Generative artificial intelligence (AI)

CANADIAN CENTRE FOR CYBER SECURITY
July 2023 | ITSAP.00.041

Many organizations use artificial intelligence (AI) for process optimization, data analysis, patient diagnosis and treatment, and customization of their user experience.

Generative AI is a type of artificial intelligence that generates new content by modelling features of data from large datasets that were fed into the model. While traditional AI systems can recognize patterns or classify existing content, generative AI can create new content in many forms, including text, image, audio, or software code.

One class of generative AIs that have seen significant improvement in recent years are large language models (LLMs). To create content the LLM is provided with a set of parameters via a query or prompt. When generative AI tools interact with users in a conversational mode via prompts, it's easier for the user to generate content. Since late 2022, several LLMs (OpenAI's ChatGPT and Google's LaMDA) and services using LLMs (Google's Bard and Microsoft's Bing) have gained the world's attention. As the interest in generative AI increases, its possible uses are being explored by many. This publication provides some information on the potential risks and mitigation measures associated with generative AI.

How is generative AI being used?

Generative AI is both a transformative and disruptive technology that may significantly alter the way consumers, industries, or businesses operate. It has the potential to enable creativity and innovation that could improve services and business operations. Some of the useful applications of generative AI are in the following areas:

Healthcare: Assists healthcare providers to make faster diagnoses. AI holds the promise to make personalized treatment plans commonplace. It can be leveraged to enable therapeutic targets and novel drug candidates.

Business: Creates personalized customer communications for existing and prospective clients as well as generates predictive sales modelling to forecast customer behaviour.

Education: Allows educators to create personalized learning plans for students tailored to their individual performance, needs, and interests. This could help teachers better support their students.

Software development: Enables software developers to generate code, assist in debugging, or offer code snippets. This can potentially help speed up the development and release of software products.

Publishing and media: Enables content creators to produce unique outputs for use in marketing campaigns, advertising, television, and video productions. On demand content can be generated quickly and with fewer resources, leading to significant cost reduction.

Cyber security: Facilitates enhancement of cyber defence tools against ransomware and other attacks. Assists cyber security practitioners to more easily scan large datasets to identify potential threats and minimize false positives by filtering out non-malicious activities.

Online marketplace: Generates human-like responses in a chatbot and conversational agents which can help organizations improve customer service and reduce support costs.

Check out our publication on Artificial intelligence (ITSAP.00.040)

Communications Security Establishment | Centre de la sécurité des télécommunications

Canada

Generative artificial intelligence (AI)

CANADIAN CENTRE FOR CYBER SECURITY
July 2023 | ITSAP.00.041

What are the risks of generative AI?

While the capabilities of generative AI technology present great opportunities, they also bring many concerns. Generative AI can enable threat actors to develop malicious exploits and potentially conduct more effective cyber attacks. A huge concern is that it can provide threat actors with great powers to influence. For example, deliberate manipulation of the underlying code and the tools using it, can introduce supply chain risks from insider threat at the design level to the distribution and patching of software. Here are some of the potential risks to be aware of:

- **Misinformation and disinformation**: Content may not clearly be identified as being AI generated, and could potentially lead to confusion (misinformation) or deception (disinformation). Threat actors can use this in scams or fraudulent campaigns against individuals and organizations.
- **Phishing**: Threat actors can craft targeted spear phishing attacks more frequently, automatically, and with a higher-level of sophistication. Highly realistic phishing emails or scam messages could lead to identify theft, financial fraud, or other forms of cybercrime.
- **Privacy of data**: Users may unknowingly provide sensitive corporate data or personally identifiable information (PII) in their queries and prompts. Threat actors could harvest this sensitive information to impersonate individuals or spread false information.
- **Malicious code**: Technically skilled threat actors can overcome restrictions within the generative AI tools to create malware for use in a targeted cyber attack. Those with little or no coding experience can use generative AI to easily write functional malware that could cause a nuisance to a business or organization.
- **Buggy code**: Software developers may deliberately or inadvertently introduce unsecured and buggy code to the development pipeline. This could happen if they omit or improperly implement error handling and security checks.
- **Poisoned datasets**: Threat actors can inject malicious code into the dataset used to train the generative AI system. This could undermine the accuracy and quality of the generated data. It could also increase the potential for large-scale supply-chain attacks.
- **Biased content**: A majority of the training dataset fed into the LLMs come from the open Internet. As such, generated content has a fundamental bias in that only limited amounts of the world's total data are online and available for AI to use. Also, generated content may be prejudiced if the training dataset lacks balanced representation of data points.
- **Loss of intellectual property (IP)**: Generative AI tools may enable sophisticated threat actors to more easily steal corporate data faster and in bulk. Loss of IP (e.g. proprietary business information, copyrighted data, software code or drug trial data) can devastate your organization's reputation, revenue, and future growth.

Be aware

Generative AI is a technology that is in the realm of machine learning rather than true "intelligence". It doesn't actually understand concepts but produces content that is statistically the best response to a prompt or query.

Always keep in mind that its outputs can be:

- wrong
- illogical
- unaware
- biased

Communications Security Establishment | Centre de la sécurité des télécommunications

Canada

Generative artificial intelligence (AI)

CANADIAN CENTRE FOR CYBER SECURITY
July 2023 | ITSAP.00.041

How to mitigate the risks?

Generative AI is another tool that threat actors can leverage to launch their cyber attacks. As this technology becomes more widely used and exploited, there will likely be increases in sophisticated cyber attacks including phishing and social engineering, misinformation/disinformation, and identity theft. While it may be difficult to identify (or positively attribute) cyber attacks that leverage generative AI, organizations and individuals can prepare for the increased challenges that these attacks may bring.

Organizations can take the following actions to minimize their risks of compromise to cyber attacks:

- ☐ **Implement strong authentication mechanisms**— Secure accounts and devices on your networks with multi-factor authentication (MFA) to prevent unauthorized access to your high-value resources and sensitive data. For more information, see Secure your accounts and devices with multi-factor authentication (ITSAP.30.030) and Steps for effectively deploying multi-factor authentication (ITSAP.00.105).
- ☐ **Apply security patches and updates** — Enable automatic updates of IT equipment and patch known exploited vulnerabilities as soon as possible. This will help to prevent AI generated malware from infecting the network.
- ☐ **Stay informed** — Keep up to date on the latest threats and vulnerabilities associated with generative AI and take proactive steps to address them.
- ☐ **Protect your network** — Use network detection tools to monitor and scan the network for abnormal activities. This enables you to quickly identify incidents and threats in order to deploy appropriate mitigation measures. Additionally, explore how AI might be deployed defensively in network protection tools and consider any ramifications. For more information, see Network Security Logging & Monitoring (ITSAP.00.085) and Top10 IT security action items - No. 5 Segment and separate information (ITSM.10.092).
- ☐ **Train your employees** — Educate all users on how to identify the warning signs of social engineering attacks and who to contact to manage these situations securely. This should include an easy way for users to report phishing attacks or suspicious communications.

Individuals can take the following actions to protect their personal data from phishing attacks:

- ☐ **Verify content** — As more data becomes available, it may not be easy to tell who is responsible for the content or how much of it is logical or factual. It's important to read and look for indication that the content was produced by a generative AI tool. Review the generated content and take the time to fact check against credible sources. For more information, see How to identify misinformation, disinformation, and malinformation (ITSAP).00.300).
- ☐ **Practice basic cyber security hygiene** — Stay informed, use strong passwords, and enable two-factor authentication to protect online accounts. Make sure to keep software up to date, use antivirus software, and avoid public Wi-Fi networks.
- ☐ **Limit exposure to social engineering or business email compromise**— Implement basic online safety practices such as:
  - reducing the amount of personal information posted online
  - avoiding opening email attachments and clicking on links from unknown sources
  - communicating via an alternate, verified channel
  - being suspicious of callers that want sensitive information.

For more information, see Don't take the bait; recognize and avoid phishing attacks (ITSAP.00.101) and What is voice phishing (vishing)? (ITSAP.00.102).

Security protections when using generative AI tools

The following security measures can help you generate quality and trusted content while mitigating privacy concerns:

- ☐ **Establish generative AI usage policies** — The policies should include the types of content that can be generated and how to use the technology to avoid compromises to your sensitive data. Your policies should also include the oversight and review processes required to ensure the technology is used appropriately. When creating solutions using generative AI, ensure practices lead to trustworthy and ethical behaviour. Be sure to implement the policies quickly and ensure they are communicated to staff.
- ☐ **Select training datasets carefully** — Obtain datasets from a trusted source and implement a robust process for validating and verifying the datasets, whether they're externally acquired or developed internally. Use diverse and representative data to avoid inaccurate and biased content. Establish a process for outputs to be reviewed by a diverse team from across your organization to look for inherent biases within the system. Continuously fine-tune or retrain the AI system with appropriate external feedback to improve quality of outputs.
- ☐ **Choose tools from security-focused vendors** — Ensure your vendors have robust security practices baked into their data collection, storage, and transfer processes.
- ☐ **Be careful what information you provide** — Avoid providing PII or sensitive corporate data as part of the queries or prompts. Determine whether the tool allows your users to delete their search prompt history.

Communications Security Establishment | Centre de la sécurité des télécommunications
Canada