NORTHWAVE
CYBER SECURITY

White Paper: Policies And Controls When Using AI Applications

Date: July 5, 2023
Author: Rob Berends

Introduction

The goal of this document is to propose practical examples of policies and controls for preventing or mitigating security-related risks when using machine learning (ML) and artificial intelligence (AI) tools, such as ChatGPT, within your company. First, we will give a brief introduction on what these AI applications can do.

Then, we list and discuss the risks related to the usage of recently released AI applications. For each of these risks, we provide practical examples of policies and controls that can serve as a reference for your company.

Finally, because it is very likely that AI applications are already being used within your organisation, we will let you know what you can do starting today.

What Are AI Applications Used For?

Generative AI Technology is a type of 'artificial intelligence' that can generate content such as images, text, and music without human intervention. These tools use deep learning and machine learning algorithms to learn from existing data, get feedback from users, and generate content based on what they have 'learned'. ChatGPT was not the first conversational AI of its kind; many have come before it. But ChatGPT was publicly available, free, and more advanced than anything before it, which is why it has become the fastest growing platform ever, growing faster than Instagram or TikTok1. Due to its popularity, it has opened doors for many other AI applications that are now being used in various fields such as art, music, and literature to create completely new content on a scale that we have not seen before. The overview below lists some examples of the most used generative AI applications:

- **ChatGPT, Bing Chat, Google Bard, Baidu E:** Large language models(LLMs) that can generate human-like responses to various prompts and questions.
- **Midjourney, DALL-E 2, Stable Diffusion:** Image generation tools that can generate images from textual descriptions.
- **Amper Music, AIVA, Jukedeck: M**usic composition tools that can generate original music tracks in various genres.

Besides these standalone applications, we observed that software companies add generative AI support to their applications:

- **Microsoft Copilot:** Will be added to all Microsoft products and will help users increase their productivity while composing documents, spreadsheets, presentations, or even autogenerating responses in Outlook or minutes from your Teams Meetings.[2]
- **GitHub Copilot:** Already available for developers to function as an AI pair programmer that aids in their development and programming and can suggest functions or parts of code.[3]
- **Zendesk AI:** Customer Service platforms like Zendesk are all rushing to implement LLMs to automate their processes and decrease the human costs of running these platforms.[4]

Considering these use cases, the most used types of AI applications that are currently being used within organisations are LLMs like ChatGPT, which help people and companies to generate emails, marketing texts, websites, code, and white papers like this one. These technologies have the potential to revolutionise many industries by automating tasks, processes, and operations.

This document aims to provide an overview of the risks these technologies pose and provide an overview of the policies and controls companies can and should take to ensure the secure usage of AI applications within their company. These can be implemented to structure governance, give guidance to employees, and provide guardrails for secure usage.

1 https://in.mashable.com/tech/46656/chatgpt-becomes-fastest-growing-platform-in-the-world-as-it-hits-100-million-users-in-2-months
2 https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/
3 https://github.com/features/copilot
4 https://www.zendesk.com/service/ai/

Risks Of Using AI

As with all innovative technology, it is likely that these AI applications will introduce new types of risks. To ensure companies take the right measures, it is of the essence to understand what these risks could be. And while some of the risks may still be unknown, the following risks could impact your company directly:

Data privacy risks: This happens when individuals or companies feed confidential or personal data into the applications, which could also include sensitive information such as medical records, financial information, or intellectual property. Because all the data is used to train and improve these algorithms, feeding personal data into an AI application on the internet must always be considered a data breach.

Data integrity risks: ChatGPT has no knowledge of what is correct or not; it performs next-token predictions on text. This means that it sometimes generates complete nonsense which is easy to detect. However, it can also produce faulty information that is much more difficult to distinguish from correct information. This is called 'hallucinating', and it can cause serious problems if the output is not checked and verified properly.

Malicious input risk: This risk involves employees unintentionally or maliciously feeding the ML/AI tool with inappropriate, biased, offensive, or illegal content and using the output for company purposes; for example, an employee could input a prompt into the application that attempts to generate hateful or discriminatory responses, which could then lead to reputational damage or legal issues for the company.

Plagiarism and copyright infringement risks: Generative AI can be used to create content that is similar or identical to existing content, which can lead to issues with plagiarism or copyright infringement. As of this writing, (inter)national laws are not clear on this topic, and the first lawsuits are just being filed.[5]

User authentication risks: If unauthorised individuals gain access to AI applications by bypassing or impersonating an employee, they could get access to data or misuse it for malicious activities. This can happen when an employee's login credentials for the AI application are compromised or if the authentication process is weak.

Compliance and legal risks: If there are legal and regulatory requirements for AI applications. it is important to comply to these. For example, if the application processes and stores personally identifiable information (PII), such as customer data, without adhering to data protection laws like the General Data Protection Regulation (GDPR), the company could face legal consequences and financial penalties. Additionally, some countries might ban the usage of AI applications as was, for example, temporarily the case in Italy.[6]

Third-party risks: Most AI applications are provided through a Software-as-a-Service (SaaS) solution.

These third-party providers are thus responsible for the development and/or hosting of the AI applications. Before deploying or using a third-party application, it is important to thoroughly evaluate their infrastructure, security controls, and practices as these could contain vulnerabilities or privacy concerns.

Monitoring and auditing risks: Monitoring and auditing the activities of the usage of AI applications can be critical to the ability to detect and respond to security incidents, policy violations, or unusual behaviours that would potentially allow malicious activities to go unnoticed.

5 https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning
6 https://www.bbc.com/news/technology-65139406

External Threats Facilitated By AI

The following threats are applicable to AI applications as well but cannot be fully prevented from a company perspective. In addition, these are intentionally malicious and therefore not the focus of this document. It is important, though, to be aware that these could be exploited by threat actors:

- **Deepfakes:** Deepfakes are videos or images that have been manipulated using AI applications to make them appear real. These can be used to create fake news or to impersonate someone by manipulating the video or audio of a phone call.[7]
- **Phishing emails and social engineering:** AI applications can be used to create convincing, highly tailored phishing emails or social media posts that trick people into giving away or granting access to sensitive information.
- **Malware and ransomware code:** AI applications can be used to create new and complex types of malware and ransomware that can avoid conventional protection measures.[8]
- **Disinformation and propaganda:** AI applications can be used to create disinformation and propaganda that can be spread through social media and other channels.[9]

7 https://consumer.ftc.gov/consumer-alerts/2023/03/scammers-use-ai-enhance-their-family-emergency-schemes
8 https://www.wired.com/story/chatgpt-ai-bots-spread-malware
9 https://www.forbes.com/sites/petersuciu/2023/06/09/the-next-threat-from-generative-ai-disinformation-campaigns

Policies & Controls

Proper controls should be implemented to manage the eight risks potentially associated with the internal use of the AI applications mentioned above. The policies and controls proposed below are focused on the risk of the usage of AI applications within your organisation, not on the use of these tools by threat actors. At Northwave, we always categorise the controls in three domains:

- **Business:** These are organisational controls and include policies, responsibilities, procedures, and (contractual) agreements.
- **Bytes:** These are technical controls supported by hardware or software, such as authentication & authorisation, logging & monitoring, and configurations.
- **Behaviour:** These are people controls that ensure the safe behaviour of employees when using information systems, including training courses, exercises, and communication.

To determine which of the policies and controls are necessary and most suitable for your application, a risk assessment should be conducted, during which it is also important to identify any existing controls that can be used or updated.

Data Privacy Risks

Policy

All employees must classify and handle data appropriately based on its sensitivity. Sharing sensitive or confidential information with the AI application is prohibited. Sensitive data should be properly labelled and protected according to the company's data classification guidelines. Failure to comply with this policy may result in disciplinary action.

Controls

Business

- **Privacy policy and terms of use:** Before using AI applications, it is important to read the privacy policy and terms of use. These applications often use the input for the further improvement and learning of the underlying model. Therefore, the sharing of PII must be considered a data breach.

Bytes

- **Data loss prevention:** Implement data loss prevention solutions that can monitor and block the transmission of sensitive data to the ML/AI application. This can help prevent accidental or unauthorised data exposure.

Behaviour

- **Communicate clear guidelines:** Develop and communicate clear guidelines on what data may and may not be used with these tools. It is considered good practice to only use data that can be considered public.

Data Integrity Risks

Policy

All output that has been generated must be checked and verified for accuracy, truth, and legal implications by a natural person. If used for decision-making purposes, always document the prompt and output and consider it advice. Always have the final decision made by a natural person. Inaccuracies or errors that could have a significant impact should be reported.

Controls

Business

- **Human oversight and fact checking:** While most AI applications can operate autonomously, human oversight of the outcome and conclusions based on that outcome are essential for ensuring the accuracy and quality of the output. Companies should ensure that personnel are available to review and correct any errors made by the AI system.

Behaviour

- **User training:** Train users on the potential negative effects of these applications. Explain and show examples that these tools do not 'copy' from sources, but rather generate text based on what they have been taught, which can result in very truthful sounding answers that are nevertheless completely wrong. It is important that they do not blindly trust a response and know how to make corrections if necessary.

Malicious Input Risks

Policy

The AI application should only be used for legitimate business purposes. Employees must not input offensive, biased, or illegal content into the application. Inappropriate use may result in disciplinary action. Employees should undergo training on acceptable use guidelines and the potential risks associated with generating harmful or inappropriate content.

Controls

Business

- **Acceptable use policy:** The policy above should be added to the acceptable use policy of the company. It can be expanded with what is allowed as well, especially if certain usage is prohibited by laws or regulations.

Bytes

- **Accountability mechanisms:** Establishing mechanisms to hold users accountable for their actions with AI applications can promote responsible use and deter misuse. However, most tools do not support extensive logging or monitoring yet, so if accountability is required, create documenting guidelines for your employees, e.g. registering in a separate document the prompts used, together with their outputs. This should always be implemented when generative AI is being used in decision-making processes.

Behaviour

- **Communication guidelines:** Train users in the potential impact of negative output being used. Give employees guidelines that help them verify whether the output adheres to the company's values and communication guidelines.

Plagiarism And Copyright Infringement Risks

Policy

Employees must not use AI applications to generate content that infringes upon copyrights or discloses proprietary information. All content generated should be reviewed for potential intellectual property violations before use or distribution. The company respects intellectual property rights and expects employees to do the same. Violations of this policy may lead to legal action and disciplinary measures.

Controls

Business

- **Human oversight and reviewing:** Similar to the data integrity risks, it is important to also verify that no plagiarism or copyright infringements are being output by the AI application. Especially in creative businesses, this should be a mandatory check for all content used for commercial purposes.

Behaviour

- **Communication guidelines:** Explain to users the potential impact of intellectual property being used. Give employees guidelines and procedures that help them verify whether the output is original or borrows from content that might result in legal implications for the company.

User Authentication Risks

Policy

All data used by the ML/AI application must be securely stored, encrypted, and protected against unauthorised access. Access controls, including role-based access and encryption mechanisms, should be implemented to ensure data confidentiality and integrity. Regular reviews of access permissions and security assessments of the data storage infrastructure are required to maintain compliance and mitigate risks.

Controls

Business

- **Roles and reviews:** Be clear on who within the company should be allowed to use these tools. If limited to a specific group of users, document and review their permissions periodically.

Bytes

- **Access controls:** Limiting who has access to the system can help prevent unauthorised use and misuse of the technology. Access controls can include password protection, role-based access, two-factor authentication, and other security measures. Since most tools are publicly available, strict access controls can be hard to implement. We advise always using or providing corporate accounts for accessing these applications. When disallowing usage is preferred, blocking the domains on the company network could be considered, but keep in mind that these tools can always be accessed from private devices as well.

Compliance and Legal Risks

Policy

The company is committed to protecting personal data and complying with applicable laws and regulations, including the GDPR, and industry-specific standards. All employees are responsible for handling personal data in accordance with the established data protection and privacy guidelines. Failure to comply with these guidelines may result in legal consequences and disciplinary action.

Controls

Business

- **Privacy policy and terms of use:** Before using AI applications, it is important to read the privacy policy and terms of use. These applications often use the input for further improvement and learning of the underlying model. Therefore, the sharing of PII must be considered a data breach.

- **Regulation checks and audits:** Because these tools are fairly recent, an increase in the regulatory requirements for the use of these technologies is expected soon. New legislation is still being drafted; therefore, it is important to verify whether new regulations have taken effect and are applicable to your organisation. Depending on the industry and location, there may be specific regulations and laws that apply to the use of AI applications.

Behaviour

- **Clear communication:** It is important to communicate clearly with employees, customers, and other stakeholders about the use of AI technology. This includes being transparent about how and when AI technology is being used, what data is being collected, and how the collected data is used. For example, add a footnote or disclaimer stating that AI technology was used when the output is used in communications or documents.

Third-party Risks

Policy

Before engaging with any third-party vendor, a thorough assessment of their security practices must be conducted. Vendors must adhere to specified security controls and standards as defined in the vendor agreements. Failure to comply with this policy may result in termination of vendor contracts.

Controls

Business

- **Due diligence:** Before using AI applications, it is important to do proper due diligence on the policies provided. Because most of these tools are SaaS solutions and freely available, it is important to check the privacy policy and terms of use.

Bytes

- **Security and vulnerability testing:** Test the systems for known vulnerabilities and verify whether they comply with your organisation's security standards.
- **Secure configurations:** When AI technologies are available in enterprise solutions (e.g. Microsoft Copilot), it is critical that the configuration is in line with the policies and security requirements of the organisation. We advise configuring these tools according to the best practices as supplied by the third-party vendor.

Monitoring and Auditing Risks

Policy

The company implements monitoring and auditing mechanisms to track user interactions and system activities related to the ML/AI application. This includes logging and reviewing user sessions, generated responses, and system events. Monitoring and auditing are essential for identifying security incidents, policy violations, or unusual behaviours. Employees should be aware that their activities within the application may be logged and audited for security and compliance purposes.

Controls

Business

- **Periodic auditing and review:** Conducting regular audits and reviews of the activities within the AI application is crucial to identifying potential security incidents or policy violations. These audits should also cover configuration settings and technical implementations such as external connections. Also conduct regular compliance assessments and audits to ensure adherence to legal and regulatory requirements.

- **Acceptable use:** Incorporate monitoring and auditing as part of the acceptable use policy, emphasising the importance of compliance and accountability.

Bytes

- **Monitoring:** Detecting possible misuse of generative AI technology is critical, whether this is use outside the allowed usage policy or the (accidental) sharing of PII or confidential data. Even though technical measures on publicly available solutions are limited, we strongly advise granting employees access through their corporate account to ensure some form of control and monitoring. We do expect that AI tools integrated in enterprise solutions will give the options for audit logs, usage monitoring, and reporting.

- **Secure configurations:** When AI technologies are available in enterprise solutions (e.g. Microsoft Copilot) it is critical that the configuration is in line with the policies and security requirements of the organisation. We advise configuring these tools according to the best practices as supplied by the third-party vendor.

Concluding Remarks

Development of these AI technologies has been going on for several years, and it is important to note that applications that use machine learning and deep learning algorithms have been around for a long time. Applications such as Google Translate, Grammarly, Facebook, TikTok, and Netflix all use these algorithms to optimise their service and user experience. Therefore, the same policies could or should apply to the usage of these applications if they are used in a corporate environment.

But these generative AI solutions, which are becoming more popular by the day, can still be considered in the early adopter stage of the technology adoption life cycle. It has already been shown that they present opportunities on a scale we could not have imagined even a few years ago. The possibilities for creating unique content, such as text, images, videos, or music, are unprecedented. The solutions will bring countless opportunities for people and companies to automate processes, streamline operations, and create new content.

AI applications are here to stay, with novel solutions and ideas being developed every day. We will be monitoring this subject and the technologies being developed around it closely to be able to identify new risks as they are introduced. Going forward, we will continue to update our recommended measures, policies, best practices, and guidelines to ensure that these technologies can be implemented securely.

We believe this technology will transform numerous industries, and we encourage everyone to start experimenting with them. But because it is uncharted territory that is accessible to anyone with an internet connection, it is important to keep in mind the risks when exploring this technology.

What Can You Do Starting Today?

The goal of this document is to propose practical examples of policies and controls for preventing or mitigating security-related risks when using ML and/or AI-based tools. We have identified eight risks, and for each one, we have proposed several policies and controls. We strongly advise taking immediate action as there is a significant chance that colleagues at your company are already using AI applications and without guidance. Therefore, it is crucial to conduct a thorough risk assessment on the usage of these applications within your company and ensure that the proper governance, guidelines, and technological guard rails are in place. You can always count on Northwave Cyber Security to assist you in implementing these security policies and controls.

Disclaimer And Acknowledgements

We used OpenAI API and ChatGPT to support the production of this document. Highly specialised colleagues from Northwave were essential in the curation and production of this meaningful and practical guidance, including Rob Berends, Jair Santanna, Christiaan Ottow, Rob Braun, Fook Hwa Tan, and Evi Hagenaars.

About Northwave Cyber Security

Founded in 2006, Northwave Cyber Security is the leading Dutch interdisciplinary specialist in cyber security, with offices in Utrecht, Leipzig, and Brussels. With their managed cyber security services, they enable European clients to remain in control while placed under the permanent protection of their confident cyber crew. Their integrated approach towards cyber risk mitigation delivers solid security and aims for cyber awareness and resilience.

Get in contact with us

Currently facing a security incident?
Call day and night: 00800 1744 000

Contact

E: info@northwave.nl
T: +31 (0) 30 303 1240
W: northwave-cybersecurity.com

NORTHWAVE
CYBER SECURITY