CERT-EU Security Guidance 23-002

Potential impact and risks of Generative AI in EUIBAs

CERT-EU Team
ver. **1.2**
May 31, 2023

Contents

1 Introduction

1.1 Disclaimer

Please note that portions of this document have been generated with the assistance of ChatGPT using the GPT-4 model developed by OpenAI. While the content provided by the language model has been significantly edited and corrected, it is one of the aims of this document to demonstrate the possible use of a Generative AI technology.

1.2 Contact

If you have suggestions that could help improve this document, please contact us at services@cert.europa.eu. We always appreciate constructive feedback.

1.3 Generative AI

A Generative AI is a type of artificial intelligence that focuses on creating new content or data, simulating humanlike creativity and adaptability. These AI systems are designed to learn from vast amounts of data and generate outputs based on their training. They can be used in a wide range of applications, from natural language processing to computer vision and beyond. For example, generative AI can be employed to create realistic images, draft humanlike text, compose music, or even design novel chemical compounds. Notable examples include large language model transformers (LLMs) such as OpenAI's GPT series, Google's Bard and Meta's LLaMA, which can generate coherent and contextually relevant text, but also text-to-image generation tools such as DALL-E from OpenAI, or Stable Diffusion form Stability AI.

Large language model transformers are advanced deep learning models that utilise the so-called *self-attention mechanisms* and multi-layer architectures to understand and generate humanlike text. They excel at tasks such as language translation, summarisation, and question-answering by analysing vast amounts of data and identifying complex patterns. Transformers' strengths include their ability to capture contextual information, generate coherent and contextually relevant responses, and adapt to a wide range of tasks. However, they have notable weaknesses, such as being data and computationally intensive, requiring significant resources for training and fine-tuning. Additionally, they may produce plausible-sounding but incorrect or nonsensical answers (*hallucinations*), and they can be sensitive to the phrasing of input prompts. Lastly, transformers may inadvertently generate *biased* or *harmful* content due to biases present in the training data.

In turn, text-to-image generation models employ deep learning techniques to create visually coherent images based on textual input. The strengths of these models include their ability to generate diverse and creative images, as well as aiding in data augmentation and visual storytelling. However, weaknesses include their dependence on large, well-annotated datasets for training, high computational requirements, and the possibility of generating unrealistic or low-quality images. Furthermore, they may struggle to accurately capture complex and abstract concepts described in the textual input, and – similar to transformers – may inadvertently propagate *biases* present in the training data.

1.4 How does it work?

Generative AI is not magic, but rather a testament to recent breakthroughs in the field of neural networks and deep learning. The substantial progress made in recent years is primarily attributed to the exponential growth of computational power and the availability of massive datasets for training.

1.4.1 Text generation models

LLMs, at their core, consist of neural networks with millions or billions of parameters (i.e., *model size*) that are trained on vast amounts of text data. The *parameters* describe a number of interconnections between nodes of the neural networks. Before these models can be used, they need to be *trained*. During the training process, they are presented with huge amounts of data (i.e. *training sets*) that allow them to learn patterns, relationships, and structures within the language by adjusting their parameters to minimise prediction errors. They are typically trained using a method called unsupervised learning, which allows them to predict the next word or token in a sequence, given the previous context.

Inference is the process of using a trained language model to generate predictions or complete tasks based on new input data. During this stage, the model leverages its internal knowledge of language patterns and relationships, acquired during training, to produce relevant and coherent output. Inference involves a series of calculations to assign probabilities to potential next words or tokens, ultimately selecting the one with the highest probability. This process is then repeated to generate subsequent words, resulting in a coherent and contextually appropriate response.

1.4.2 Image generation models

Generative AI text-to-image models are designed to create visual representations of textual descriptions. These models consist of neural networks, specifically built upon architectures like Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs). Usually, these models consist of two interconnected networks – a generator that synthesises images, and a discriminator that evaluates the quality and relevance of generated images to the input text. To train such models, large datasets containing paired text and image samples are used. The training process involves teaching the model to understand the relationships between textual descriptions and their corresponding visual representations by minimising the discrepancies between generated images and the actual images associated with the input text. Again, as in case of LLMs, the parameters of the neural network are adjusted during training to minimise prediction errors.

During inference, when a text description is provided to the trained model, it generates a visual representation that best matches the input. The model uses its understanding of the relationships between text and images, acquired during training, to generate plausible and relevant visual content. Inference typically involves sampling from the model's *latent space*, which is a compressed representation of the complex relationships between text and images. The model then maps this latent space back to the image space, producing a coherent and contextually appropriate visual output.

1.5 Future outlook

In recent months, there has been a substantial leap forward in the development of Generative AI technology. Further rapid advancements are expected as researchers continue to push the boundaries of AI capabilities, explore new techniques such as unsupervised and self-supervised learning, and tap into the potential of quantum computing. These future improvements could lead to more sophisticated language models, enhanced creativity and problem-solving skills, and the ability to generate highly realistic images, audio, video, and even virtual environments.

At CERT-EU, we assess that the development of Generative AI technology will continue, and as such, it is crucial to embrace this transformative innovation. It presents numerous opportunities, such as enhancing creative processes, personalising user experiences, and automating mundane tasks, thereby increasing efficiency. However, it also comes with risks, including the potential

for AI-generated misinformation, information leakage, and ethical concerns surrounding AI-generated content.

2 Focus area

As the potential impact of Generative AIs on society as a whole seems quite transformative and is a huge topic in itself, the scope of this document has been intentionally limited to those areas most relevant to European institutions, bodies, and agencies (EUIBAs). Specifically, this means focusing on the potential impact and valuable insights into the benefits and risks along with some practical recommendations. In addition, due to the volatility of the domain, the information provided here is valid as of the first half of 2023, as further possible advances in the field may require significant changes to be introduced in the future.

Since the core role of CERT-EU is focused on cybersecurity, the cybersecurity-related aspects of Generative AIs is in the centre of our attention, as they have the potential to significantly transform both the defensive and the offensive parts of our field.

On the defensive side, Generative AIs can be utilised to generate realistic attack scenarios for training purposes, enabling security teams to better prepare for and respond to potential threats. Other ideas include the creation of more robust and adaptive honeypots, which can confuse and slow down attackers, or even facilitate the identification and tracking of malicious actors. LLMs have also been shown to be quite efficient in identifying correlations and patterns, as well as various analyses of data that can significantly help analysts in their job.

On the offensive side, Generative AIs can be employed to generate sophisticated social engineering attacks, such as highly personalised phishing emails that are more likely to deceive targets. Furthermore, this kind of AIs can assist in automating the discovery and exploitation of vulnerabilities in target systems, streamlining the process and potentially uncovering previously unknown attack vectors. Additionally, LLMs have also shown to be capable of creating code that can be used for malicious purposes.

With all the potential benefits of this new technology, Generative AI also raises several concerns. Among others, they are related to data protection and copyright issues. Since these AI models can generate realistic and creative content by analysing vast amounts of data, there is a risk that they may inadvertently expose sensitive or private information, violating individuals' right to privacy.

Additionally, Generative AI's ability to produce content resembling human-generated works challenges traditional copyright frameworks. The line between originality and infringement can become blurred, making it difficult to determine the rightful owner of intellectual property. There have been already several lawsuits filled to clarify these matters.

2.1 Deployment considerations

Large Language Models (LLMs) can be deployed and utilised in a variety of ways. Examples include:

- public closed-source models - paid or free;
- locally-hosted open-source models;
- privacy-focused commercial closed-source models with specific conditions of use.

The model dominating the market today is one where a 'free' service is offered as a closed-source *black box*. ChatGPT, Microsoft New Bing, DALL-E, Midjourney, and Google's Bard all fit in this category. Their terms-of-use often make it clear that the input data provided as well

as the output is stored outside EU and may be further used for training and fine-tuning of the models. As such, users of these services should operate under the assumption that all data provided as part of a prompt will become public knowledge.

On the other hand, there are a number of open source models available that can be deployed and run locally – either in a cloud environment or on-premise. LLaMA, Alpaca, BLOOM, Dolly, and various other models available through Hugging Face community[1] all fit in this category. As these systems can be self-hosted, the data inside them or provided to them can be contained by applying general security provisions to the hosted environment.

EUIBAs might also consider the third option - more privacy-focused commercial models with specifically negotiated terms-of-use. Such commercial models, but using different configuration and terms-and-conditions than the public ones – with severe consequences for not respecting those conditions – can be potentially quite interesting.

For example, the European Commission Cloud Centre of Excellence has recently cleared the Azure OpenAI service for Sensitive Non-Classified data for the Commission. In particular, the user prompts are not sent outside EU and are not allowed to be used for training. This model may be used as part of regionally hosted cloud service, under the Cloud Broker's Contracts. Such privacy-tuned hosting modes may be preferable for many organisations over hosting an entirely local model, as the skills and effort needed to train and maintain a local model cannot be underestimated.

3 Benefits

3.1 Benefits of using Generative AI

Generative AI can be a tool to enhance the efficiency and effectiveness of organisations. By automating repetitive tasks, these systems can potentially reduce human error, save time, and allow staff to focus on more complex, value-added activities. Furthermore, the technology's capacity to produce natural language text can aid in the creation of clear, concise, and coherent communication materials, as well as improve analysis and correlation of information.

Considering the specific focus of this document – i.e. impact of Generative AI technology on cybersecurity and EUIBAs in particular – several use cases can be identified where potential benefits stand to be gained.

3.1.1 Enhancing learning

The sophisticated natural language processing abilities of LLM transformers can help in training and bridging the knowledge gap faced by junior cybersecurity staff. By providing contextualised explanations of complex cybersecurity concepts and detailed analysis of examples, LLMs facilitate a more rapid learning process, enabling junior team members to understand and respond to threats more effectively. Moreover, the AI models can offer real-time guidance, ensuring that less experienced staff can contribute to the team's efforts in a meaningful way.

3.1.2 Improving detection rules

AI systems can potentially assist in fine-tuning existing detection algorithms and creating new rules based on the latest threat intelligence. By incorporating the insights gleaned from such analysis, cybersecurity teams can improve their detection capabilities, leading to a more robust defence against cyberattacks.

[1] https://huggingface.co/

Another crucial aspect of a cybersecurity analyst's job is log analysis. LLMs can be used to sift through massive amounts of log data to identify potential anomalies, outliers, or correlations that might indicate a security breach. By automating the log analysis process, LLMs reduce the time and effort required to perform this task, allowing analysts to focus on more pressing issues instead of trying to spot a needle in a haystack. Furthermore, LLMs can help identify correlations between seemingly unrelated events, giving analysts a holistic view of the security landscape and enabling them to respond more effectively to potential threats.

There is a hope that AI systems can aid in enhancing detection rules by analysing massive datasets – such as logs – and identifying patterns that might be missed by human analysts. However, for this particular task, the number of false positives has to be accounted for. Even, if the False Positive (FP) rate is a mere 0.1%, then this still results in a large number of FPs over multiple hundreds of millions of log lines.

According to Microsoft, which has recently released Security Copilot, AI technology should in the future tip the scales in favour of the defenders[2]. Other similar initiatives will undoubtedly follow.

3.1.3 Supporting analysis

LLMs have also enabled analysts to work more efficiently and accurately in various aspects of their jobs. One key area where they have demonstrated their potential is in the deobfuscation of malicious code. Attackers often obfuscate their code to evade detection, but LLMs can assist analysts in identifying patterns and decoding hidden algorithms, providing valuable insights into the attacker's intent and revealing the true nature of the threat, thereby significantly speeding up the investigation.

Additionally, LLMs transformers show a remarkable ability to correlate data from various sources and fields as they have been trained on diverse datasets. This vast training corpus enables the AI model to extract and synthesise information from various seemingly unrelated sources. By leveraging its deep learning capabilities, LLMs can then identify connections, patterns, and insights across these different sources. As a result, the AI model has proven to be a valuable tool in solving problems, providing novel insights, and identifying correlations that would otherwise be easy to miss.

Code analysis and reverse engineering are also areas where LLMs can provide substantial assistance. With their extensive knowledge base, LLMs can evaluate software code and explain its operation. In the case of reverse engineering, LLMs can help dissect complex obfuscated code and provide insights into its functionality and purpose. By understanding how a piece of malware or exploit operates, analysts can develop effective countermeasures to protect their systems and networks, while also saving time during the investigation.

VirusTotal recently announced integration of a new feature based on Sec-PaLM model to produce natural language summaries of code snippets[3]. Similar such features are likely to be integrated into many reverse engineering, sandboxes, and analysis tools in the future.

3.1.4 Automating threat intelligence

LLM transformers can greatly enhance the process of generating threat intelligence reports by automating the collection, analysis, and summarisation of relevant data. This not only saves time and effort, but it also ensures that the information presented to cybersecurity teams is

[2] https://blogs.microsoft.com/blog/2023/03/28/introducing-microsoft-security-copilot-empowering-defenders-at-the-speed-of-ai/

[3] https://blog.virustotal.com/2023/04/introducing-virustotal-code-insight.html

accurate, up-to-date, and easily digestible. Armed with this intelligence, defenders can make more informed decisions and take proactive measures to protect their systems.

3.1.5 Coding and documentation

LLM transformers have already shown their impact on the field of software development. These AI models can assist developers in various ways, including code completion, bug detection, and automatic documentation generation. By accurately predicting and suggesting relevant code snippets, LLM transformers expedite the coding process, reduce human error, and improve overall code quality. Additionally, these models can potentially analyse complex codebases, identifying vulnerabilities or inefficiencies that may otherwise go unnoticed.

Several offerings have sprung up that specifically aim at supporting developers in their tasks. Examples include Github Copilot X[4], Amazon CodeWhisperer[5], Google Bard[6] or Tabnine[7]. Initial feedback from most developers indicates that while these tools are not yet completely replacing the human developers, they clearly provide a significant boost to the developers' productivity.

In addition to actual code generation, Generative AI appears to be particularly well suited to help in several mundane tasks that are sometimes shunned by developers, such as writing documentation or unit tests. As LLMs can accomplish a large part of such tasks autonomously, they can provide a significant productivity boost.

At the same time, it is worth remembering that the generated code is based on large training datasets of public code discussion forums such as Stackoverflow[8] or other uncurated sources. This means that there is a possibility that both trivial and non-trivial bugs present in the code in the sources will also be represented in the LLM-generated code.

3.1.6 Content generation

Finally, one of the most obvious ways that Generative AI can be used is in creating high-quality content across various domains, including marketing materials, corporate communications, and presentations. The ability of AI-driven content generation platforms to understand context and produce humanlike text can enhance organisations' approach to content creation. The expected widespread adoption of generative AI can be attributed to its impressive capacity to save time, reduce costs, and increase overall efficiency in producing diverse forms of content.

In the realm of marketing, generative AI enables the creation of targeted and persuasive texts, advertisements, and social media content tailored to specific audiences. This allows marketeers to better engage with their target demographic, leading to improved conversion rates and ROI. Additionally, generative AI could be a powerful tool in corporate communication, as it helps to develop clear, concise, and accurate messaging, while maintaining brand consistency and a professional tone. This is leading to a significant reduction in the need for manual editing and proofreading, thereby increasing productivity.

Similarly, AI-driven content generation can be used to improve presentations. By leveraging data, generative AI can dynamically construct visuals, generate relevant talking points, and even suggest persuasive storytelling techniques to captivate audiences. This can not only streamline the process of creating presentations but also enhance their overall impact and effectiveness.

[4] https://github.com/features/preview/copilot-x
[5] https://aws.amazon.com/codewhisperer/
[6] https://blog.google/technology/ai/code-with-bard/
[7] https://www.tabnine.com/
[8] https://stackoverflow.com

Similarly to code generation examples, there are already several products either available or being rolled out that propose such AI-based enhancements and solutions. These include for instance Microsoft 365 Copilot[9] and Google AI-powered Workspace Features[10].

4 Risks

Risks associated with the use of Generative AI can be subdivided into two categories:

- the risks of using it, and
- the risks of it being used by others.

4.1 Risks of using Generative AI

Specific risks can be identified that stem from the possible use of Generative AI technology by the staff of EUIBAs. As in case of the benefits of this technology, the focus is squarely on the cybersecurity aspects and EUIBAs in particular.

4.1.1 Indirect prompt-injection attacks

As the Generative AI technology evolves, new possibilities, but also risks emerge. Recent emergence of various plugins that may be used in conjunction with some of the large language models increase their capabilities, but also introduce new risks.

One of these is the possibility of *indirect prompt-injection attacks*. Plugins generally allow language models to use external data - websites, documents, emails, etc. Such external data which may be under the control of malicious actor may allow an attacker to attempt influencing the model's output by carefully crafting their input or 'prompt', often embedding hidden instructions or biases. The AI model then inadvertently generates output that could potentially spread misinformation, reveal sensitive information, or produce other undesirable outcomes. Despite the input appearing harmless or neutral to a human observer, it can result in manipulated outputs, thus presenting significant security concerns in the application of AI technologies.

Indirect prompt-injection attacks are already occurring in the wild, with various plugins allowing Large Language Models (LLMs) to access external data sources, such as webpages, being used as vectors for these attacks. Examples include hidden text in webpages, or inside documents or e-mails that are then provided as input to LLMs buy unsuspecting users.

A significant challenge here is that the existing defences are not currently equipped to effectively counter these attacks. The subtlety of the manipulation makes detection extremely difficult, especially as the injected prompts often appear harmless or neutral to human observers, or are not easily visible at all. While it is possible to configure the models to ignore certain types of these attacks, or specific prompts, there is no obvious way to create a permanent fix. Users should be aware of using AI tools on any input that may have been subject to malicious modification (e.g., webpages, external documents, incoming e-mails, etc.).

4.1.2 Disclosure of sensitive data

The use of freely available, closed-source AI language models, such as ChatGPT, poses potential risks to sensitive data provided in user prompts. As users interact with the model, they might inadvertently input confidential or personally identifiable information (PII) while seeking assistance or answers. Since this information is usually stored for the model to process and

[9] https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/
[10] https://workspace.google.com/blog/product-announcements/generative-ai

generate responses, there is a chance that this sensitive information could be exposed, either through data breaches or during the training of future iterations of the AI. Subsequently, without proper data anonymisation and privacy measures in place, such information could be misused by unauthorised parties, leading to identity theft, financial fraud, or reputational damage for both individuals and organisations involved.

For instance, current OpenAI terms-of-use[11] specify in particular that while OpenAI does not use the *API Content* for improving their services, it *may* use the *Non-API-Content* (i.e. prompts and output of ChatGPT, for instance) to do so. Hence, if confidential or sensitive data is entered as part of a ChatGPT prompt, it may eventually be leaked into the public domain. OpenAI claims – as of the time of this writing – that the requests submitted to their *API* will be stored for 30 days[12] and not be used for training. But we have no proof of compliance nor any insight regarding the future plans of OpenAI.

In the event of a cyberattack on the infrastructure of an AI language model, there is a significant risk associated with the potential leakage of data. Such a breach could expose sensitive and private user information, including personal details, confidential conversations, and intellectual property. The fallout from this exposure could have wide-ranging consequences, including compromised privacy, loss of user trust, and potential legal ramifications.

The risk of sensitive data disclosure while using large language models (LLMs) can be significantly mitigated by carefuly choosing the model deploymnet method. Employing a local, open-source model, hosted on-premise and under the direct control of the organisation utilising it, or using a vetted commercial service offering a privacy-focused environment could be a potential solution.

Such approach allows for enhanced data security, as information is processed and stored within the organisation's own infrastructure or in a dedicated tenant with specific governing rules. Dedicated privacy-focused services may require additional investments. Local models in turn require investment and specific skills. In either case, this approach reduces the possibility of unauthorised access or data breaches, as the organisation can implement strict security protocols and closely monitor the model's usage. Furthermore, the organisation can ensure compliance with relevant data protection regulations and adapt the model to suit its specific requirements, ultimately fostering trust and confidence in the system's ability to handle sensitive information securely.

4.1.3 Copyright violations

Generative AI technologies, such as text and image generation models, have raised concerns about potential copyright violations as they become increasingly adept at creating content that closely resembles human-authored works. In the realm of text generation, AI-powered tools can produce articles, stories, or even poetry, often blurring the lines between human creativity and synthetic, machine-generated output. This raises questions about the originality of the content, and whether or not the AI system has unintentionally reproduced or closely mimicked copyrighted materials.

For instance, if a text-generation AI model creates a story that closely resembles a popular novel, the copyright holder of the original novel may claim infringement, as the AI-generated work could be perceived as a derivative of their copyrighted material.

Similarly, image-generation models have the ability to create visually attractive artwork, designs, and even photorealistic images. These AI-generated images could infringe upon copy-

[11] https://openai.com/policies/terms-of-use
[12] https://platform.openai.com/docs/guides/chat/do-you-store-the-data-that-is-passed-into-the-api

righted visual content if they closely resemble existing works, such as paintings, photographs, or graphic designs.

For example, if an image-generation AI model were to create an artwork strikingly similar to a famous painting, it could lead to copyright disputes between the original artist and the creator of the AI-generated piece. Moreover, these concerns extend to the potential appropriation of elements from multiple copyrighted works to create a new image, which could lead to multiple copyright violation claims.

In both cases, the increasing sophistication of generative AI technologies complicates the legal landscape surrounding copyright protection, as it becomes more challenging to determine the true authorship and originality of content.

Additionally, in some cases, the models powering generative AI technologies are known to be trained on copyrighted content without the explicit approval of the authors. This raises additional concerns, as the organisations behind these models could be held liable for potential copyright infringement. By using copyrighted material to train their AI systems, organisations may inadvertently propagate the unauthorised reproduction or adaptation of protected works, opening themselves up to potential litigation. As a result, there is a growing need for more robust and transparent content acquisition policies, ensuring that the data used to train AI models is either appropriately licensed or falls under the scope of fair use.

4.1.4 False or inaccurate information

AI language models have become increasingly adept at generating high quality text. However, these models have flaws, and the dangers of providing false or inaccurate information remains quite significant[13].

As AI language models are trained on vast amounts of data from the Internet, they are susceptible to absorbing and perpetuating the biases, misconceptions, and inaccuracies that may be present in that data. We should also not confuse Natural Language Processing (NLP), which these models excel at, with Natural Language Understanding (NLU), a significant challenge of AI research. A system trained only on form has *a priori* no way to learn meaning[14]. Consequently, users of these models must be aware of the potential pitfalls and exercise critical thinking when interpreting generated text.

One of the primary concerns with AI language models is the possibility of bias. As these models learn from the data they are trained on, any biases present in the training data will likely be absorbed and perpetuated by the model. This could manifest in the form of gender, racial, or political biases, among others, and can lead to the generation of text that is offensive or perpetuates harmful stereotypes. In some cases, these biases may even cause the AI to provide misleading or outright false information, leading users astray and potentially reinforcing pre-existing biases.

Similarly, when generating computer code in various programming languages, LLMs often provide code that contains errors or is insecure. This is primarily due to the training data that these models are exposed to, which may include a diverse array of programming languages, styles, and quality levels. As a result, LLM-generated code may not always adhere to best practices or conform to the latest security standards. Additionally, these models lack the ability to inherently understand the context or the specific requirements of a given task, which may lead to the production of code that is unsuitable, flawed, or even dangerous. Therefore, it is crucial for developers to carefully review and validate any code generated by LLMs before incorporating

[13] https://dl.acm.org/doi/abs/10.1145/3442188.3445922
[14] https://aclanthology.org/2020.acl-main.463.pdf

it into their projects in order to mitigate potential risks and ensure the safety and integrity of their software applications.

Another concern is the phenomenon of *hallucinations*, where AI language models generate text that appears plausible but is entirely fabricated or lacks a factual basis. These hallucinations can occur for various reasons, such as the model trying to fill in gaps in its knowledge or attempting to provide a coherent response to an ambiguous or unfamiliar prompt. While these hallucinations can sometimes be relatively harmless, in other instances, they can lead to the dissemination of false information or contribute to the spread of misinformation.

4.1.5 Hype abuse

The rapid advancements in Generative AI technology and the surrounding hype have led to a surge in public interest and adoption. While these innovations undoubtedly offer numerous benefits and transformative potential, this excitement can also lead to possible pitfalls. With increased hype, bad actors may exploit the situation by creating fake applications or investment schemes, capitalising on the popularity of recognisable AI brand names to deceive users and fulfil malicious objectives.

One such pitfall is the emergence of fake ChatGPT apps on Android and iOS platforms. These counterfeit apps, disguised as popular AI language models, may carry out invasive data harvesting activities. Unsuspecting users who download and interact with these malicious apps may inadvertently expose their personal information, including messages, contacts, and browsing history. The harvested data can then be used for a wide range of nefarious purposes, such as identity theft, targeted advertising, or even extortion. This highlights the importance of exercising caution when downloading mobile applications and ensuring they originate from trusted sources and developers.

Another potential pitfall linked to the hype around Generative AI is the creation of fictitious cryptocurrency tokens using recognisable AI brand names. Bad actors may design and market these tokens to lure in unsuspecting investors, who may believe they are investing in a promising AI venture. Once the scammers accumulate a substantial amount of funds, they may disappear, leaving the investors with worthless tokens and significant financial losses. This highlights the need for investors to conduct thorough research and due diligence before committing to any investment, particularly in emerging technologies like AI and cryptocurrencies.

4.1.6 Over-relying on technology

Over-relying on Generative AI technology presents several potential dangers that could have a profound impact. One significant concern is the possible loss of competence among staff. As AI systems become more adept at handling tasks traditionally performed by humans, employees may become increasingly dependent on these technologies. This reliance could lead to a decline in critical thinking and problem-solving skills, making staff less versatile and adaptive in the face of novel challenges. Moreover, as AI takes over routine tasks, workers may lose the ability to perform them manually, resulting in a loss of valuable expertise.

Another danger lies in the overconfidence in the quality of output provided by Generative AI. Due to the inherent limitations in AI models, such as the token limit that restricts the amount of information a language model can *remember*, the generated content might not be as accurate, comprehensive, or contextually appropriate as users expect. This could lead to situations where AI-generated content is taken at face value, potentially leading to misinformation or poorly informed decisions.

The over-reliance on AI technologies may also present itself in the form of a failure to account

for policy or political decisions that limit their use. Governments and regulatory bodies are increasingly scrutinising the implications of AI on privacy, security, and social equality. As a result, they may implement policies or regulations that impose restrictions on the development, deployment, or use of AI technologies. Organisations that would become heavily reliant on AI systems might find themselves unprepared to adapt to such changes, leading to operational disruptions.

Finally, as mentioned in the benefits chapter, especially when using LLM tools for programming, it is paramount to remember that the generated code may contain bugs or be otherwise insecure or unsuitable. Hence, extra care must be taken when allowing staff and contractors to use LLMs for developing applications. The focus of attention should shift to ensuring proper validation and testing of the supplied code.

4.1.7 LLMs opinions, advice, and moral values

Large Language Models (LLMs) such as ChatGPT should not be consulted for opinions, advice, or moral values due to the inherent limitations of their design and the nature of their training data. LLMs are powerful AI tools, but they are not human beings with emotions, life experiences, or ethical systems. Instead, they are complex algorithms designed to generate humanlike text based on patterns and associations found in vast amounts of data.

One of the primary reasons that LLMs are not suitable for providing opinions, advice, or moral guidance is that their responses are formulated based on the datasets used in their training. These datasets consist of vast amounts of text from a diverse array of sources, which may contain conflicting opinions, values, and perspectives. When an LLM encounters such conflicts in its training data, it may struggle to generate a coherent and consistent response. The output of the LLM can be quite random, as it attempts to find a balance between opposing viewpoints or may simply reproduce popular opinions without understanding the underlying reasons or nuances.

Moreover, LLMs are not capable of forming independent opinions or moral judgements. They do not possess the ability to critically analyse complex issues or empathise with human emotions, which are essential components of providing sound advice or ethical guidance. Relying on an LLM for such matters could lead to misguided or superficial conclusions that fail to account for the unique complexities of a given situation.

It is hence not surprising that for example China is setting guard rails on what values LLMs must follow (in this case the ones by the Chinese Communist Party)[15]. After all, an LLM will always reflect the moral values that were given by its training data and by the human feedback (Reinforcement Learning with Human Feedback - RLHF). Hence, for any generated text that is intended to be used in a political context, it might also make sense to check if it aligns with the EUIBA's general vision, policy, and strategy.

4.2 Risks of others using Generative AI technology

Specific risks stem from the use of Generative AI technology by malicious actors. As before, the focus is on cybersecurity and EUIBAs in particular.

4.2.1 Privacy issues

Personal Identifiable Information (PII) can inadvertently become part of the training datasets of Generative AI models when data is collected from a wide range of sources, such as websites,

[15] https://www.nytimes.com/2023/04/24/world/asia/china-chatbots-ai.html

forums, social media, and other digital platforms. This data may not be adequately anonymised or sanitised before being fed into the AI model. As a result, PII may be embedded within the model's training data, which could include names, addresses, phone numbers, email addresses, or other sensitive information that can be linked back to specific individuals.

When these Generative AI models are used, this PII may be unintentionally disclosed or made public, which poses serious privacy concerns. Furthermore, the information generated by the AI might be incomplete or inaccurate, which could lead to misinformation or misidentification of individuals. This creates a dual problem: on one hand, the disclosure of sensitive information can have severe consequences for affected individuals, while, on the other hand, the generated data may be unreliable, potentially causing harm to both the individuals and entities relying on it.

4.2.2 More advanced cyberattacks

Generative AI technologies could also enable new methods for conducting cyberattacks. As AIs become more sophisticated, they can increasingly be utilised by malicious actors to facilitate their attacks and exploit vulnerabilities in various ways.

One such method involves using AI to generate phishing content. By leveraging natural language processing and generation capabilities, cybercriminals can create highly convincing emails, text messages, and social media posts that appear to come from legitimate sources. These AI-generated messages can be tailored to target specific individuals, making them more likely to fall for the scam. Additionally, the AI can be used to automate the process of sending phishing messages, enabling attackers to target a broader range of potential victims.

Social engineering attacks can also be enhanced by using AI-generated voice and video deep fakes. These realistic forgeries can be used to impersonate executives, celebrities, or other influential individuals to manipulate victims into providing sensitive information or performing actions that benefit the attacker. Deep-fake technology can also be employed in creating more believable phone scams or video calls, increasing the likelihood of a successful attack.

Moreover, AI technologies can be used to improve malware by making it more difficult to detect and more effective in its operations. For instance, AI algorithms can be employed to analyse existing malware and identify patterns that are likely to be flagged by antivirus software. Based on this analysis, AI can then generate new, stealthier malware variants that can evade detection and better exploit system vulnerabilities.

Another way AI can facilitate cyberattacks is through more efficient vulnerability detection and fuzzing. By using AI-powered tools, attackers could automatically discover security weaknesses in software or network infrastructure at a much faster rate than traditional methods. This would allow them to identify and exploit vulnerabilities before they are patched, increasing the likelihood of a successful attack.

As yet another example, AI can be used to automate and optimise the process of password cracking. By using machine learning algorithms, attackers can identify patterns in password creation and generate more effective password dictionaries to speed up the cracking process. This can significantly reduce the time it takes to gain unauthorised access to accounts, making it more difficult for security professionals to respond to attacks.

Finally, the development of freely available Generative AI tools has inadvertently lowered the barrier for entry for new malicious actors in the cybercrime ecosystem. With minimal technical expertise, these individuals can exploit the capabilities of advanced AI models to conduct a variety of nefarious activities, such as generating phishing emails, crafting realistic deep-fake content, or creating fake news. This democratisation of access to powerful AI-driven tools

amplifies the potential reach and impact of cybercrime, cyberespionage, and other forms of malicious activities as it allows a broader range of actors to participate in these activities, posing significant challenges for cybersecurity professionals, law enforcement, and policymakers.

4.2.3 Disinformation

Generative AI models' powerful capabilities come with significant dangers when they are used maliciously for disinformation campaigns. These models are capable of impersonating public figures and crafting highly convincing narratives, making them potent tools for spreading false and misleading information. For instance, deep-fake technology has allowed bad actors to create fake videos and audio clips of politicians and celebrities, manipulating their words and actions in order to deceive the public and create confusion.

Generative AI models can be employed to design believable disinformation campaigns, which have the potential to undermine trust in institutions, destabilise social cohesion, and disrupt democratic processes. For example, during election periods, a sophisticated AI-generated disinformation campaign could manipulate the public discourse by disseminating false news stories, conspiracy theories, and divisive content. This can have far-reaching consequences as it sways public opinion based on falsehoods, ultimately leading to an erosion of trust in the democratic process.

The fact that these disinformation campaigns can be planned ahead and automated significantly exacerbates the problem, as it enables bad actors to generate and disseminate false information at an overwhelming scale. It could be extremely challenging for fact-checkers, journalists, and social media platforms to identify and counteract the spread of disinformation in a timely manner. Additionally, the speed and efficiency of AI-generated content can make it difficult for users to discern between legitimate and fake news, further facilitating the spread of disinformation.

4.2.4 Censorship and control

Large AI models can also be employed by authoritarian governments to manipulate public opinion and suppress democratic processes. By using these sophisticated technologies to generate fake news, propaganda, and deep-fake content, such regimes can create an illusion of reality that suits their interests. This disinformation can sow confusion and mistrust among the public, undermining the credibility of democratic institutions and opposition leaders.

Additionally, authoritarian governments can use AI-powered surveillance systems to identify and monitor the activities of opposition members and dissidents. By analysing vast amounts of data from social media, communications, and location tracking, these models can create detailed profiles of individuals deemed a threat to the regime. The regime can then use this information to suppress opposition voices through harassment, arrests, and other forms of repression.

5 Conclusions

Generative AI technology has emerged as an important innovation with potential to disrupt various industries and aspects of society. It is a tool capable of creating high-quality content, designs, and simulations by learning patterns from vast amounts of data. This technology has far-reaching applications across diverse fields.

However, the immense potential of generative AI technology is not without its inherent dangers. The capability of generating realistic content raises ethical concerns, as it can be exploited to produce fake news, deep fakes, and misinformation, potentially disrupting societies

and undermining trust in institutions. Additionally, the technology's rapid progress may result in unprecedented job displacement, as AI-generated content and automation replace human labour in various sectors. Furthermore, biases present in training data may lead to unintended consequences, perpetuating discrimination and inequality.

Generative AI is not a product of magic; rather, it is built on solid technical foundations derived from years of research and advancements in machine learning, neural networks, and computational power. This technology utilises sophisticated algorithms and models to process, analyse, and learn from vast amounts of data, enabling the creation of intricate and nuanced outputs. The current state of the art in generative AI, impressive as it is, represents only a stepping stone towards even more efficient and capable tools in the future. As the understanding of artificial intelligence deepens, and as computational capabilities continue to grow, it can be expected that further breakthroughs will push the limits of what generative AI can achieve, unlocking new possibilities and applications across various domains.

The inevitability of progress has always been a driving force in human history, pushing the boundaries of what is possible and transforming the way humans live and work. Generative AI, as a part of this relentless march of innovation, is most likely here to stay. Its potential to reshape industries and create new opportunities is immense, making it a technology that organisations will have to embrace and harness in order to remain competitive and relevant. Ignoring or resisting this technological revolution is not a viable option, as others will inevitably capitalise on its benefits, including those with malicious intent. It is therefore imperative for organisations to proactively integrate generative AI into their strategies, while simultaneously working towards establishing ethical guidelines, security measures, and risk mitigation strategies to ensure the responsible and secure utilisation of this new technology.

5.1 Recommendations

We provide below recommendations that can help EUIBAs in directing and coordinating their efforts with regards to Generative AI. Given the volatile nature of the field of Generative AI, they can be divided in short, medium, and long-term recommendations.

5.1.1 Short-term

- Follow attentively the developments in the field of Generative AI, as it will most likely have significant impacts on several aspects of your operation.
- Invest in user awareness to ensure that this technology is used in a safe and responsible way. Make sure the staff correctly understands the benefits and risks.
- Have a clear policy to only use TLP:CLEAR (public) data in any prompt submitted to public large language models provided online by commercial companies.
- Investigate services available under the Cloud Broker PaaS Contracts, possibly Azure Open AI or other services that may soon join, can provide the best of both worlds – the convenience of using commercial models with additional privacy and security provisions.
- Local, open-source models are making rapid progress. It is worth monitoring these developments, as they possibly offer additional fine-tuning (training) possibilities on local (also sensitive) data.

5.1.2 Medium- to long-term

- Create policies ensuring responsible use of Generative AI technology – including the definition of acceptable use cases and proper validation of the output.
- If useful, plan a strategy to deploy local, open-source Generative AI technology models on-premise or in private Cloud, which would allow for a much better control over the

data.

6 Credits

We would like to warmly thank our colleagues from European Commission for the useful input, feedback and suggestions they have provided to improve this guidance.

TLP Definition

TLP	Disclosure	Message
RED	Not for disclosure, restricted to participants only.	Recipients may not share TLP:RED information with any parties outside of the specific exchange, meeting, or conversation in which it was originally disclosed.
AMBER	Limited disclosure, recipients can only spread this on a need-to-know basis within their organisation and its clients.	Recipients may share TLP:AMBER information only with members of their own organisation.
AMBER+STRICT	Limited disclosure, recipients can only spread this on a need-to-know basis within their organisation only.	Recipients may share TLP:AMBER+STRICT information only with members of their own organisation.
GREEN	Limited disclosure, restricted to the community.	Subject to standard copyright rules, TLP:GREEN information may be distributed with peers and partner organisations within their sector or community, but not via publicly accessible channels.
CLEAR	Disclosure is not limited.	TLP:CLEAR Recipients can spread this to the world, there is no limit on disclosure.