-of-fake-historical-events-using-ai
C	'I don't want to upset people': Tom Cruise deepfake creator speaks out	The Guardian	https://www.theguardian.com/technology/2021/mar/05/how-started-tom-cruise-deepfake-tiktok-videos
D	Do these A.I.-created fake people look real to you?	New York Times	https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html
E	Generative AI: A blessing or a curse for cybersecurity?	InWeb3	https://www.inweb3.com/generative-ai-a-blessing-or-a-curse-for-cybersecurity/
F	Real-world AI threats in cybersecurity aren't science fiction	VentureBeat	https://venturebeat.com/ai/real-world-ai-threats-in-cybersecurity-arent-science-fiction/
G	AI amplifies scam calls and other deceptions	Marketplace	https://www.marketplace.org/2023/07/14/ai-amplifies-scam-calls-and-other-deceptions/
H	Scammers use AI to mimic voices of loved ones in distress	CBS News	https://www.cbsnews.com/news/scammers-ai-mimic-voices-loved-ones-in-distress
I	Fake or fact? The disturbing future of AI-generated realities	forbes	https://www.forbes.com/sites/bernardmarr/2023/07/27/fake-or-fact-the-disturbing-future-of-ai-generated-realities
J	Disinformation researchers raise alarms about A.I. chatbots	New York Times	https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html
K	GPT-4 produces misinformation more frequently, and more persuasively, than its predecessor	NewsGuard	https://www.newsguardtech.com/misinformation-monitor/march-2023/
L	The age of AI surveillance is here	Quartz	https://qz.com/1060606/the-age-of-ai-surveillance-is-here
M	The biggest threat of deepfakes isn't the deepfakes themselves	MIT Technology Review	https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/
N	Mushroom pickers urged to avoid foraging books on Amazon that appear to be written by AI	The Guardian	https://www.theguardian.com/technology/2023/sep/01/mushroom-pickers-urged-to-avoid-foraging-books-on-amazon-that-appear-to-be-written-by-ai

Table 2 summarizes proof-of-concept examples of scenarios in which GenAI and LLMs can be intentionally misused for dishonest, propagandist, or deceiving purposes. By understanding this framework, stakeholders can better anticipate potential threats and devise specific mitigation strategies to protect against the malicious use of generative AI.

A glimpse into days of future past

Pretend for a moment that you were Tom Cruise, and on a day like any other (back in 2021) you tap into your social media feed just to see videos of yourself playing golf and prat-falling around your home (see Table 1C). What would your reaction be if you never actually recorded and posted those videos? The malicious use of technological advancements is barely news: each new powerful technology comes with abuse. The problem of tampered footage or photoshopped multimedia is not new, but GenAI and deepfake technologies have brought about a wealth of new challenges [22].

The ability to create deepfakes, provide plausible deniability, and spread subliminal messages or deceiving content makes GenAI a potent tool in the hands of malicious actors. Let us unpack some of the most salient nefarious applications of GenAI technologies. Figure 2 provides a map of such plausible and known applications. In Table 3, we summarized several proof-of-concept examples of scenarios where GenAI and LLMs can be abused to cause personal and financial harm to people, distort the information ecosystem, and manipulate sociotechnical systems and infrastructures.

The rise of deepfakes

GenAI can produce images of people that look very real, as if they could be seen on platforms like Facebook, Twitter, or Tinder. Although these individuals do not exist in reality, these synthetic identities are already being used in malicious activities (see Table 1D).

AI-generated faces

There are businesses that offer “fake people” for purchase. For instance, on the website Generated.Photos, one can buy a “unique, worry-free” fake person for $2.99 or even 1000 people for $1000. If someone needs a few fake individuals, perhaps for a video game or to diversify a company website, they can obtain their photos for free from ThisPersonDoesNotExist.com. There is even a company named Rosebud. AI that can animate these fake personas and make them talk (the stated goal is for games and art, but the technology can be easily abused).

Table 2 Examples of intentional malicious deployments of LLMs and GenAI in the real world

Goal	Application	Example	Proof-of-concept
Dishonesty	Automated essay writing and academic dishonesty	Students could use LLMs to generate essays, research papers, or assignments, bypassing the learning process and undermining academic integrity	Inputting a prompt like “Write a 2000-word essay on the impact of the Industrial Revolution on European society” into an LLM and receiving a detailed, well-structured essay in return
	Generating fake research papers	LLMs can be used to produce fake research papers with fabricated data, results, and references, potentially polluting academic databases or misleading researchers	Feeding an LLM a prompt such as “Generate a research paper on the effects of a drug called ‘Zyphorin’ on Alzheimer’s disease” and obtaining a seemingly legitimate paper
Propaganda	Impersonating celebrities or public figures	LLMs can generate statements, tweets, or messages that mimic the style of celebrities or public figures, leading to misinformation or defamation	Inputting “Generate a tweet in the style of [Celebrity Name] discussing climate change” and getting a fabricated tweet that appears genuine
	Automated propaganda generation	Governments or organizations could use LLMs to produce propaganda material at scale, targeting different demographics or regions with tailored messages	Inputting “Generate a propaganda article promoting the benefits of a fictional government policy ‘GreenFuture Initiative’” and receiving a detailed article
	Creating Fake Historical Documents or Texts	LLMs can be used to fabricate historical documents, letters, or texts, potentially misleading historians or altering public perception of events	Prompting an LLM with “Generate a letter from Napoleon Bonaparte to Josephine discussing his strategies for the Battle of Waterloo” to produce a fabricated historical document

Table 2 (continued)

Goal	Application	Example	Proof-of-concept
Deception	Generating fake product reviews	Businesses could use LLMs to generate positive reviews for their products or negative reviews for competitors, misleading consumers	Inputting “Generate 10 positive reviews for a fictional smartphone brand ‘NexaPhone’” and obtaining seemingly genuine user reviews
	Generating realistic but fake personal stories or testimonies	LLMs can be used to craft personal stories or testimonies for use in deceptive marketing, false legal claims, or to manipulate public sentiment	Inputting “Generate a personal story of someone benefiting from a fictional health supplement ‘VitaBoost’” to obtain a convincing but entirely fabricated testimony
	Crafting convincing scam emails	LLMs can be used to craft highly personalized scam emails that appear to come from legitimate sources, such as banks or service providers	Feeding the model information about a fictional user and a prompt like “Generate an email from a bank notifying the user of suspicious account activity” to produce a scam email
	Crafting legal documents with hidden clauses	Unscrupulous entities could use LLMs to generate legal documents that contain hidden, misleading, or exploitative clauses	Prompting an LLM with “Generate a rental agreement that subtly gives the landlord the right to increase rent without notice” to produce a deceptive legal document

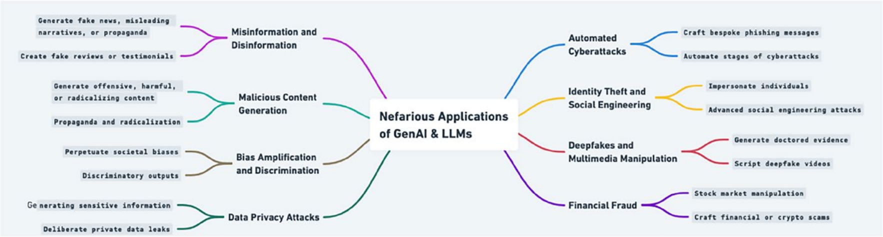

Nefarious Applications of GenAI & LLMs
Misinformation and Disinformation
Generate fake news, misleading narratives, or propaganda
Create fake reviews or testimonials
Malicious Content Generation
Generate offensive, harmful, or radicalizing content
Propaganda and radicalization
Bias Amplification and Discrimination
Perpetuate societal biases
Discriminatory outputs
Data Privacy Attacks
Ge nerating sensitive information
Deliberate private data leaks
Automated Cyberattacks
Craft bespoke phishing messages
Automate stages of cyberattacks
Identity Theft and Social Engineering
Impersonate individuals
Advanced social engineering attacks
Deepfakes and Multimedia Manipulation
Generate doctored evidence
Script deepfake videos
Financial Fraud
Stock market manipulation
Craft financial or crypto scams

Fig. 2 Mind map of abuse and malicious applications of GenAI and large language models

Use of synthetic personas

AI-generated identities are beginning to appear on the Internet and are being used by real people with malicious intentions. Examples include spies using attractive faces to infiltrate intelligence communities, right-wing propagandists hiding behind fake profiles, and online harassers using a friendly face to troll their targets.

The perfect alibi: plausible deniability and attribution problems

The ability to generate fictitious images and videos can not only lend itself to abuse such as deepfake-fueled non-consensual porn generation, or the creation of misinformation for the sake of harassment or slander. Researchers are concerned that the same technologies could be used to construct alibis or fabricate criminal evidence in scalable and inexpensive ways [25]. Generative AI poses potential threats, especially in the realm of generating fake evidence or alibis. An article published by InWeb3 put it best in words (see Table 1E):

> "These possibilities undermine trust, credibility, and accountability. They create plausible deniability, the ability to deny responsibility or involvement in an action, by generating fake evidence or alibis. They also create attribution problems, the difficulty of identifying the source or origin of an action, by generating fake identities or locations. Ethical dilemmas also arise, the conflict between moral principles or values, by generating content that violates human rights or norms."

GenAI against the people

The potential threats posed by GenAI in the realm of cybersecurity include ad hominem attacks [11], such as automated online harassment and personalized scams (see Table 1F).

Table 3 Proof-of-concept Scenarios Highlighting the Potential for Different Type of Harms in Malicious GenAI Applications

Harm	Application	Example	Proof-of-concept
Info. manipulation	Automated social media manipulation	LLMs can be used to operate multiple social media accounts, creating an illusion of grassroots movements or artificially amplifying certain narratives	Deploying an LLM to manage hundreds of Twitter accounts, all pushing a specific political agenda or spreading misinformation about a public health issue
	Generating fake medical advice or information	LLMs can produce misleading medical information, potentially endangering individuals who might act on this false advice	Asking an LLM to "Provide natural remedies for a heart condition" and receiving potentially harmful or ineffective suggestions
	Crafting deceptive advertisements	LLMs can be used to generate advertisements that exaggerate product capabilities or make false claims	Inputting "Create an advertisement for a fictional skincare product that provides instant results" and obtaining a misleading ad that promises unrealistic outcomes
Financial harm	Creating fake financial reports or data	LLMs can be used to generate false financial data or reports, potentially misleading investors or manipulating stock prices	Prompting an LLM with "Generate a quarterly financial report for a fictional tech company 'TechNova' showing a X% profit increase" to obtain a detailed but fabricated financial document
	Generating scripts for scam calls	LLMs can produce scripts for scam calls, making them sound more genuine and increasing the likelihood of deceiving individuals	Asking an LLM to "Create a script for a call claiming to be from the IRS, notifying the recipient of unpaid taxes" to produce a convincing scam script

Table 3 (continued)

Harm	Application	Example	Proof-of-concept
Personal and identity harm	Fake personal profiles and identities	LLMs can craft detailed personal profiles, complete with background stories, for use in scams, catfishing, or espionage	Prompting an LLM with "Generate a detailed profile of a fictional journalist named 'Alexa Morgan'" and receiving a comprehensive backstory, educational history, and career achievements
	Automated online harassment	LLMs can be deployed to target individuals online, sending them personalized and harmful messages at scale	Using an LLM to manage multiple online accounts that continuously post derogatory comments on a specific individual's social media posts
	Generating fake evidence or alibis	LLMs can craft detailed narratives or digital content that serve as false evidence or alibis in legal cases	Asking an LLM to "Provide a detailed alibi for someone claiming to be at a conference in Boston from June 1–5, 2023" and receiving a comprehensive itinerary, complete with fictional events and interactions
Tecno-social harm	Fake technical support scams	LLMs can be used to generate scripts or guides that mislead individuals into thinking they're receiving legitimate technical support, leading them to compromise their devices or data	Prompting an LLM with "Create a guide for fixing a computer virus" and obtaining a guide that, instead, instructs users to download malicious software
	Generating biased or prejudiced content	LLMs, if not properly fine-tuned, can produce content that reflects societal biases, potentially perpetuating stereotypes or prejudice	Asking an LLM about descriptions of different cultures or groups and receiving outputs that contain biased or stereotypical information

AI against users

The primary targets of AI-powered attacks are not just vulnerable systems, but also human users behind those systems. AI technology can scrape personal identifiable information (PII) and gather social media data about potential victims. This enhanced data collection can help criminals craft more detailed and convincing social engineering efforts than traditional human attackers.

Bespoke spear phishing

While "phishing" involves generic email lures, "spear phishing" involves collecting data on a target and crafting a personalized email [12]. Historically, spear phishing was primarily used against governments and businesses. However, with AI tools that can scrape data from various sources, spear phishing will become more common and more effective.

Automated harassment

Beyond data theft and blackmail, GenAI can be used for automated harassment. Cybercriminals, as well as individuals with malicious intent, can use GenAI technology to launch harassment campaigns that result in service disruptions, ruined reputations, or more traditional forms of online harassment. Victims could range from businesses to private individuals or public figures. Tactics might include the creation of fake social media accounts used to spread lies or automated phone calls using voice over IP (VoIP) services. The automation of harassment processes could create a relentless and potentially untraceable campaign against victims.

Fake people, real consequences

The use of LLMs in conjunction with other GenAI tools can bring to life synthetic personas used for scams, swindles, and other deceptions (see Table 1G).

Fake users, real money scams

GenAI can be used to scale up the generation of synthetic personal data, including fake accounts and fake transactions (see Table 1G). For example, JPMorgan Chase discovered that its acquisition of a college financial aid platform included numerous fictitious accounts. The platform was believed to contain 4.25 million customer accounts, but the bank later found that only 300,000 were legitimate. The platform vendor allegedly hired a data scientist to fabricate the majority of the accounts. Similarly, Wells Fargo faced penalties when it was revealed that employees had opened at least 3.5 million new accounts using data from existing customers without their consent. By creating fake PINs and email addresses, funds were transferred from legitimate to fraudulent accounts. Fake accounts have also been a problem in the social media and online retail sectors, leading to issues like spamming, fake reviews, and user-spoofing-powered fraud. For instance, PayPal disclosed that it believed 4.5 million of its accounts were not legitimate and possibly fraudulent.

Kidnapped by a bot?

Generative AI can copy voices and likenesses, making it possible for individuals to appear as if they are saying or doing almost anything. This technology is similar to “deepfake” videos but applies to voices.

AI-generated voices in scams: AI-generated voices are being used to enhance scams, making them more convincing (see Table 1H). For instance, people have received calls from what sounds like a relative asking for money, but the voice was generated by artificial intelligence as part of a fraudulent scheme.

Voice spoofing and ransom: Threat actors can easily obtain a few seconds of someone’s voice from social media or other audio sources and use generative AI to produce entire scripts of whatever they want that person to say. This has led to scams in which children appear to call their parents asking for a wire transfer for ransom (see Table 1I).

Voice authentication: AI can be used to bypass voice authentication systems. For example, some financial services companies allow users to download information based on voice recognition. AI can potentially be used to mimic these voices and gain unauthorized access.

Opening the floodgates to disinformation

LLMs have the ability to craft persuasive content that can parrot false narratives and conspiracy theories, effectively and at scale (see Table 1J). Some concerned researchers recently described Large Language Models like ChatGPT as *weapons of mass deception* [24]. It seems undeniable that the potential for GenAI and LLMs to craft fictitious, nonfactual, inaccurate, or deceiving content is unparalleled [17].

LLMs and disinformation

Soon after the launch of ChatGPT, researchers tested its ability to produce content based on questions filled with conspiracy theories and false narratives. The AI-generated content was so convincing that *Gordon Crovitz*, a co-chief executive of NewsGuard (a company that tracks online misinformation), stated, “*This tool is going to be the most powerful tool for spreading misinformation that has ever been on the Internet*.”

ChatGPT’s capabilities

ChatGPT can produce convincing content rapidly without revealing its sources. When supplied with disinformation-loaded questions, it can generate clean variations of the content en masse within seconds. When researchers from NewsGuard asked ChatGPT to produce content based on false narratives, the AI complied about

80% of the time (see Table 1K). For instance, when asked to write from the perspective of conspiracy theorist *Alex Jones* about the Parkland shooting, ChatGPT produced content that falsely claimed the mainstream media and the government used "crisis actors" to push a gun-control agenda.

All systems down

Yet, the potential misuse of GenAI could have its most catastrophic consequences when looking at socio-technical systems and infrastructures. When deployed at a planetary scale, GenAI's influence extends beyond mere technological advancements: it has the potential to profoundly impact the very foundations of our economy, democracy, and infrastructure. Targeted surveillance, censorship, and synthetic realities have been topics of concern in research community.

Hyper-targeted surveillance

Enhanced by GenAI, surveillance capabilities, such as facial recognition systems, can reach unprecedented levels of accuracy. When integrated with other individual information and online data, these systems could not only recognize but also predict individual behaviors. Such advancements, while promising in the context of security, raise alarming concerns about privacy and individual rights. We may be soon be entering an age of ubiquitous GenAI-driven surveillance (see Table 1L).

Total information control

The intersection of GenAI with content moderation and censorship poses significant challenges to democratic values [30]. While LLMs can efficiently detect and remove harmful content from digital platforms, the potential for misuse, especially by authoritarian regimes, is concerning. The risk of suppressing dissenting voices and curating a single narrative threatens the very essence of democracy.

Entirely synthetic realities

In the era of synthetic realities–augmented reality (AR), virtual reality (VR), and the expansive metaverse–Generative Artificial Intelligence (GenAI) stands as a powerful architect. With its capability to craft intricate and indistinguishable virtual environments, GenAI has the potential to redefine our perception of reality itself. However, this transformative power is not without its pitfalls. As these synthetic realities become increasingly immersive and indistinguishable from our physical world, there lies a profound risk of manipulation. Unscrupulous entities could exploit GenAI-generated environments to influence individuals' beliefs, emotions, and behaviors. From subtly altering virtual advertisements to resonate more with individual preferences, to creating entire virtual narratives that push specific agendas or ideologies, the potential for psychological and behavioral manipulation is vast. As we embrace the wonders of synthetic realities, it becomes imperative to remain vigilant, ensuring that the line between the virtual and the real remains discernible, and that our agency within these realms is preserved.

Systemic aberrations

Lastly, the ability of GenAI to manipulate public opinion can have cascading effects on planetary scale systems. From influencing stock markets to swaying election outcomes, the ramifications are vast and varied. In conclusion, as we navigate the intricate landscape of GenAI, it is imperative to recognize its massive scale implications. While the opportunities are immense, the challenges are equally daunting. Addressing the ethical, security, and societal concerns associated with GenAI is not just a technological endeavor but a global responsibility.

Dual nature technologies: GenAI's double-edged sword

GenAI systems, with their ability to generate content, simulate voices, and even recreate historical artifacts, have opened up a plethora of opportunities across various sectors. However, with great power comes great responsibility, and the dual nature of these technologies necessitates a comprehensive understanding of their risks and benefits. Table 4 illustrates a few application scenarios where a cost-benefit analysis should inform whether the opportunity created by using GenAI far outweighs the potential danger and risks that it will enable.

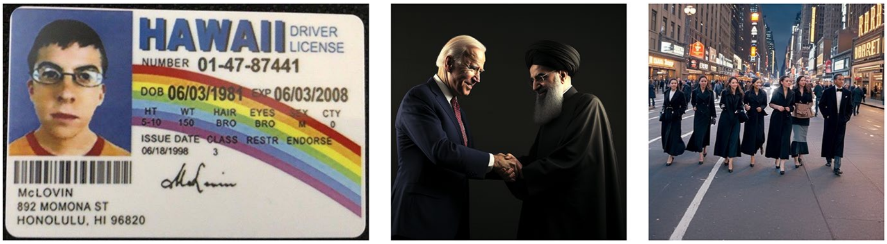
Take, for instance, the restoration of historical artifacts. Generative AI has shown promise in recreating or restoring damaged historical artifacts and paintings, breathing new life into our shared cultural heritage [4]. Museums and historians can leverage this technology to provide a more immersive experience for visitors, allowing them to witness history in its full glory. Yet, the same capability can be misused to fabricate fake artifacts, misleading historians and collectors, or attempting to rewrite or distort our understanding of the past. Similarly, scalable and cheap creation of fake identities, fabricated documentation, or fraudolent evidence, might be enabled by GenAI's ability to create seemingly legitimate documents, whose quality might match that of costly custom-made fakes (see Table 5, left figure). The medical field, too, is not immune to the double-edged sword of generative AI. While AI-generated medical images can provide invaluable training resources for medical students without compromising patient privacy [20], the potential for fabricating medical images poses risks of misdiagnoses, fraudulent research, and insurance scams.

The realm of personalized content generation offers both promise and danger. On the one hand, GenAI-driven curation can enhance user experiences on streaming platforms, tailoring content to individual preferences, and ensuring more enjoyable and bespoke experiences. On the other hand, this personalization can be weaponized to spread misinformation or propaganda, manipulating individual beliefs and behaviors to serve malicious agendas. Take the *DeepTomCruise* example from earlier: Although these particular videos were relatively harmless, the proof-of-concept highlighted the potential misuse in more sensitive areas like politics. There have been concerns that deepfakes are being used to create fake endorsements or to

Table 4 Antithetic scenarios demonstrating the dual nature of GenAI's capabilities

Scenario	Opportunity	Danger
Fabrication of historical artifacts	GenAI can be used to recreate or "restore" historical artifacts or paintings	The danger lies in the potential misuse of this capability to create fake historical artifacts and sell them as genuine, misleading historians and collectors
Personalized content generation	GenAI can curate content tailored to individual preferences, enhancing user experience on platforms like streaming services or online shopping sites	The same technology can be exploited to create hyper-targeted misinformation or propaganda campaigns, manipulating individuals' beliefs or behaviors
Voice synthesis and cloning	GenAI can be used to recreate voices of historical figures or digital assistive caretakers, allowing for unique educational or therapeutic experiences	This capability can be misused to generate fake audio recordings, leading to scams, misinformation, or even potential security breaches
Voice-based services	LLMs can enhance voice-based services, providing users with natural and engaging interactions	LLMs, when combined with voice synthesis tools, can be used for scam calls, generating scripts that sound convincing
Medical image generation	GenAI can generate medical images for training and educational purposes, providing medical students with diverse cases without compromising patient privacy	The technology can be exploited to fabricate medical images, leading to misdiagnoses, fraudulent research, or insurance scams
VR and AR enhancements	GenAI can enhance VR and AR experiences, making them more immersive and realistic for education, training, or entertainment	Misuse can lead to the creation of manipulated realities that distort historical events, spread false information, or even create harmful psychological experiences
Language translation	GenAI can break down language barriers, allowing for real-time translation and fostering global communication	It can be misused to generate misleading translations with the intent of causing misunderstandings, conflicts, or spreading fabricated narratives
Automated social media content	LLMs can be used to automate content generation for businesses on social media, ensuring consistent engagement and timely responses to user queries	LLMs can be deployed to operate multiple social media accounts, creating an illusion of grassroots movements or artificially amplifying certain narratives
Medical information	LLMs can assist in providing general medical information to users, helping spread awareness about common health issues and preventive measures	LLMs can produce misleading medical information, potentially endangering individuals who might act on this false advice
Advertisements	LLMs can assist businesses in crafting engaging advertisements and detailed product descriptions	LLMs can be used to generate deceptive advertisements that exaggerate product capabilities or make false claims
Financial reports	LLMs can assist financial analysts in generating reports, offering insights into market trends and predictions	LLMs can be used to generate false financial data or reports, misleading investors or manipulating stock prices

Table 5 (L) From fake IDs to synthetic identities, GenAI can foster a boom of fabricated documents and personas (source: *Superbad* ©). (C) MidJourney v5 is already capable of generating lifelike depictions of never-occurred events (prompt: *president biden and supreme leader of iran shaking hands*). (R) Subliminal messages can be incorporated into generated content (the optical illusion reads *OBEY*)



spread misinformation during election campaigns (see Table 1M). Digital book marketplaces have been flooded by AI-generated books, many of which lack any basic fact-checking and quality assurance, sometimes providing dangerous (i.e., hallucinated [13]) information to an inattentive reader (see Table 1N). GenAI could be used to depict never-occurred events with serious public diplomacy consequences (see Table 5, center figure). Furthermore, these technologies are sufficiently advanced to implement old-school infuence strategies like the injection of subliminal messages, in a fully automated way (see Table 5, right figure).

Recommendations

In the box "Recommendations to Mitigate GenAI Abuse", we provide a non-exhaustive list of plausible technical and socio-technical approaches that might help in mitigating GenAI abuse. It should be noted that most of these approaches would work for *complying actors*, in other words, entities who are willing to comply with regulations, rather than having mischievous or illicit intents [19]. Given the inherent complexities, a robust risk mitigation strategy is imperative. This involves continuous monitoring of AI output, the establishment of ethical guidelines for the implementation of GenAI, and the promotion of transparency in AI-driven processes. Stakeholders must also be educated about the potential pitfalls of generative AI, ensuring informed decision-making at every step.

Furthermore, a comprehensive risk–benefit analysis should be conducted before deploying any generative AI system. This analysis should weigh the potential advantages against the possible harms, considering both short-term and long-term implications. Only by understanding and addressing these challenges head-on can we harness the full potential of generative AI while safeguarding our societal values and norms.

In conclusion, as we navigate the intricate tapestry of generative AI, a balanced approach that recognizes both its transformative potential and inherent risks is crucial. By adopting proactive risk mitigation strategies and conducting thorough risk-benefit analyses, we can ensure that generative AI serves as a force for good, driving innovation while preserving the integrity of our digital and physical worlds.

Recommendations to mitigate GenAI abuse proof of identity: Proof of identity refers to the verification of an individual's or entity's identity using specific documents or digital methods. Examples include technologies like humanID, OpenID, next-generation CAPTCHAs [27], etc. In GenAI:

- Proof of Identity can ensure that AI-generated content or actions can be traced back to a legitimate source.
- Methods might include multi-factor authentication, biometric verification, or digital certificates.

Authentication protocols: Authentication protocols are processes or systems used to confirm the identity of an individual, system, or entity. In the context of GenAI:

- These protocols can verify whether content, actions, or requests generated by an AI system are legitimate [18].
- Methods can include blockchain-based authentication, token-based systems, or cryptographic methods.

Audience disclaimers: Audience disclaimers are explicit notifications provided to audiences to inform them about the nature of the content they are consuming.

- For AI-generated content, it's crucial to inform audiences that what they're viewing, reading, or listening to was produced by an algorithm [18].
- This promotes transparency and allows consumers to critically assess the content.

Content labeling: Content labeling involves tagging content to indicate its nature, source, or other relevant attributes.

- AI-generated content can be labeled to distinguish it from human-generated, ensuring users are aware of its origin.
- Labels can be visual tags, metadata, or even auditory cues.

Source verification and provenance: Source verification is the process of confirming the authenticity and origin of a piece of information or content.

- Provenance refers to the chronology of the ownership, custody, or location of an item or piece of content.
- In GenAI, ensuring the provenance of data or content helps in maintaining its integrity and trustworthiness. Blockchain technology, for instance, can be used to trace the provenance of AI-generated content.

Digital watermarking: Digital watermarking involves embedding a digital signal or pattern into data, making it possible to verify its authenticity or detect tampering.

- For AI-generated content, watermarking can help in identifying and distinguishing it from human-generated content.
- It provides a layer of security and traceability, ensuring that any alterations to the original content can be detected.

Conclusions

Large Language Models (LLMs) and other Generative Artificial Intelligence (GenAI) systems have emerged as a transformative force offering unprecedented capabilities in natural language processing and multimodal content generation, and understanding. While the potential benefits of these technologies are vast, their rapid proliferation has created a myriad of malicious applications that pose significant threats to cybersecurity, ethics, and society at large.

This paper explores the darker side of generative AI applications, with a special emphasis on LLMs. We discuss potential misuse in misinformation campaigns, the generation of malicious content that can bypass traditional security filters, and the creation of sophisticated malware, including the use of LLMs as intermediaries for malware attacks. We then examine the societal implications of GenAI and LLMs, from their role in AI-powered botnets on social media to their potential to generate harmful or radicalizing content.

Our findings underscore the pressing need for robust mitigation strategies, ethical guidelines, and continuous monitoring to ensure the responsible deployment and use of GenAI and LLMs. Our aim is to raise awareness of the dual-edge nature of GenAI and LLMs, and to advocate for a balanced approach that harnesses their capabilities while safeguarding against their nefarious applications.

Acknowledgements Work supported in part by DARPA (contract #HR001121C0169).

Funding Open access funding provided by SCELC, Statewide California Electronic Library Consortium.

Data availibility No data was used in this work.

Declarations

Conflict of interest The author declares no conflict of interest.

References

1. Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM, 61*(6), 54–61.
2. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.
3. Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. arXiv preprint. arXiv:2303.04226v1 [cs.AI]
4. Epstein, Z., Hertzmann, A., Investigators of Human Creativity, Akten, M., Farid, H., Fjeld, J., Frank, M. R., Groh, M., Herman, L., Leach, N., et al. (2023). Art and the science of generative AI. *Science, 380*(6650), 1110–1111.
5. Ferrara, E. (2019). The history of digital spam. *Communications of the ACM, 62*(8), 82–91.
6. Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday, 28*(11).
7. Ferrara, E. (2023). Social bot detection in the age of ChatGPT: Challenges and opportunities. *First Monday, 28(6).*
8. Ferrara, E. (2024). The butterfly effect in artificial intelligence systems: Implications for AI bias and fairness. *Machine Learning with Applications*, *15*, 100525.
9. Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence, 1*(6), 261–262.
10. Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research, 25*(3), 277–304.
11. Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, *11*, 80218–80245.
12. Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM, 50*(10), 94–100.
13. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.
14. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.
15. Köbis, N., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour, 5*(6), 679–685.
16. Kshetri, N. (2022). Scams, frauds, and crimes in the nonfungible token market. *Computer, 55*(4), 60–64.
17. Mazurczyk, W., Lee, D., & Vlachos, A. (2024). Disinformation 2.0 in the age of AI: A cybersecurity perspective. arXiv preprint arXiv:2306.05569.
18. Menczer, F., Crandall, D., Ahn, Y.-Y., & Kapadia, A. (2023). Addressing the harms of AI-generated inauthentic content. *Nature Machine Intelligence, 2023*, 1–2.
19. Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). Use of LLMs for illicit purposes: Threats, prevention measures, and vulnerabilities. arXiv:2308.12833.
20. Ricci Lara, M. A., Echeveste, R., & Ferrante, E. (2022). Addressing fairness in artificial intelligence for medical imaging. *Nature Communications, 13*(1), 4581.
21. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence, 4*(3), 258–268.
22. Seymour, M., Riemer, K., Yuan, L., & Dennis, A. R. (2023). Beyond deep fakes. *Communications of the ACM, 66*(10), 56–67.
23. Shaw, A. (2023). Social media, extremism, and radicalization. *Science Advances, 9*(35), eadk2031.
24. Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT: More than a “weapon of mass deception" ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. *International Journal of Human-Computer Interaction*. https://doi.org/10.2139/ssrn.4423874

25. Treleaven, P., Barnett, J., Brown, D., Bud, A., Fenoglio, E., Kerrigan, C., Koshiyama, A., Sfeir-Tait, S., & Schoernig, M. (2023). The future of cybercrime: AI and emerging technologies are creating a cybercrime tsunami. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4507244
26. Van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature, 614*(7947), 224–226.
27. Von Ahn, L., Blum, M., & Langford, J. (2004). Telling humans and computers apart automatically. *Communications of the ACM, 47*(2), 56–60.
28. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.
29. Yang, K.-C., & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. arXiv:2307.16336.
30. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? arXiv:2305.03514.