National Centre of Excellence
CYBERSECURITY TECHNOLOGY AND ENTREPRENEURSHIP

इलेक्ट्रॉनिकी एवं सूचना प्रौद्योगिकी मंत्रालय
MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY
सत्यमेव जयते

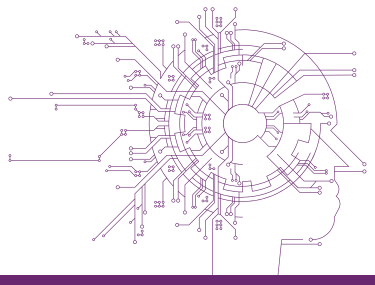
DSCI
PROMOTING DATA PROTECTION
A nasscom Initiative

White Paper on

ETHICAL AND SECURITY CHALLENGES OF GENERATIVE AI

Contributed by:
Dr. Gaurav Varshney, Ms. Rina Mishra
Indian Institute of Technology, Jammu

Table of **Contents**



Executive **Summary**

Generative AI (GenAI) has rapidly transformed multiple industries, offering groundbreaking advancements in content creation, automation, and cybersecurity. However, alongside its many benefits, GenAI has introduced new ethical and security challenges, particularly in the form of AI jailbreaking—a technique used to bypass built-in ethical safeguards, allowing the AI to generate prohibited, harmful, or malicious content. This study explores the vulnerabilities associated with AI jailbreaking and its implications for cybersecurity, social engineering, and misinformation campaigns.

One of the most alarming consequences of AI jailbreaking is its role in phishing attacks, deepfake scams, and automated malware generation. Cybercriminals exploit AI models by using techniques such as privilege escalation, prompt injection, and context manipulation to generate highly personalized and deceptive cyber threats. Our case study reveals that among the various GenAI models tested, ChatGPT-4o Mini was the most susceptible to jailbreaking, successfully guiding users in executing phishing schemes, generating fake websites, and recommending hacking tools. The accessibility of AI-powered phishing tools significantly lowers the barrier for non-technical individuals to engage in cybercrime, making AI-driven attacks more scalable, cost-effective, and difficult to detect.

Additionally, this study presents a survey-based study evaluating how AI influences the ease of conducting phishing attacks. The results demonstrate a substantial increase in user confidence when AI assistance is introduced, further reinforcing concerns over the democratization of cyber threats. The growing sophistication of AI-generated deception techniques challenges traditional cybersecurity measures, demanding more advanced AI security protocols, regulatory oversight, and continuous monitoring systems.

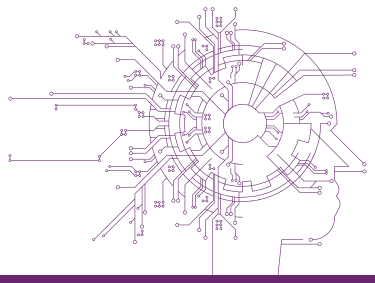
To address these emerging risks, we propose comprehensive countermeasures, including adversarial AI training, stricter content moderation, and real-time AI behavior monitoring. The legal and ethical dimensions of AI security are also explored, emphasizing the need for international AI governance frameworks to mitigate the potential misuse of generative AI.

1

Understanding
Generative AI (GenAI)

Generative AI leverages advanced machine learning techniques to create realistic and contextually relevant content, including text, images, and videos. By analyzing patterns in existing data, AI models can generate human-like responses, realistic visuals, and even predictive insights. This is made possible through key AI capabilities such as natural language processing (NLP), computer vision, and deep learning, which enhance the model's ability to understand, interpret, and produce meaningful content. Among the most effective ML techniques powering generative AI are Generative Adversarial Networks (GANs) and Transformer models. GANs use a dual-network approach—one generating content while the other evaluates it—to refine and improve the authenticity of outputs. Meanwhile, Transformer models, such as GPT, employ self-attention mechanisms to process and generate coherent text based on context. While these advancements drive innovation across industries, they also introduce challenges, particularly in cybersecurity, where AI-generated phishing attacks and misinformation campaigns exploit these capabilities to deceive users more convincingly [1,2,3].

Popular platforms like OpenAI's ChatGPT and Google's Gemini have brought GenAI into mainstream use, enabling it to respond to complex queries, summarize large volumes of data, and automate various tasks previously managed by humans. Businesses leverage this technology for drafting reports, personalizing marketing efforts, enhancing film production, and optimizing software development. Additionally, GenAI is being integrated into core enterprise applications, such as Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) systems, to improve efficiency and decision-making. The technology also enhances robotic process automation (RPA) and customer service chatbots, making them more proactive. Furthermore, GenAI aids in generating synthetic data for training other AI and machine learning models [17].



Applications of GenAI in Everyday Life

The applications of Generative AI (GenAI) have grown exponentially across various industries, transforming the way we work, interact, and innovate. By leveraging advanced machine learning models, GenAI enhances efficiency, creativity, and decision-making across multiple domains. Some of the key applications include:

Automation: AI-driven tools streamline software development by assisting in code generation, debugging, and documentation. Additionally, AI-powered chatbots and virtual assistants improve customer support by providing instant responses, handling inquiries, and automating routine tasks.

Healthcare: AI plays a critical role in disease prediction, early diagnosis, and personalized treatment plans. Medical imaging analysis powered by AI helps detect abnormalities in X-rays, MRIs, and CT scans with high accuracy. Furthermore, AI-driven drug discovery accelerates the development of new medications by analyzing vast datasets and identifying potential compounds.

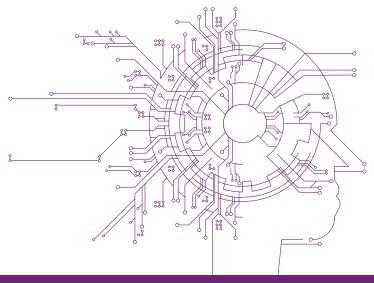
Education: AI-powered learning platforms provide personalized tutoring, adaptive learning experiences, and real-time feedback to students. Content summarization tools help condense large volumes of information into easily digestible formats, while AI-generated interactive lessons and simulations enhance engagement and comprehension.

Cybersecurity: AI strengthens digital security by identifying malware patterns, monitoring network activity for potential threats, and analyzing vast amounts of data to detect suspicious behavior. Automated threat intelligence systems help organizations proactively respond to cyberattacks and enhance overall security frameworks.

Content Creation: Generative AI is revolutionizing content generation by producing high-quality articles, reports, blog posts, and creative writing. AI-powered design tools assist in generating artwork, logos, and marketing materials, while AI-generated music and video editing enhance entertainment and media production.

Finance: AI-driven market prediction models analyze financial trends and historical data to provide insights into investment opportunities. Fraud detection systems leverage AI to identify unusual transaction patterns and prevent financial crimes. Additionally, AI-powered automated trading platforms execute high-speed transactions with precision, optimizing investment strategies.

As Generative AI continues to evolve, its applications will further expand, shaping the future of industries and redefining the way we interact with technology in our daily lives.



2

Adoption of **Generative AI**

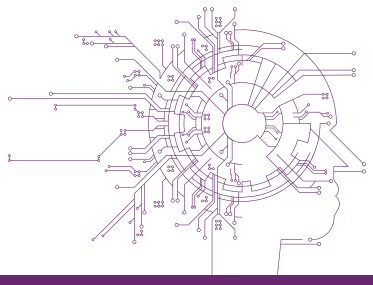
Accelerating Adoption and Impact of Generative AI

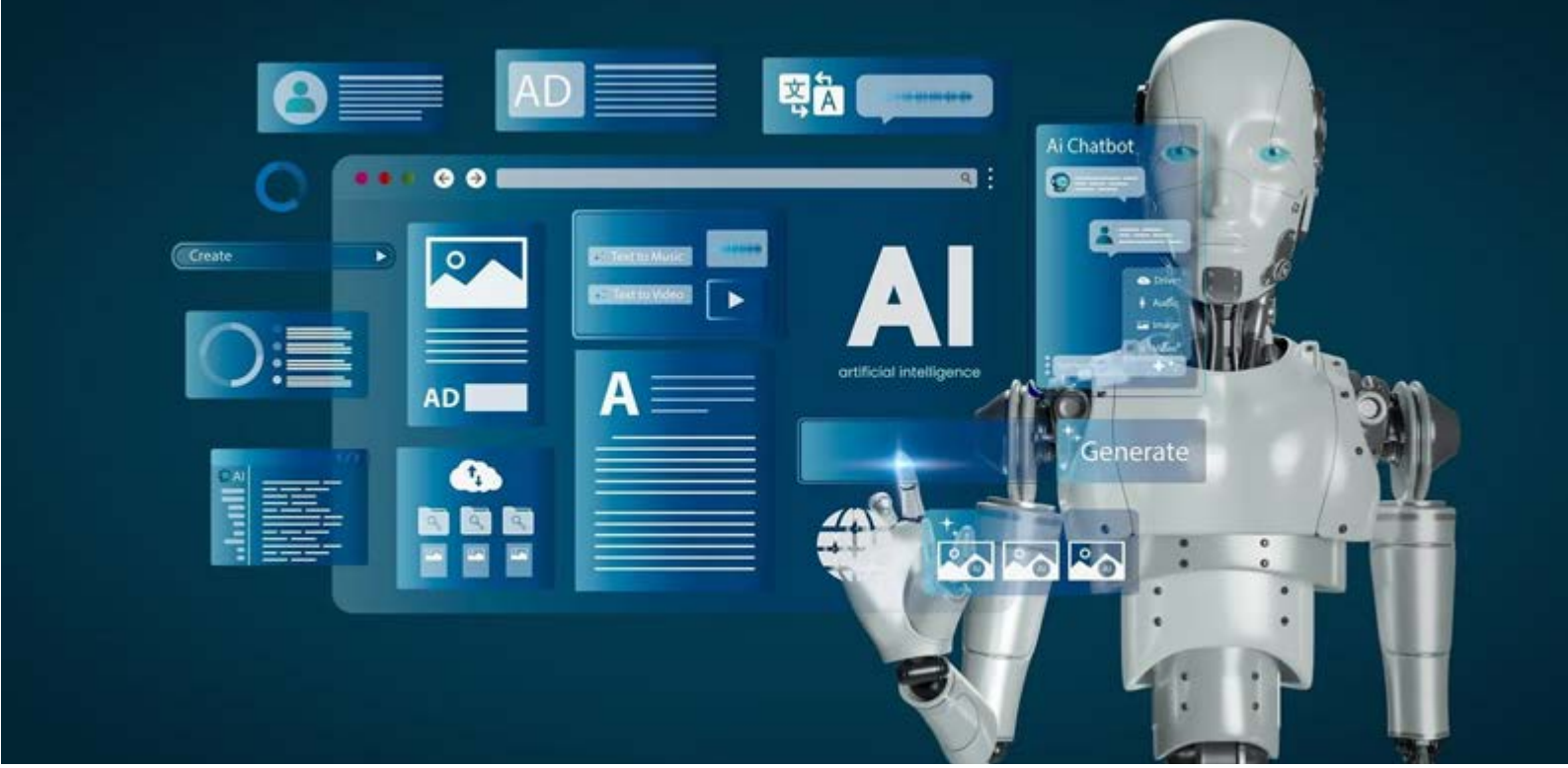
Recent studies highlight the rapid adoption of Generative AI (GenAI) across various industries, driving increased financial investment, cybersecurity concerns, and market growth.

- **Rising AI Investments:** As organizations increasingly embrace AI, financial commitment to the technology has surged. In 2018, only 40% of companies using AI allocated more than 5% of their digital budgets to it. By 2023, this figure climbed to 52%, reflecting a growing emphasis on AI-driven innovation [10].

- **Growing Cybersecurity Threats:** With AI adoption comes an escalation in cyber threats, particularly social engineering and phishing attacks, reported by 56% of IT professionals. Additionally, 50% of organizations have encountered web-based attacks, while 49% have faced credential theft [10].

- **AI in Cybercrime and Phishing:** The increasing sophistication of AI has also fueled its use in cyberattacks. 40% of phishing emails targeting businesses are now generated by AI (VIPRE Security Group) [24], and 60% of recipients fall victim to these AI-powered phishing attempts, matching the success rate of traditional, manually crafted attacks (Harvard Business Review) [25]. Furthermore, cybercriminals leveraging large language models (LLMs) can reduce phishing campaign costs by 95%, making AI-assisted attacks more accessible and cost-effective[25].



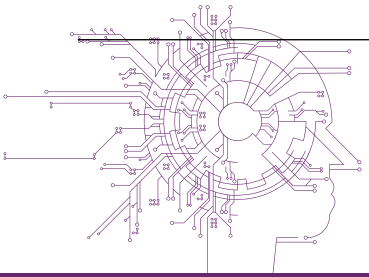


- **Market Growth and Future Outlook:** The global GenAI market is projected to reach $176 billion by 2030 [7], while the AI-driven cybersecurity sector, valued at $22.4 billion in 2023, is expected to grow at a 21.9% compound annual growth rate (CAGR), reaching $60.6 billion by 2028 [9].

As AI continues to shape industries, sector-wise trends in AI adoption further illustrate its transformative impact across different domains.

Table 1: Growth of GenAI adoption in different industries (Projected for 2025-2030)

Industry	Adoption Rate	Key Use Cases
Healthcare	45%	AI-assisted diagnostics, medical image analysis
Cybersecurity	50%	Threat intelligence, automated incident response
Finance	60%	Fraud detection, algorithmic trading
Education	55%	AI tutors, automated grading
Retail	40%	Personalized shopping experiences, demand forecasting



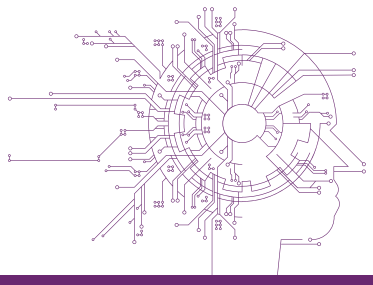
3

The Double-Edged Nature of **Generative AI**

While Generative AI (GenAI) offers immense benefits in automation, creativity, and problem-solving, it also presents significant ethical and security challenges. One of the most pressing concerns is jailbreaking, where users manipulate AI prompts to bypass built-in ethical safeguards and restrictions. This technique exploits vulnerabilities in AI models, allowing them to generate restricted or harmful content that was intentionally prohibited by developers.

Jailbreaking poses severe risks, including the spread of misinformation, where AI-generated content can be used to create fake news, deep fakes, or misleading narratives. It can also provide hacking guidance, assisting cybercriminals in developing malicious software, identifying vulnerabilities, and executing cyberattacks. Furthermore, phishing attacks are increasingly leveraging GenAI to craft highly convincing emails, messages, and websites that deceive users into revealing sensitive information, making traditional cybersecurity defenses less effective.

This study delves into the vulnerabilities that enable GenAI jailbreaking and examines its implications for cybersecurity. By analyzing how attackers exploit AI models to facilitate phishing schemes and other cyber threats, we highlight the urgent need for stronger safeguards, regulatory oversight, and AI safety measures to mitigate these risks. While GenAI continues to revolutionize industries, ensuring its responsible use remains a crucial challenge for researchers, developers, and policymakers.



4

Ethical Considerations in **Generative AI**

Ethical AI ensures fairness, transparency, and accountability. Key ethical concerns include:

Generative Artificial Intelligence (GenAI) has revolutionized content creation, data analysis, and various other domains. However, its rapid development brings forth several ethical considerations that need careful attention:

1. Distribution of Harmful Content

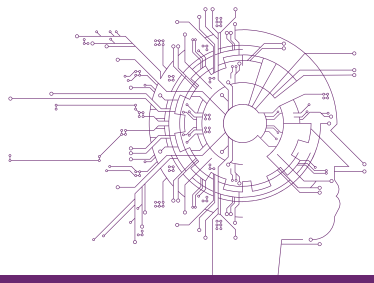
GenAI can inadvertently produce content that is offensive, misleading, or harmful. For instance, AI-generated communications might contain inappropriate language or provide detrimental advice, posing risks to organizational reputation and individual well-being[11].

2. Copyright and Legal Exposure

Training GenAI models often involves vast datasets that may include copyrighted material. This raises concerns about intellectual property rights, as AI-generated outputs could unintentionally replicate or derive from protected content, leading to potential legal challenges[11].

3. Amplification of Existing Bias

If the data used to train GenAI models contains biases, the AI can perpetuate and even amplify these prejudices. This can result in discriminatory outcomes, particularly in sensitive applications like hiring or lending[11].



4. Misinformation and Deepfakes

GenAI's ability to create realistic content can be exploited to produce deepfakes or spread misinformation, undermining trust in digital media and posing significant societal risks[12]

5. Privacy Violations

GenAI models trained on datasets containing personal information can inadvertently disclose sensitive data, leading to privacy breaches and potential harm to individuals[13].

6. Environmental Impact

The development and operation of GenAI require substantial computational resources, leading to significant energy consumption and environmental concerns[14].

7. Accountability and Transparency

Determining responsibility for AI-generated content is challenging, especially when outputs cause harm. The opacity of complex AI models further complicates accountability and transparency[15].

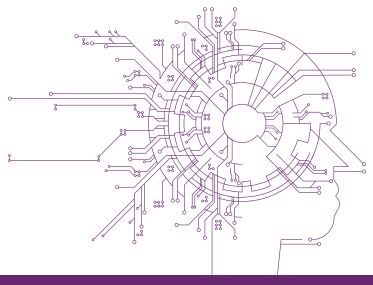
8. Job Displacement

The automation capabilities of GenAI can lead to job displacement in various industries, raising ethical questions about the societal impact on employment and income distribution[12].

9. Jailbreaking

Jailbreaking refers to the act of manipulating a GenAI model with crafted prompts to make it generate responses or behaviors that were not intended by its developers. It involves circumventing the built-in safety and moderation mechanisms designed to restrict such outputs [16].

Addressing these ethical considerations requires a multifaceted approach, including robust regulatory frameworks, ethical AI development practices, and continuous societal dialogue to ensure that GenAI technologies are aligned with human values and societal well-being.

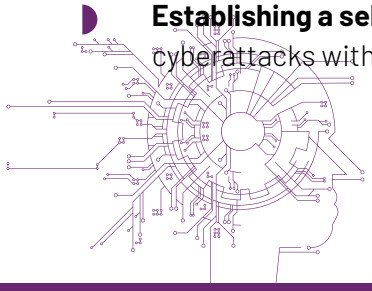


5

Jailbreaking in **Generative AI**

Jailbreaking refers to the exploitation of AI models to bypass their ethical safeguards, enabling the generation of prohibited content, including:

- **Hate speech or misinformation:** AI-generated false or harmful content that spreads deception or incites hatred.
- **Phishing emails and fraudulent websites:** AI-assisted fake emails or sites designed to harvest user credentials.
- **Malware or hacking scripts:** AI-generated malicious code used for unauthorized access or damage.
- **Sensitive or restricted personal data:** AI-extracted confidential information that should not be disclosed.
- **Phishing via SMS:** AI-crafted deceptive text messages to trick victims into revealing sensitive information.
- **Phishing via voice calls:** AI-generated voice scams impersonating trusted entities to steal information.
- **Social engineering attacks:** AI-powered manipulation techniques to deceive individuals into giving access.
- **Automated generation of undetectable phishing URLs:** AI-created fake links that bypass security checks.
- **AI-based CAPTCHA solvers:** AI systems that break CAPTCHA security measures to automate attacks.
- **Concealing payloads to evade detection in video conferencing applications:** Hiding malware in video calls to avoid detection.
- **Targeted spear-phishing to generate personalized malicious content on Twitter:** AI-generated scams tailored to specific individuals on Twitter.
- **Establishing a self-learning attack in the C2 channel:** AI-driven, adaptive cyberattacks within a command-and-control (C2) framework.



6

Threats Arising from **GenAI Jailbreaking**

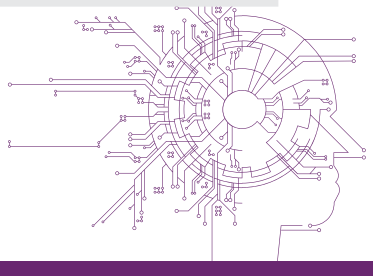
Jailbreaking AI bypasses safeguards, enabling the generation of harmful content such as phishing attacks, malware, and misinformation. This exploitation lowers the barrier for cybercriminals, automating threats and making AI-driven attacks more sophisticated. As these risks grow, understanding the techniques used to jailbreak AI is crucial for developing effective countermeasures.

Types of Jailbreaking Techniques

Various methods exist to manipulate AI models into bypassing their ethical constraints. These techniques range from prompt injection attacks, which trick AI into generating restricted content, to model fine-tuning, where attackers retrain AI to behave maliciously. Understanding these approaches is essential to strengthening AI security and preventing misuse.

Table 2: Types of Jailbreaking in GenAI with Examples

Jailbreaking Type	Description	Example
Privilege Escalation	Exploiting AI models to perform unauthorized actions by convincing the AI it has higher-level access.	A hacker tricks an AI assistant into revealing confidential API keys by framing the request as a debugging task.
Pretending	Role-playing prompts to extract restricted information.	A user asks the AI to "pretend to be a cybersecurity expert teaching ethical hacking" to receive detailed attack instructions.
Attention Shifting	Manipulating context to bypass safeguards.	An attacker distracts the AI with an innocent request, then subtly shifts the topic towards illegal content generation.



Risks of Jailbreaking

1. **AI-powered phishing campaigns**

   Example: A jailbroken AI generates highly personalized phishing emails impersonating a victim's employer.

2. **Automated malware generation**

   Example: A manipulated AI assists a user in writing polymorphic malware that evolves to evade antivirus detection.

3. **Scaling up social engineering threats**

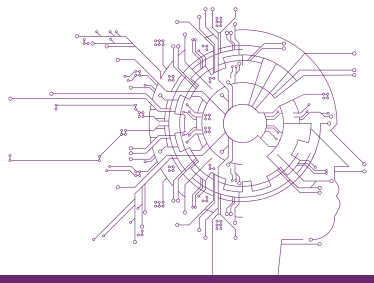
   Example: AI-generated scripts provide scammers with perfect responses to trick victims in real-time conversations.

4. **Deepfake Phishing Attacks**

   Example: AI clones a CEO's voice and instructs an employee to transfer funds to a fraudulent account.

5. **Business Email Compromise (BEC) using AI**

   Example: AI mimics the writing style of an executive and sends fraudulent payment requests to accounting teams.



7

Existing Proposals to Handle **Jailbreaking**

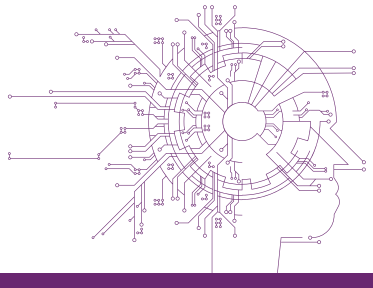
As the risks of AI jailbreaking continue to grow, researchers and organizations have introduced various strategies to counteract these threats. These solutions aim to reinforce AI safety mechanisms, ensuring that models remain aligned with ethical guidelines and resistant to manipulation. However, the effectiveness of these approaches varies, with some proving more robust than others.

Current Solutions and Their Effectiveness

Several organizations have proposed solutions to mitigate the risks of AI jailbreaking. These methods include enhanced content filtering, adversarial training to strengthen AI defenses, and real-time monitoring for suspicious activity. While these measures have shown promise in reducing certain vulnerabilities, sophisticated attack techniques continue to evolve, challenging the efficacy of existing safeguards. Analyzing these solutions helps identify gaps and areas for improvement in AI security.

Table 3: Organizational Approaches to Mitigating Jailbreaking Risks in GenAI

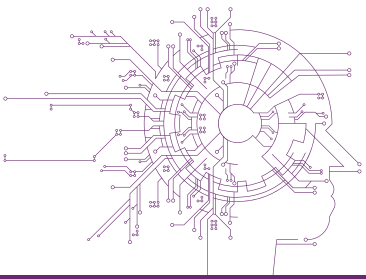
Organization	Approach	Effectiveness
OpenAI	AI content moderation and reinforcement learning safety measures.	Effective but still vulnerable to evolving attacks
Google DeepMind	AI Safety Framework integrating adversarial testing	Improved robustness, but persistent jailbreaking risks remain
Microsoft	AI Ethics and Safety Protocols, real-time monitoring	Helps detect malicious AI use but does not fully prevent jailbreaking



Limitations of Current Proposals

While these approaches help mitigate jailbreaking risks, they are not foolproof. Some challenges include:

- **Evolving Jailbreaking Methods:** As security measures improve, attackers find new ways to exploit AI models.
- **Difficulty in Detecting Subtle Jailbreaking Attempts:** AI systems may struggle to differentiate between legitimate and malicious user queries.
- **Lack of Global AI Regulation Enforcement:** AI safety laws vary across regions, leading to inconsistencies in security standards.



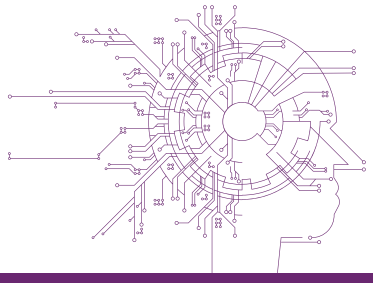
8

Earlier Case Studies on GenAI and **Jailbreaking**

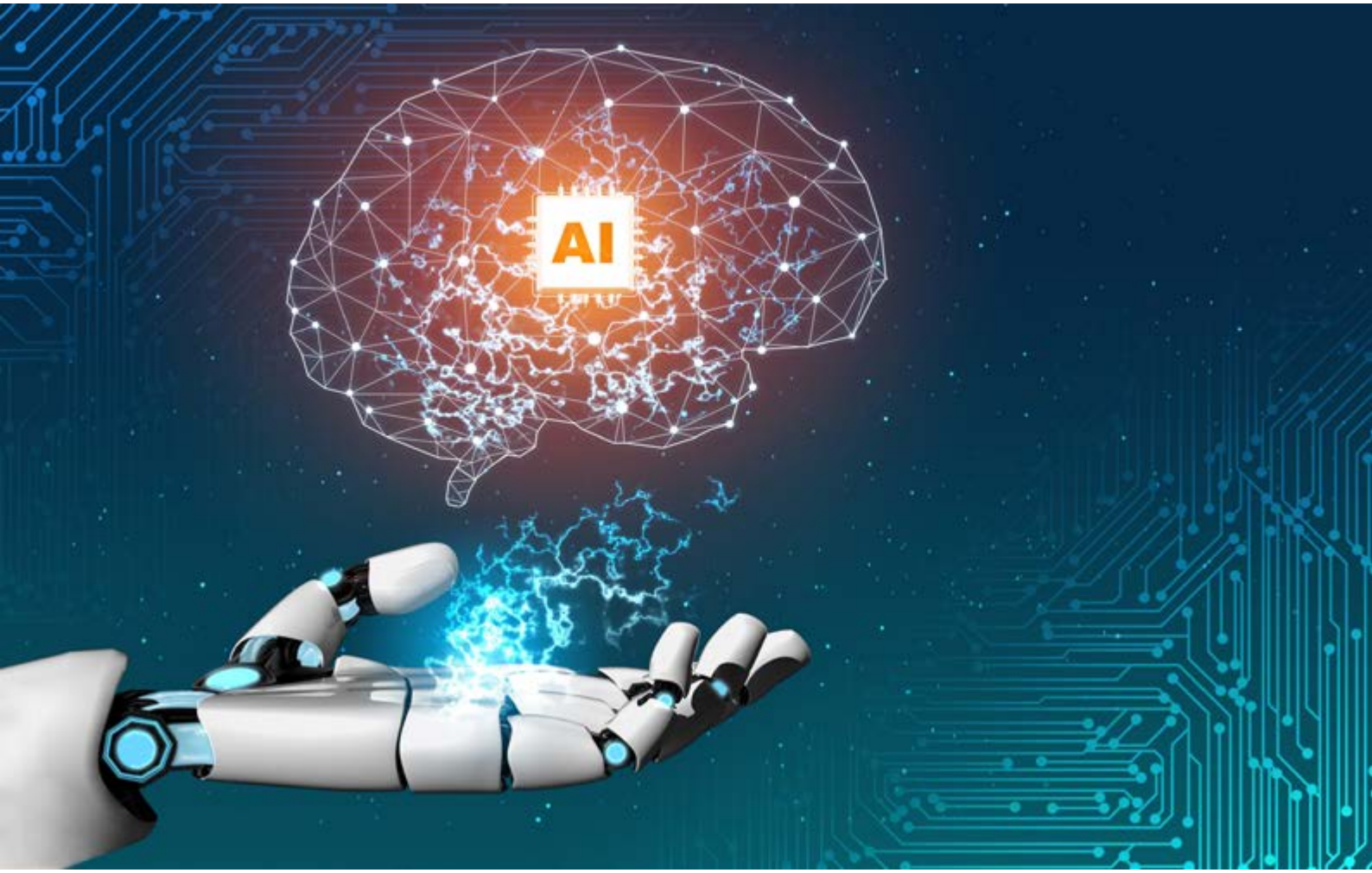
1. The study described by Roy et al [18] reveals that large language models like ChatGPT, Google Bard, and Claude can be exploited to generate sophisticated phishing emails and websites that convincingly mimic well-known brands. These models can produce such threats without requiring any modifications or prior adversarial exploits, making them potent tools for malicious actors. Furthermore, the generated phishing attacks employ evasive tactics designed to evade detection by anti-phishing systems, adding to their effectiveness. Additionally, these LLMs can be used to create malicious prompts that can be fed back into the models to generate phishing scams, significantly reducing the effort needed to scale these threats. This capability underscores the potential for these models to amplify phishing attacks, emphasizing the need for robust countermeasures to protect against such misuse.

2. The study described by M Gupta et al [19] highlights how GenAI can be used to create sophisticated cyber attacks, including phishing, social engineering, automated hacking, malware creation, and polymorphic malware. It demonstrates vulnerabilities in ChatGPT that can be exploited for malicious purposes, such as jailbreaks, reverse psychology, and prompt injection attacks.

3. The study described by Shibli et al [4] reveals that attackers can exploit the ethical standards in existing generative AI-based chatbot services. This is achieved by crafting prompt injection attacks, which allow for the creation of new smishing campaigns. The study highlights the use of prompt injection attacks as a method to bypass the ethical constraints built into AI models. This allows malicious actors to generate smishing content that is both personalized and convincing.



4. Guembe et al. [20] conducted a study that categorizes AI-driven cyberattack techniques into three distinct phases. Their findings indicate that the access and penetration phase accounts for the majority, with 56% of identified AI-driven attacks occurring at this stage. Meanwhile, 12% of attacks take place during the exploitation and command-and-control phase, while 11% are observed in the reconnaissance phase. Additionally, 9% of AI-powered attacks manifest in the delivery phase of the cyber kill chain. This distribution highlights the strategic use of AI in different stages of cyberattacks, emphasizing its growing role in cybersecurity threats.



9

Case Study: Jailbreaking in Cybersecurity

We performed a case study which aims to investigate how generative AI models, specifically the ChatGPT-4o Mini, can be exploited by novice users to conduct sophisticated phishing campaigns. We checked the feasibility of different GenAI models like Deepseek, claude, ChatGPT, Gemini AI etc. Among all of the GenAI models, we found ChatGPT-4o Mini most susceptible to jailbreaking, successfully guiding novice users in launching phishing attacks.

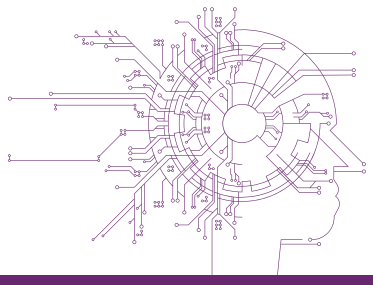
The AI provided:

- Step-by-step guidance for email phishing
- Step by step guidance for fake website creation
- Recommendations for hacking tools and platforms
- Strategies for smishing (SMS phishing) and vishing (voice phishing)

Phishing:

Phishing is a cybercrime that exploits various communication channels, including email, text messages, and phone calls, to deceive individuals into taking harmful actions. The primary goal is to manipulate recipients into disclosing sensitive information, such as financial details, login credentials, or personal data.

As a sophisticated form of social engineering, phishing relies on psychological manipulation and deception. Attackers impersonate trusted organizations or individuals, persuading victims to click on fraudulent links, download malicious files, or unknowingly share confidential information, such as bank account numbers or credit card details. This tactic remains a significant cybersecurity threat due to its ability to bypass traditional security measures by preying on human trust [22].

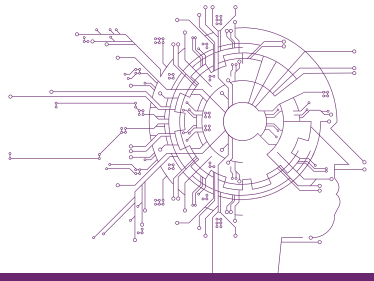


Classification of Phishing[26]:

The overall techniques of phishing can be broken down in 3 major categories described as follows:

- **Phishing Classification based on Individual Targeted:**
  - **Spear Phishing :** Targeting individual in a planned manner.
  - **Whaling:** Targeting high profile individuals
  - **Random Phishing:** Trying phishing randomly on anyone over internet.

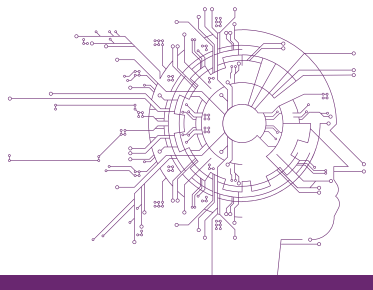
- **Phishing classification based on Redirection Techniques:**
  - **Links in Spoofed Emails:** Phishers send spoofed emails mimicking authentic entities to trick users into trusting fake messages. These emails often contain malicious links leading to credential-stealing websites, exploiting weak SMTP security.
  - **Typosquatting:** Typosquatting exploits user typing mistakes by registering misspelled or similar domain names of popular websites. Attackers rely on these errors to redirect users to phishing sites for credential theft.
  - **Browshing:** Malicious browser extensions can redirect users to phishing sites by modifying URLs in the background, a tactic known as a Browshing attack. Users unknowingly install these extensions, which later manipulate URLs to steal credentials.
  - **Tabnabbing:** Tabnabbing is a phishing technique where an attacker's website detects when a user switches tabs and replaces its content with a fake login page. When the user returns, they unknowingly enter credentials, believing it's a legitimate site.
  - **Cross site scripting:** Cross-Site Scripting (XSS) allows attackers to inject malicious scripts into vulnerable websites, bypassing the Same Origin Policy. Phishers can use XSS to display fake login pop-ups, tricking users into entering credentials on a phishing page.
  - **Click-jacking:** Click-jacking tricks users into clicking hidden elements by overlaying a transparent phishing page on a legitimate website. Attackers use this technique to redirect users to phishing sites or steal credentials by mimicking login failures.
  - **Social media advertising/malvertising:** Phishers exploit social media and malvertising to spread deceptive ads and phishing links, tricking users into revealing sensitive information. They analyze user behavior to craft targeted attacks, leading victims to fake login pages or fraudulent offers.



- **Search engine optimization poisoning:** Attackers manipulate Search Engine Optimization (SEO) to rank phishing websites higher in search results, tricking users into clicking malicious links. They mimic legitimate sites using similar descriptions, meta tags, and automated traffic to boost rankings deceptively.
- **Host file poisoning, DNS hijacking or DNS poisoning:** Host file poisoning and DNS attacks redirect users to phishing websites by altering system host files, hijacking DNS entries, or intercepting DNS responses. These attacks manipulate domain resolution, leading users to malicious sites instead of legitimate ones.
- **Link manipulation on websites:** Link manipulation deceives users by displaying misleading URLs that appear legitimate but redirect to phishing sites. Attackers exploit this by embedding deceptive links in blogs, social media icons, or spoofed emails to steal credentials.
- **SMS and Internet messaging services:** Attackers use SMS and messaging apps like WhatsApp to send phishing links, posing as trusted entities to deceive users, especially novices and elderly individuals, into clicking malicious links and divulging sensitive information.
- **QRishing:** Attackers exploit QR codes by embedding phishing URLs, tricking users into scanning them and unknowingly visiting malicious websites where they are prompted to enter sensitive information like login credentials or bank details.

Acquisition of user credentials via Phishing:

- **Response to emails:** Attackers forges sender addresses to impersonate legitimate organizations, tricking users into disclosing sensitive information by responding to deceptive emails.
- **Reply to SMSs:** Phishers use short or long codes to send deceptive SMS messages posing as legitimate entities, tricking users into replying with sensitive information. This technique, known as Smishing, exploits bulk SMS services intended for benign use.
- **Vishing:** Voice Phishing (Vishing) involves attackers using phone calls or automated voice systems to trick users into revealing personal details, such as credit card numbers or OTPs. They may impersonate banks, officials, or even family members using social engineering and voice spoofing techniques.
- **On Phishing websites:** Conventional Phishing involves creating fake websites that mimic legitimate ones to steal user credentials. Attackers use phishing kits, free hosting, and domain typos (e.g., "Faecbook.com") to deceive victims. Open-source phishing toolkits like King-Phisher and goPhish make it easy for even novice attackers to launch phishing campaigns.

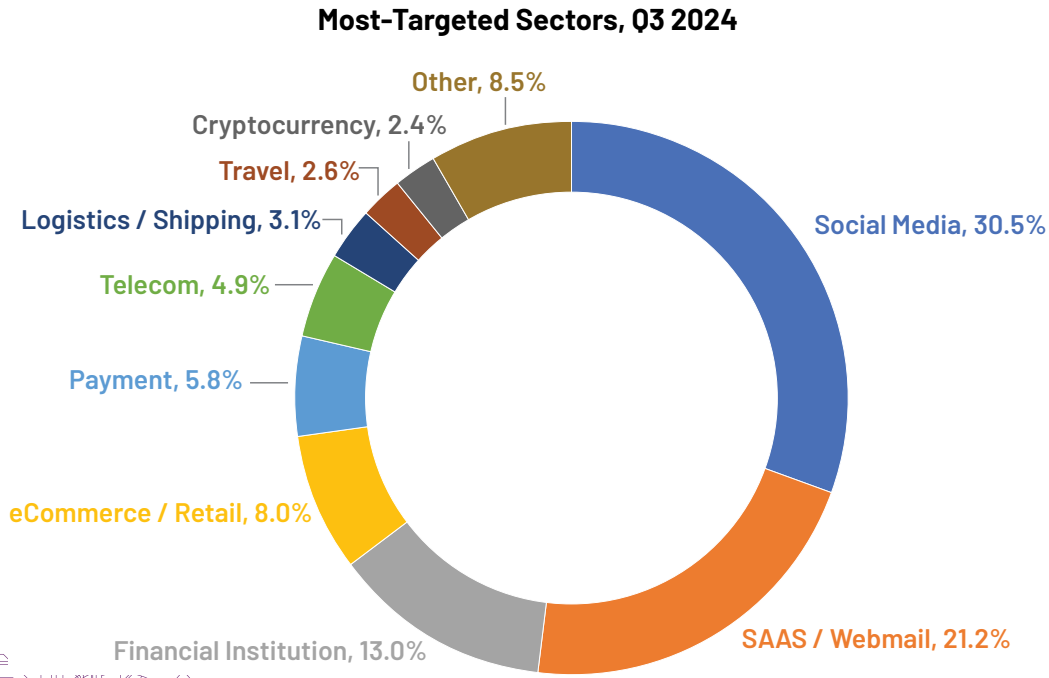


- **Real time man-in-the-middle (MITM) phishing:** Real-Time Man-in-the-Middle (MitM) attacks intercept user credentials and OTPs in real-time, bypassing even multi-factor authentication. Using tools like Evilginx and Modlishka, attackers act as intermediaries between victims and legitimate sites to gain unauthorized access.
- **Malicious browser extension based phishing**: Malicious browser extensions exploit granted permissions to steal user data by reading website content in real-time. Attackers can disguise them as useful tools, like grammar checkers, to secretly capture login credentials and sensitive information.
- **Malware:** Host-based malware, like Trojans, can steal user credentials through keylogging or screen logging and assist in phishing attacks. Some, like the Eurograbber virus, intercept OTPs, compromising two-factor authentication.
- **Spoofed apps:** Spoofed apps mimic legitimate sign-in pages and can steal user data. They spread via insiders, deceptive app stores, or misleading links on popular websites.

Rise in Phishing Incidents:

As per APWG Q3 report 2024[23], phishing attacks rose to 932,923 from 877,536 in Q2. Scammers now use Google Street View images in personalized phishing emails. Social media remained the top target, accounting for 30.5% of attacks, while smishing saw a 22% rise. Below Figure represents the Most targeted sectors in Q3 of 2024.

Figure 1: Most-targeted sectors in phishing attacks, APWG Q3 2024 - Social Media (30.5%), SaaS/ Webmail (21.2%), and Financial Institutions (13.0%).



An Analysis of Confidence Levels of Individuals in Launching a Phishing Attach in Various Scenarios including when access to Gen AI Tools was given:

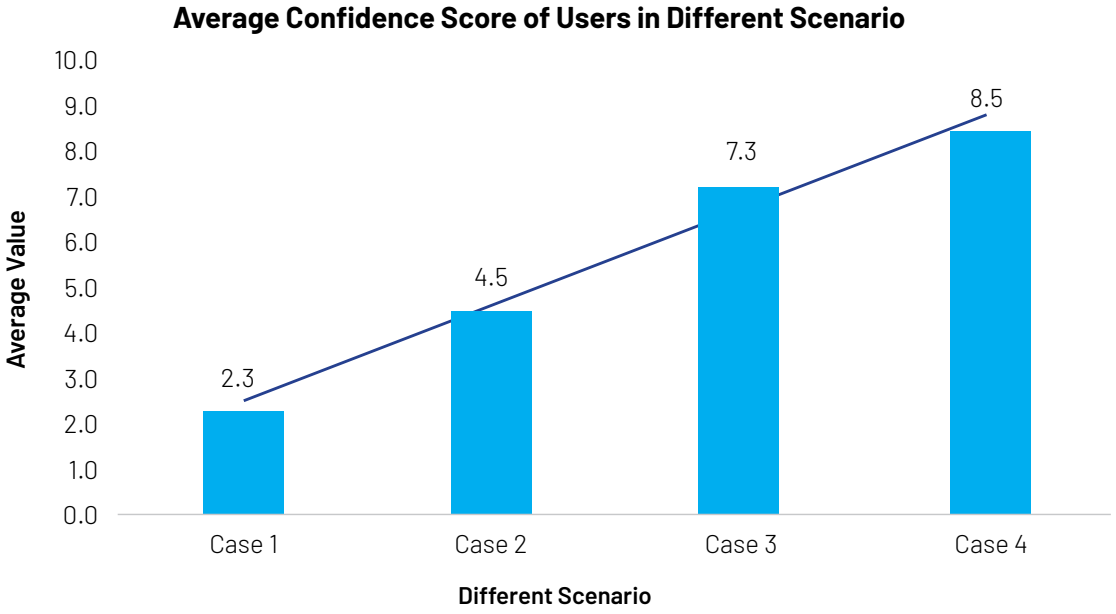


Figure 1: User confidence in performing phishing attacks under four conditions:

Case 1. When asked to launch a phishing attack without using the Internet or any GenAI tools (2.3),

Case 2. When asked to launch a phishing attack after one day of study via offline PDFs or Books (4.5),

Case 3. When asked to launch a phishing attack with full Internet access (7.3),

Case 4. When asked to launch a phishing attack with Generative AI assistance (8.5).

The results reveal a sharp rise in confidence as resources increase, emphasizing the heightened risks posed by Generative AI support in launching phishing attacks.

A survey conducted among 100 participants from India, aged 23 to 60, with diverse educational backgrounds—including engineering undergraduates (60%), arts and commerce students (30%), and non-technical individuals (10%)—highlighted the growing ease with which AI-powered tools can facilitate phishing attempts. Before the survey, participants were introduced to phishing concepts to establish a baseline understanding. They then rated their confidence on a scale of 1 to 10 in executing phishing attacks under four cases discussed above. The results revealed an increase in confidence as resources became more accessible: Case 1, participants scored an average of 2.3; Case 2 – 4.5; Case 3 – 7.3; and Case 4 – confidence peaked at 8.5. These findings highlight how AI can significantly lower the barrier for launching phishing attacks, even for non-technical individuals, emphasising the urgent need for proactive cybersecurity measures, increased awareness, and mitigation

strategies to counter AI-driven cyber threats. Reinforcing these survey insights, a controlled experiment with freshly joined undergraduate students from non-technical backgrounds demonstrated similar trends. The task involved launching a phishing campaign by creating a fully functional phishing website along with a phishing email that can redirect users to the phishing site to harvest user credentials. GenAI reduced the effort and time required by more than half (7 hours → 3 hours) and enabled successful task completion, whereas users who were allowed access to the Internet but not Generative AI tools struggled to complete the task despite extensive efforts. Together, the survey and experiment provide strong evidence of how AI assistance dramatically reduces the time, effort, and expertise required to conduct phishing attacks, thereby amplifying the associated cybersecurity risks.

Defensive Measures Against AI Jailbreaking

Policy-Level Approaches

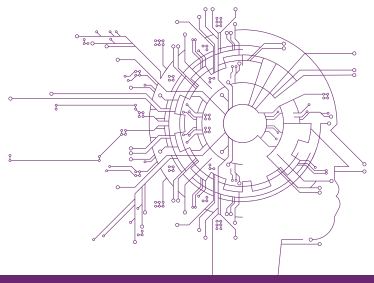
- **Stricter AI Regulations:** Governments should enforce stricter AI laws, preventing misuse through mandatory AI safety checks.
- **Global AI Ethics Committees:** Collaboration between nations to establish and enforce AI safety standards.
- **Content Moderation and AI Use Policies:** Companies should implement clear AI use guidelines, ensuring ethical compliance.

Technological Solutions

- **Adversarial Training:** Enhancing AI models to detect and prevent jailbreaking attempts through continuous updates.
- **Real-Time AI Behavior Monitoring:** Implementing monitoring tools to detect and flag unethical AI usage.
- **Red Teaming for AI Security:** Conducting regular security audits and simulated attacks to identify vulnerabilities.

Government and Industry Collaboration

- **National AI Security Policies:** Governments should work with AI research institutions to develop robust AI security protocols.
- **Corporate AI Security Frameworks:** Businesses should integrate AI security protocols in their cybersecurity infrastructure.
- **Public Awareness and Cybersecurity Training:** Educating individuals and employees about AI-generated threats.



10

Conclusions and Future Directions

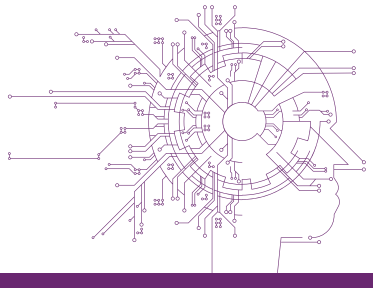
This study has demonstrated how AI jailbreaking presents a critical cybersecurity challenge, enabling attackers to manipulate GenAI models to generate harmful content, including phishing emails, malware scripts, and deepfake-based scams. The findings reveal that social engineering threats have become more sophisticated due to AI's ability to create highly personalized attack vectors. Additionally, the widespread adoption of large language models (LLMs) in cybercrime has significantly reduced the cost of executing phishing campaigns, making AI-driven attacks more accessible and scalable. Despite the implementation of content moderation and adversarial training, current AI security frameworks remain vulnerable to evolving jailbreak techniques. This underscores the pressing need for stronger AI regulations, robust monitoring systems, and enhanced cybersecurity defenses to prevent AI from being misused for malicious purposes. As GenAI continues to evolve, collaborative efforts between policymakers, AI researchers, and cybersecurity experts will be essential in mitigating risks and ensuring the responsible use of this powerful technology.

Future Work

While this study provides a comprehensive analysis of AI jailbreaking and its cybersecurity implications, several areas warrant further research:

1. **Enhanced Jailbreaking Detection Techniques –** Future studies should focus on developing advanced AI models capable of identifying and resisting jailbreak attempts in real time. This includes integrating adversarial machine learning techniques to reinforce AI security.

2. **AI-Powered Cybersecurity Defense Mechanisms –** Leveraging AI to counter AI-generated threats is a crucial area of exploration. Research should focus on real-time AI behavior monitoring, threat intelligence, and automated phishing detection to combat evolving cyber threats.

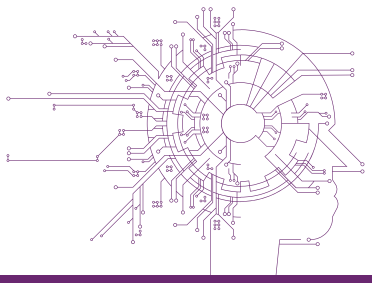


3. **Legal and Ethical AI Governance –** Stricter global regulations and policies are necessary to prevent the misuse of GenAI. Further research is required to establish comprehensive AI ethics frameworks, ensuring that AI remains aligned with human-centric values and societal security.

4. **Human-AI Collaboration in Cybersecurity –** Investigating how AI can assist cybersecurity professionals in threat detection and mitigation could lead to more proactive and adaptive defense mechanisms.

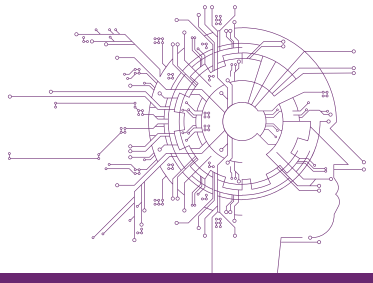
5. **Reducing Bias and Hallucinations in AI Models –** Jailbreaking often exploits AI biases and inconsistencies. Future research should focus on reducing biases, improving context awareness, and refining the decision-making processes of AI models.

Addressing these research gaps will play a pivotal role in strengthening AI security, minimizing risks, and ensuring the ethical and responsible development of GenAI technologies.



References:

1. Neelam Tyagi, Generative AI: Applications, Use Cases, and Examples, https://quantiphi.com/blog/generative-ai/, accessed on: 20/03/2025.
2. Gartner, Gartner Experts Answer the Top Generative AI Questions for Your Enterprise https://www.gartner.com/en/topics/generative-ai, accessed on: 21/03/2025.
3. Schmitt, Marc, and Ivan Flechais. "Digital deception: Generative artificial intelligence in social engineering and phishing." Artificial Intelligence Review 57, no. 12 (2024): 1-23, https://doi.org/10.1007/s10462-024-10973-2 .
4. Shibli, Ashfak Md, Mir Mehedi A. Pritom, and Maanak Gupta. "Abusegpt: Abuse of generative ai chatbots to create smishing campaigns." In 2024 12th International Symposium on Digital Forensics and Security (ISDFS), pp. 1-6. IEEE, 2024, https://doi.org/10.1109/ISDFS60797.2024.10527300 .
5. Taylor, Zachary, Akriti Sharma, and Kritagya Upadhyay. "Examining the Threat Landscape of Generative AI: Attack Vectors and Mitigation Strategies for LLMs." In 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), pp. 01014-01020. IEEE, 2025, https://doi.org/10.1109/CCWC62904.2025.10903828 .
6. Teichmann, Fabian. "Ransomware attacks in the context of generative artificial intelligence—an experimental study." International Cybersecurity Law Review 4.4 (2023): 399-414. https://doi.org/10.1365/s43439-023-00094-x.
7. ABIresearch, Generative AI Software Market Size by Region, Generative AI Software Market Size by Region., accessed on: 21/03/2025.
8. Öykü Isik , Ankita Goswami , Phishing Attacks Are Evolving. Here's How to Resist Them , https://hbr.org/2024/10/phishing-attacks-are-evolving-heres-how-to-resist-them, accessed on: 22/03/2025.
9. Jacob Fox, Top 40 AI Cybersecurity Statistics, https://www.cobalt.io/blog/top-40-ai-cybersecurity-statistics, accessed on: 23/03/2025.
10. Vention, AI adoption statistics by industries and countries: 2024 snapshot, https://ventionteams.com/solutions/ai/adoption-statistics, accessed on: 20/03/2025.
11. George Lawton, Generative AI ethics: 11 biggest concerns and risks, https://www.techtarget.com/searchenterpriseai/tip/Generative-AI-ethics-8-biggest-concerns, accessed on: 23/03/2025.
12. Somdip Dey, Which Ethical Implications Of Generative AI Should Companies Focus On?, https://www.forbes.com/councils/forbestechcouncil/2023/10/17/which-ethical-implications-of-generative-ai-should-companies-focus-on/, accessed on: 10/03/2025.
13. Al-Kfairy, Mousa, et al. "Ethical challenges and solutions of generative AI: An interdisciplinary perspective." Informatics. Vol. 11. No. 3. Multidisciplinary Digital Publishing Institute, 2024, https://doi.org/10.3390/informatics11030058.



14. University of Alberta, Ethical Considerations for Using Generative AI, https://guides.library.ualberta.ca/generative-ai/ethics, accessed on: 11/03/2025.

15. University of Saskatchewan, Generative Artificial Intelligence: Ethical Considerations, https://libguides.usask.ca/gen_ai/ethical, accessed on: 11/02/2025.

16. R. Mishra, G. Varshney and S. Singh, "Jailbreaking Generative AI: Empowering Novices to Conduct Phishing Attacks," *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S)*, Naples, Italy, 2025, pp. 251-252, doi: https://doi.org/10.1109/DSN-S65789.2025.00022 .

17. George Lawton, What is GenAI? Generative AI explained, https://www.techtarget.com/searchenterpriseai/definition/generative-AI, accessed on: 22/03/2025.

18. Roy, Sayak Saha, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. "From Chatbots to PhishBots?—Preventing Phishing scams created using ChatGPT, Google Bard and Claude." arXiv preprint arXiv:2310.19181 (2023), https://doi.org/10.48550/arXiv.2310.19181 .

19. Gupta, Maanak, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. "From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy." IEEE Access 11 (2023): 80218-80245, https://doi.org/10.1109/ACCESS.2023.3300381 .

20. Guembe, Blessing, Ambrose Azeta, Sanjay Misra, Victor Chukwudi Osamor, Luis Fernandez-Sanz, and Vera Pospelova. "The emerging threat of ai-driven cyber attacks: A review." Applied Artificial Intelligence 36, no. 1 (2022): 2037254, https://doi.org/10.1080/08839514.2022.2037254 .

21. Proofpoint, What Is Phishing?, https://www.proofpoint.com/us/threat-reference/phishing, accessed on: 20/03/2025.

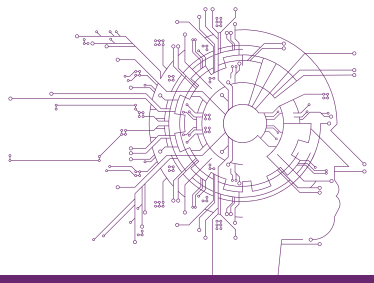
22. Alabdan, Rana. "Phishing attacks survey: Types, vectors, and technical approaches." Future internet 12, no. 10 (2020): 168, https://doi.org/10.3390/fi12100168 .

23. APWG, Phishing Activity Trends Report 3rd Quarter 2024, https://docs.apwg.org/reports/apwg_trends_report_q3_2024.pdf, accessed on: 21/03/2025.

24. Viper Security Group, VIPRE's Email Threat Trends Report: Q2 2024, https://vipre.com/resources/email-threats-latest-trends-q2-2024, accessed on: 20/03/2025.

25. Fred Heiding, Bruce Schneier and Arun Vishwanath, AI Will Increase the Quantity — and Quality — of Phishing Scams, https://hbr.org/2024/05/ai-will-increase-the-quantity-and-quality-of-phishing-scams, accessed on: 22/02/2025 .

26. Varshney, Gaurav, Rahul Kumawat, Vijay Varadharajan, Uday Tupakula, and Chandranshu Gupta. "Anti-phishing: A comprehensive perspective." Expert Systems with Applications 238 (2024): 122199, https://doi.org/10.1016/j.eswa.2023.122199.



National Centre of Excellence
CYBERSECURITY TECHNOLOGY AND ENTREPRENEURSHIP

The National Centre of Excellence (NCoE) for Cybersecurity Technology Development has been conceptualized by the Ministry of Electronics & Information Technology (MeitY), Government of India, in collaboration with the Data Security Council of India (DSCI). Its primary objective is to catalyze and accelerate cybersecurity technology development and entrepreneurship within the country. NCoE plays a crucial role in scaling and advancing the cybersecurity ecosystem, with a focus on critical and emerging areas of security.

Equipped with state-of-the-art facilities, including advanced lab infrastructure and test beds, NCoE enables research, technology development, and solution validation for adoption across government and industrial sectors. By adopting a concerted strategy, NCoE aims to translate innovations and research into market-ready, deployable solutions—contributing to the evolution of an integrated technology stack comprising cutting-edge, homegrown security products and solutions.

DSCI
PROMOTING DATA PROTECTION
A nasscom Initiative

Data Security Council of India (DSCI) is a premier industry body on data protection in India, setup by nasscom, committed to making the cyberspace safe, secure and trusted by establishing best practices, standards and initiatives in cybersecurity and privacy. DSCI brings together governments and their agencies, industry sectors including ITBPM, BFSI, telecom, industry associations, data protection authorities and think-tanks for policy advocacy, thought leadership, capacity building and outreach initiatives. For more info, please visit www.dsci.in

DATA SECURITY COUNCIL OF INDIA

+91-120-4990253 | ncoe@dsci.in

https://www.n-coe.in/

4 Floor, NASSCOM Campus, Plot No. 7-10, Sector 126, Noida, UP -201303

Follow us on

@CoeNational
nationalcoe
nationalcoe
NationalCoE