NCSC
National Cyber Security Centre

29082301000-NCSC

Department of the Environment, Climate & Communications

Cyber Security Guidance on Generative AI for Public Sector Bodies

01 June 2023

Status: TLP-CLEAR

Recipients can spread this to the world, there is no limit on disclosure. Sources may use TLP-CLEAR *when information carries minimal or no foreseeable risk of misuse, in accordance with applicable rules and procedures for public release. Subject to standard copyright rules,* TLP-CLEAR *information may be shared without restriction.*

For more information on the Traffic Light Protocol, see https://www.first.org/tlp/. Please treat this document in accordance with the TLP assigned.

Issue

In the past number of months there have been rapid developments and advancements in the field of Artificial Intelligence, and in particular in the field of Generative AI (Chatbots, Large-Language Models, Text/Image/Video/Voice generation etc.) As the capabilities of Generative AI (GenAI) continue to advance and with concerns around its increasing availability and use, Departments have sought guidance on the cyber security risks associated with the use of GenAI.

Recommendation

All new technology, such as GenAI should only be adopted based on a clearly defined business need following an appropriate risk assessment.

Each department will likely have a different view in terms of the business use case of GenAI tools and platforms, as well as the risk appetite for such use.

The NCSC recommends that access is restricted by default to GenAI tools and platforms and allowed only as an exception based on an appropriate approved business case and needs. It is also recommended that its use by any staff should not be permitted until such time as Departments have conducted the relevant risk assessments, have appropriate usage policies in place and staff awareness on safe usage has been implemented.

By developing an appropriate strategy now to understand the risks and benefits associated with GenAI, Departments can better protect themselves from risks while maximising the possible benefits with these platforms. In conducting a risk assessment and determination on the use of GenAI within a Department, it is recommended that among other factors the following are examined, (i) Business Need, (ii) Permissions, (iii) Data & Privacy, (iv) Cybersecurity and (v) Non-technical factors.

The recently published Mobile Device Management Guide provides guidance on the vetting of third-party applications and software, and can be used as a reference when conducting a risk assessment in respect of GenAI.

It should be noted that there are wider factors, beyond cyber security, that need to be considered such as human oversight, ethics, non-discrimination, transparency etc. Guidance on the trustworthy and ethical use of AI in the public service is being developed by the Department of Public Expenditure, NDP Delivery and Reform with the Department of Enterprise, Trade and Employment and will cover these topics in greater detail.

The NCSC has suggested a concise **"Do & Don't"** list for staff using GenAI, which can be found at the end of this document.

Discussion

National AI Strategy

Government, through DETE, have set out policy associated with Artificial Intelligence through the National AI Strategy: AI - Here for Good.

The strategy outlines how technologies and services such as Artificial Intelligence, including Machine Learning, open up vast opportunities for potential improvements in public services. In particular, Strand 4 of the Strategy's objective is better public service outcomes through a step change in AI adoption by the Irish public sector.

Recent Developments

In recent months, many publicly accessible Generative AI products, platforms and frameworks have become available, making AI increasingly accessible. In addition to the wider availability of tools, the integration of AI into various search engines increases the likelihood of user interaction with AI , all of which runs the risk that public servants could use these tools without having the appropriate guidance or safeguards in place. Whilst most of the concerns relate to data protection, copyright infringement, bias, ethics, and the future of work – there are also **cyber security** considerations. This guidance focuses on the cyber security risks associated with the use of Generative AI, as well as some of the potential misuses of the technology by malicious actors.

Risks associated with use of GenAI

Data in

Most of the current GenAI models available for use are public cloud-based systems. The information you put into the model through prompts **will** be visible to the company that owns the model. The queries that are put into the model will almost certainly be used to train future iterations of the model. Therefore, it is not unreasonable to assume that information you put into the model will at some point surface at the future when others are querying the technology.

It is therefore imperative that data that your organisation does not want in the public domain is never entered in to a public GenAI model. This includes classified information, personal data, commercially sensitive data, private Government business etc.

In the context of cyber security, any information around the network topology, software source code, assets lists or details of deployed hardware or software etc. would be information which should never be inputted to a GenAI model.

It is also worth noting that the query itself could be a something which brings your organisation in to disrepute. Staff should be given clear instructions on what GenAI can be used for, and what it cannot – e.g., GenAI should not be used to respond to correspondence such as PQs or queries from the public.

At present most GenAI models provide very little control over the data that goes into the model. In the future there may be private GenAI models – either self-hosted or cloud-hosted - that may provide a better alternative than providing information directly to the company providing the model.

Data out

The information that comes out of a GenAI model is only as good as the information that it has been trained on. Often the data sets used for training the model will include false, bias, misleading and potentially harmful information.

GenAI models have been demonstrated to at times produce responses which are inaccurate, misleading, false, or sometimes harmful. The phenomenon of GenAI models producing this type of false information is known as *'hallucination'*. Often this information is presented articulately and confidently – making it difficult to discern fact from fiction. It is therefore important that any staff utilising GenAI are aware of its limitations and that all information is thoroughly fact checked before relying on it in any way.

While GenAI has been lauded for its ability to accelerate the process of writing computer code, some research indicates that it will often produce insecure code, and therefore should only be used by competent and skilled developers for this purpose.

Technical Security

The high-profile nature of these tools and the extensive usage of them makes them an attractive target for cyber criminals. As vast quantities of data continue to be inputted into these platforms, it is inevitable that this will include confidential, personal, and private data types, which may be targeted by cyber criminals.

The concern is that queries submitted, and the subsequent responses are stored and remain associated with a user's account. In the event of a breach of these systems, this data could be used by cyber criminals. The ability of GenAI to process vast quantities of data and determine patterns from it could assist cyber criminals in processing all acquired data and generating target profiles for the purposes of social engineering, form the basis for cyber extortion or conduct network intrusion attacks from information obtained that could be used in credentialed based attacks.

Therefore, the cyber security controls that Public Service Bodies (PSBs) place on access to GenAI models are important. General cybersecurity practices as outlined in the Cyber Security Baseline Standards, such as strong passwords, multi-factor authentication, logging and monitoring, vulnerability management and patching etc. are all very relevant when it comes to protecting the data inputted, generated or stored by these models.

Equally the security of providers of GenAI models should be subject to strict scrutiny on their own security, data, and privacy policies and procedures. The provider should be able to demonstrate that they have implemented state of the art security, ideally through appropriate international cyber security certifications. The NCSC Guidelines on Cyber Security Specifications (ICT Procurement Criteria for Public

Service Bodies) report provides detailed advice on the cyber security requirements through all phases of procurement and is relevant for the procurement of GenAI services.

Misuse by malicious actors

As with all technology, GenAI is open to misuse and abuse and could provide a powerful tool for cyber criminals to leverage. Its potential use in crafting specifically targeted and convincing phishing emails or text and voice messages could increase the likelihood of unsuspecting individuals being deceived into providing information such as login credentials or banking details. The ability to then automate large scale targeted campaigns could further increase the number of victims by using AI chatbots to interact with individuals believing they are in contact with a real person and are convinced into falling for the intended scam. The potential for impersonation or the creation of deepfake media could be used by bad actors to generate media to discredit public figures, spread a false narrative, and blackmail or defraud individuals. Personal data could be processed to establish a more efficient strategy in guessing an individual's passwords to compromise accounts. The use of GenAI for generating application code solutions has already been highlighted however, this could equally be abused to develop more sophisticated malware that is harder to detect and respond to by current systems.

Hence, cyber security awareness training should remind staff of the potential misuses of the technology by malicious actors. The NCSC is currently drafting a short guidance document for the general public relating to the risks associated with AI, which PSBs can leverage to inform staff.

GenAI as a cyber security enabler

GenAI is likely to be a very powerful tool for those responsible for the security of network and information systems. The ability of these tools to process large quantities of data make them an ideal solution in analysing log files and monitoring network traffic to identify anomalous traffic patterns and potential threats. Routine time intensive tasks such as the analysis of the varying forms of threat intelligence data, vulnerability reports, malware samples and product security advisories could be performed by GenAI for the purposes of developing proactive cyber defence strategies. The automated generation of reports from large data processing could assist cyber security staff in identifying and prioritising responses. Similarly, to the potential use of GenAI by cyber criminals in the formulation of more complex and believable phishing campaigns, the same technology could be used in training employees through AI generated phishing simulations. The benefits of GenAI in cyber security will likely continue to develop with tools likely available to determine vulnerabilities and adapt to increasingly more complex malware attacks. However, PSBs considering adopting these tools to defend their network should do so carefully, following a risk assessment, and ensure staff who have access use them safely.

Staff Advice

If the decision is taken to allow staff to have access to GenAI tools it is essential that its use is covered by Security, Data and Privacy policies that outline under what circumstances, its use is permitted and how the technology can be used within the organisation.

Some suggested DOs and DONTs from the NCSC

DO

- Do validate all generated output for accuracy, copyright infringement and bias.
- Do switch off chat history or regularly delete interactions to limit potential data breach exposure.
- Do ensure you have selected a legitimate site or an official mobile device app. There are many unofficial sites and apps available that could be malicious.
- Do understand the limitations in responses due to incomplete or insufficient data available to the platform.
- Do thoroughly validate all computer code outputs for bugs and security issues.
- Do treat your account security as a priority: Use a strong unique password and enable multi-factor authentication.

DON'T

- Do not use public versions of GenAI services for business purposes.[1]
- Do not create accounts with corporate email addresses, unless you are using an enterprise version for which you have an approved business case.
- Do not rely on GenAI to directly create, design or draft Government policy.
- Do not use GenAI to generate responses to representations made to Ministers.
- Do not enter any sensitive information such as personal data, business data, propriety information (like software source code) or any government information.
- Do not enter data that you would not normally want to be made publicly available.

[1] Departments wishing to use GenAI for business purposes should use enterprise versions which allow controlled access and safer adoption following detailed risk assessments.

Conclusion

With the increasing advances in Generative AI there is the potential for both risks and opportunities. As with the introduction of all new technologies and software into an organisation, appropriate precautionary measures should be taken to implement suitable safeguards to mitigate risks. The benefits of Generative AI should not be overlooked and could be positively embraced, subject to the necessary safeguards. There are clear benefits to productivity and effectiveness, and GenAI will likely be built into many future product offerings. The area of GenAI will continue to rapidly develop, and it will be necessary to continuously monitor and evaluate these developments to better understand changes to the risks and benefits of these tools.

DISCLAIMER: *This document is provided "as is" without warranty of any kind, expressed or implied, including, but not limited to, the implied warranty of fitness for a particular purpose. NCSC-IE does not endorse any commercial product or service, referenced in this document or otherwise.*

National Cyber Security Centre
29-31 Adelaide Road,
Dublin, D02 X285,
Ireland
Tel: +353 (0)1 6782333
Mail: certreport@decc.gov.ie
Web: ncsc.gov.ie
Twitter: ncsc_gov_ie
LinkedIn: ncsc-ie