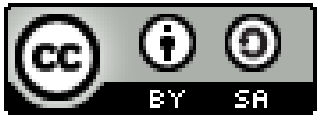
# THREAT MODELING GENERATIVE AI SYSTEMS

## USING PLOT4AI

ISABEL BARBERÁ & MARTIJN KORSE
RHITE
Rhite.tech

# Threat Modeling Generative AI Systems

*Report authored by: Isabel Barberá & Martijn Korse*
*24 April 2023*





This report contains an overview of a non-exhaustive library/list of potential threats of generative AI systems. The library has been published for the purpose of guidance and it does not guarantee the coverage of all existing potential threats.

## SUMMARY

This document offers an overview of different potential threats of generative AI systems. The threats were identified during a privacy threat modeling session we held at Rhite using the AI risk assessment tool PLOT4ai.

The overview offers 63 potential threats classified in the 8 categories of PLOT4ai: Technique & Processes (9), Accessibility (6), Identifiability & Linkability (3), Security (12), Safety (3), Unawareness (3), Ethics & Human Rights (14), Non-compliance (13).
Every threat has also been assigned to different subcategories. This sub-classification is based on Rhite's current research SARAI™, a self-assessment tool for Responsible AI that uses subcategories that have been aligned with the EU Ethics Guidelines for Trustworthy AI and the OECD principles.

For every identified threat we provide extra information together with guiding questions. These questions aim to help stakeholders identify if the potential threat could apply to them.

Each threat has also been assigned to an Action Owner. In this report, Action Owners are the stakeholders that need to take action to mitigate the threats. As Action Owners, we have identified three main roles:

| | |
|---|---|
| **Provider** | The one that develops the AI system, puts it into service or places it into the market. |
| **(Business) user** | The one distributing, putting into service or placing into the market an AI system from a provider. Business users can also customize and fine-tune the original AI system and in some cases, they can also use their own training data. This results in a new product using the AI system from the provider as a foundation. |
| **End-user**[1] | The one using the system in the course of a personal or professional activity and that is often the recipient of the output. |

[1] The main goal of this report is to raise awareness, not only among providers, but also among any type of users of the technology.

## HOW TO USE THE LIBRARY OF THREATS

During the threat modeling session, we identified 63 different potential threats that could apply to generative AI systems.
**Stakeholders that want to develop, implement or make use of generative AI can use this library to identify potential threats.**

Who should use this library:

- People and Organisations developing generative AI systems
- People and organisations that want to implement generative AI systems at work or clients
- Any user of generative AI that wants to become aware of what could go wrong using generative AI systems

### STEPS

- Gather with stakeholders that have the necessary knowledge and expertise, members of your team that are part of the decision-making process, and any stakeholder that could eventually be impacted by the implementation of the AI system. More guidance can be found in Appendix I.
- Go through the library and agree with your stakeholders about which threats could apply to you. When you are not sure, mark the potential threat to further investigate at a later moment.
- Once you have selected the threats that could apply to you, ideally you would classify them depending on their severity and likelihood. More guidance on the classification can be found in Appendix II.
- Once threats are classified, you can decide with your stakeholders which mitigation measures to apply, make a plan to implement the measures, and assign action owners and deadlines.
- Threat Modeling should be an iterative process: once you have implemented mitigation measures it is advisable to do a new threat modeling session to check if new threats have appeared.

## LIBRARY OF POTENTIAL THREATS FOR GENERATIVE AI SYSTEMS

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Technique & Processes | Scope / Data Governance / Accountability | You have not defined the tasks you will use the tool for. | ▪ Have you defined the scope of the tool and implemented safeguards to stay within the scope? (For instance, excluding malicious usage)<br>▪ Is the business problem you want to solve with the tool well-defined? Are the possible benefits clear?<br>▪ Have you defined which types of use are permitted and have employees been informed about this? | Provider/ Business user / End-user |
| Technique & Processes | Bias / Fairness | The input/output data is not representative of different groups/populations. | ▪ It is important to reduce the risk of bias and different types of discrimination. Did you consider if there is enough diversity and representativeness of users/individuals in your training data?<br>▪ Generative AI models can replicate stereotypes and discrimination through biases contained in training data and models (also pre-trained models).<br>▪ Model validation and verification are crucial to assess and eliminate biases before a system's deployment. | Provider / Business user |
| Technique & Processes | Explainability / Transparency / Accountability | The model needs to be explainable to the users or affected persons. | Do you need to be able to give a clear explanation to the user about the logic that the generative AI system used to reach a certain decision/output? And can that decision have a big impact on the user? | Provider / Business user |
| Technique & Processes | Technique / Robustness | You might not be preventing Data Leakage. | Data Leakage is present when your features contain information that your model should not legitimately be allowed to use, leading to an overestimation of the model's performance. Are you applying measures to prevent it? | Provider / Business user |
| Technique & Processes | Technique / Robustness | You might not be preventing Concept and Data Drift. | Data Drift weakens performance because the model receives data on which it hasn't been trained. With Concept Drift the statistical properties of the target variable, which the model is trying to predict, change over time in | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| | | | unforeseen ways causing accuracy issues. Are you applying measures to prevent it? | |
| Technique & Processes | Human Oversight | Human intervention is necessary to oversee the decision-making process or the output of the generative AI system. | ▪ Do humans need to review the process and the decisions/outputs of the generative AI systems? Consider the impact that this could have on the organisation.<br>▪ Do you have enough capacitated employees available for this role? | Provider / Business user |
| Technique & Processes | Data Quality / Data Governance | You cannot confirm the legitimacy of the data sources that you use. | ▪ Data lineage can be necessary to demonstrate trust as part of your information transparency policy, but it can also be very important when it comes to assessing the impact on the data flow.<br>▪ Do you know where you got the data from? Who is responsible for the collection, maintenance, and dissemination? Are the sources verified? Do you have the right agreements in place? Are you allowed to receive or collect that data? Can your generative AI system provide the correct sources of information it is basing its outputs on, also to end-users? | Provider / Business user / End-user |
| Technique & Processes | Human Oversight | You don't have enough dedicated resources to monitor the algorithm. | Do you already have a process in place to monitor the quality of the output and system errors? Do you have the resources to do this? | Provider / Business user |
| Technique & Processes | Data Availability / Data Governance | You cannot collect all the data that you need for the purpose of the algorithm. | Could you face difficulties obtaining certain types of data? This could be due to different reasons such as legal, proprietary, financial, physical, technical, etc. | Provider / Business user |
| Accessibility | Human Interaction / Fairness | Your system's user interface cannot be used by those with special needs or disabilities. | Can your AI system be accessible and usable for users of assistive technologies? | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Accessibility | Human Interaction / Human Agency / Accountability | A redress mechanism might need to be offered to the users. | For applications that can adversely affect individuals, you might need to consider implementing a redress-by-design mechanism where affected individuals can request remedy or compensation. | Provider / Business user |
| Accessibility | Human Interaction / Accountability | An age gate might need to be implemented to use your product. | Is your product not meant to be used by children? You might need to implement an age verification mechanism to prevent children from accessing the product. | Provider / Business user |
| Accessibility | Human Interaction / Human Agency / Accountability | The required information that needs to be provided to users when they need to provide consent, is not made easily available (based on GDPR). | Can the information be easily accessible and readable for end-users? | Provider / Business user |
| Accessibility | Human Interaction / Human Agency | The users might perceive the message from the AI system in a different way than intended. | ▪ Is the perception of the provided information the same as the one intended? You might need to create ways to warn users about this.<br>▪ Explainability is critical for end-users to take informed and accountable actions. | Provider / Business user |
| Accessibility | Awareness / Human Interaction /Accountability | The learning curve of the product could be an issue. | ▪ Does usage of the generative AI system require new (digital) skills? How quickly are users expected to learn how to use the product?<br>▪ Difficulties to learn how the system works could also bring the users in danger and have consequences for the reputation of the product or organisation.<br>▪ Users should learn how to use appropriate prompts to avoid generating output that is harmful or inaccurate.<br>▪ Offer instructions and training to employees and end-users. | Provider / Business user |
| Identifiability & Linkability | Privacy / Data Protection | The data used to feed the model could be linked to individuals. | Do you need to use unique identifiers in your training dataset? If personal data is not necessary for the model you would not really have a legal justification for using it. | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Identifiability & Linkability | Privacy / Data Protection / Human Agency | Actions could be incorrectly attributed to an individual or group. | Your generative AI system could have an adverse impact on individuals by incorrectly attributing them to facts or actions. | Provider / Business user |
| Identifiability & Linkability | Privacy / Data Protection | You could be revealing information that a person has not chosen to share. | How can you make sure the product doesn't inadvertently disclose sensitive or private information during use? | Provider / Business user |
| Security | Security / Robustness | Your generative AI system might need to be red team/pen tested. | You need to test the security of your generative AI system before and after deployment. | Provider / Business user |
| Security | Security / Robustness | Your generative AI system might not be protecting the queries that are stored online. | ▪ Queries stored online might be hacked, leaked, or made publicly accessible (also by accident).<br>▪ End-users should be informed that sensitive information should not be inserted in the queries/prompts. | Provider / Business user / End-user |
| Security | Security / Robustness | Your APIs might not be securely implemented. | APIs enable software systems to interact and share data. APIs are common attack targets in security and are in some sense your public front door. They should not expose information about your system or model. | Provider / Business user |
| Security | Security / Robustness | Your data storage might not be well protected. | Is your data stored and managed in a secure way? Think about training data, tables, models, etc. Are you the only one with access to your data sources? | Provider / Business user |
| Security | Security / Robustness | You might not be protected from insider threats. | AI designers and developers may deliberately expose data and models for a variety of reasons, e.g., revenge or extortion. | Provider / Business user |
| Security | Security / Robustness | You might not be protected against model sabotage. | Sabotaging the model is a nefarious threat that refers to exploitation or physical damage of libraries and machine learning platforms that host or supply AI services and systems. | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Security | Security / Human Rights | You might not be protected against possible malicious use, misuse, or inappropriate use of your generative AI system. | ▪ Your AI system could be used to spread misinformation and disinformation; for example, a chatbot being misused to spread fake news.<br>▪ Generative AI models can be used to create more effective cyberattacks, such as spear-phishing emails, and they can be misused to develop malicious software code or create personalised scams and fraud.<br>▪ End-users might fall victim to maliciously used AI systems. | Provider / Business user / End-user |
| Security | Security / Robustness | Environmental phenomena or natural disasters could have a negative impact on your generative AI system. | Environmental phenomena may adversely influence the operation of IT infrastructure and hardware systems that support AI systems. Natural disasters may lead to unavailability or destruction of the IT infrastructures and hardware that enables the operation, deployment, and maintenance of AI systems.<br>Such outages may lead to delays in decision-making, delays in the processing of data streams, and entire AI systems being placed offline. | Provider / Business user |
| Security | Security / Robustness | You might not be protected from poisoning attacks. | ▪ In a poisoning attack, the goal of the attacker is to contaminate the machine model generated in the training phase. This attack could also be caused by insiders.<br>▪ Data tampering: Actors like AI/ML designers and engineers can deliberately or unintentionally manipulate and expose data. Data can also be manipulated during the storage procedure and using some processes like feature selection. This type of threat can also bring severe discriminatory issues by introducing bias. An attacker who knows how a raw data filtration scheme is set up may be able to leverage | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| | | | that knowledge into malicious input later in system deployment.<br>▪ Adversaries may fine-tune hyper-parameters and thus influence the AI system's behaviour. Hyper-parameters can be a vector for accidental overfitting. In addition, hard-to-detect changes to hyper-parameters would make an ideal insider attack. | |
| Security | Security / Robustness | You might not be protected from adversarial example. | An adversarial example is input from a malicious entity sent with the sole aim of misleading the machine learning system. Deep learning architectures are known to be vulnerable to adversarial examples. | Provider / Business user |
| Security | Security / Robustness | You might not be protected from malicious AI/ML providers who could recover training data. | Malicious ML providers could query the model used by a customer and recover this customer's training data. This could be the case if the training process is either fully or partially outsourced to a malicious third party. | Provider / Business user |
| Security | Security / Robustness | You might not be protected from exploits on software dependencies of your generative AI systems. | In this case, the attacker does not manipulate the algorithms, but instead exploits traditional software vulnerabilities such as buffer overflows or cross-site scripting. | Provider / Business user |
| Safety | Safety | You do not have a mechanism implemented to stop the processing in case of harm. | Do you have a way to identify when your AI system is causing harm, and do you have a mechanism to mitigate the adverse impacts? | Provider / Business user |
| Safety | Environmental Wellbeing | You are using generative AI models that demand the consumption of energy or natural resources beyond what is sustainable. | Could your generative AI system have an adverse impact on the environment?<br>Your product should be designed with the dimension of environmental protection and improvement in mind. | Provider / Business user / End-user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Safety | Human Agency | You do not have measures in place to avoid the generative AI system to become persuasive causing harm to the individual. | ▪ If the generative AI system can achieve reciprocity when interacting with humans, could there be a risk of manipulation and human complacency?<br>▪ Reciprocity is a social norm of responding to a positive action with another positive action, rewarding kind actions. As a social construct, reciprocity means that in response to friendly actions, people are frequently much nicer and much more cooperative than predicted by the self-interest model; conversely, in response to hostile actions, they are frequently much more nasty and even brutal. | Provider / Business user |
| Unawareness | Human Agency / Transparency / Accountability | You are not informing users that they are interacting with a generative AI system. | ▪ Are users adequately made aware that a decision, content, advice, or outcome is the result of an algorithmic decision?<br>▪ Could the AI system generate confusion for some or all users on whether they are interacting with a human or AI system?<br>▪ Users may not be able to distinguish between human and AI-generated text or images. | Provider / Business user / End-user |
| Unawareness | Human Agency / Explainability / Transparency / Accountability | You are not providing the necessary information to the users about possible impacts, benefits, and potential risks. | Did you establish mechanisms to inform users about the purpose, criteria, and limitations of decisions generated by the AI system? | Provider / Business user |
| Unawareness | Human Interaction / Human Agency | Users cannot anticipate the actions of the generative AI system. | Are users aware of the capabilities of the AI system? Users need to be informed about what to expect, not only for transparency reasons but in some cases also for safety precautions. | Provider / Business user |
| Ethics & Human Rights | Bias / Discrimination | There could be groups that might be disproportionately | ▪ Could the generative AI system potentially discriminate against people based on any of the following grounds: sex, race, colour, ethnic or social origin, genetic | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| | | affected by the outcomes of the generative AI system. | features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age, gender or sexual orientation?<br>▪ If your model is learning from data specific to some cultural background, then the output could be discriminating for members of other cultural backgrounds. | |
| Ethics & Human Rights | Social Impact | The generative AI system could have an impact on human work. | ▪ Could the use of your generative AI system affect the safety conditions of employees?<br>▪ Could the AI system create the risk of de-skilling the workforce? (Skilled people being replaced by AI systems). | Provider / Business user |
| Ethics & Human Rights | Social Impact | The generative AI system could have an adverse impact on society at large. | ▪ Could your product be used for monitoring and surveillance purposes?<br>▪ Could the generative AI system affect society at large due to its innovative character?<br>▪ Could the generative AI system affect the right to democracy? | Provider / Business user |
| Ethics & Human Rights | Human Rights | The generative AI system could limit the right to be heard. | Consider for instance the risk if your system makes decisions that could have a negative impact on an individual and you do not offer any way to contest that decision. | Provider / Business user |
| Ethics & Human Rights | Human Rights | The generative AI system could have a big impact on decisions regarding the right to life. | ▪ Consider for instance the risk if your AI system is used in the health sector for choosing the right treatment for a patient. Is the output of the model accurate and fair?<br>▪ Are your datasets representative enough and free from bias? | Provider / Business user |
| Ethics & Human Rights | Human Rights | The generative AI system could affect the freedom of expression of its users. | ▪ Is the output of the model accurate, fair, and not discriminatory? | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| | | | ▪ Consider the risk if this could be used, intended or unintended, to prevent the freedom of expression of individuals, for instance by wrongly labelling text as hate speech. In an example like this, users would not be able to freely express their opinions because their text is wrongly labelled as hate speech and the system blocks the opinion automatically. Imagine the use of generative AI to identify hate speech in the comments of people in a forum. | |
| Ethics & Human Rights | Human Rights | The generative AI system could affect the freedom of its users. | ▪ Is the output of the model accurate, fair, and not discriminatory?<br>▪ Consider the risk if this could be used for monitoring or surveillance purposes; for instance, systems that can spread fake news putting the life of somebody in danger. | Provider / Business user |
| Ethics & Human Rights | Human Rights | The generative AI system could affect the right to a fair hearing. | ▪ Is the output of the model accurate and fair? Consider the risk if this could be used in a criminal case and the consequences if the wrong information is used to condemn someone.<br>▪ Do you have a mechanism to challenge the output of your generative AI system? | Provider / Business user |
| Ethics & Human Rights | Children Rights | Children could be part of your users' group. | ▪ Could your system be used by children?<br>▪ Does the generative AI system respect the rights of the child, for example with respect to child protection and taking the child's best interests into account? | Provider / Business user |
| Ethics & Human Rights | Fairness / Diversity & Inclusiveness | Your generative AI system cannot represent different norms and values without creating ambiguity. | ▪ Is your AI system inclusive?<br>▪ Could cultural and language differences be an issue when it comes to the ethical nuance of your algorithm?<br>▪ Well-meaning values can create unintended consequences. | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| | | | ▪ Must the AI system understand the world in all its different contexts?<br>▪ Could ambiguity in the rules you teach the generative AI system be a problem?<br>▪ Can your system interact equitably with users from different cultures and with different abilities? | |
| Ethics & Human Rights | Fairness / Diversity & Inclusiveness | Your generative AI system is not representing current social needs and social context. | ▪ The datasets that you want to use might not be representative of the current social situation. In that case, the output of the model is also not representative of the current reality.<br>▪ Due to societal and cultural differences, the "ground truth" for generative AI systems is often contextual. | Provider / Business user |
| Ethics & Human Rights | Human Rights | Your generative AI system can have an impact by denying access to jobs, housing, insurance, benefits, or education. | ▪ The output of your model could be used to deny access to certain fundamental rights.<br>▪ How can you be sure that the outputs of your generative AI system are always fair and correct?<br>▪ How can you prevent causing harm to individuals? | Provider / Business user |
| Ethics & Human Rights | Human Interaction / Human Agency | Your AI system can affect human autonomy by interfering with the user's decision-making process in an unintended and undesirable way. | ▪ Could your system affect which choices and which information is made available to people?<br>▪ Humans tend to trust AI systems' output. Could the AI system affect human agency by generating over-reliance by users (too much trust in the technology)?<br>▪ Could this reinforce their beliefs or encourage certain behaviours?<br>▪ Could the AI system create human attachment, stimulate addictive behaviour, or manipulate user behaviour?<br>▪ Generative AI systems can be designed to maximize user engagement and foster addictive behaviours, resulting in negative effects on mental health and well-being. | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Ethics & Human Rights | Human Rights | The labelling process of your training data does not respect the dignity and well-being of the labour force involved. | The need for labelling data grows and unfortunately with that the number of companies providing cheap labelling services at the cost of the dignity and labour rights of their workforce. Is the data that you are using labelled under such conditions? | Provider / Business user |
| Non-compliance | Data Protection / Privacy | You do not comply with the data minimisation principle. | Although it appears to contradict the principle of data minimisation, not using enough data could sometimes have an impact on the accuracy and performance of the model. A low level of accuracy of the AI system could result in critical, adversarial, or damaging consequences. How can you comply with the data minimisation principle? | Provider / Business user |
| Non-compliance | Data Protection / Privacy | You are processing sensitive data. | ▪ According to Art. 9 of the GDPR you might not be allowed to process, under certain circumstances, personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, health data, or data concerning a person's sex life or sexual orientation.<br>▪ You might be processing sensitive data if the model includes features that are correlated with these protected characteristics (these are called proxies). | Provider / Business user |
| Non-compliance | Data Protection / Privacy | Your data might not be accurate or up to date. | ▪ According to Article 5(d) of the GDPR, personal data shall be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy').<br>▪ Input data should be accurate, complete, and trustworthy in order to avoid the known principle of "garbage in, garbage out". Your AI system is only as | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| | | | reliable as the data it works with. Take measures to limit the impact of inaccuracies in your input data.<br>▪ Output data should be correct and accurate. Generative AI models could suffer from "hallucinations" and produce incorrect (personal) information as output. Take measures to limit the impact of inaccuracies in your output data. | |
| Non-compliance | Data Protection / Accountability | You do not have a lawful basis for processing the personal data. | Do you know which GDPR legal ground you can apply?<br>▪ Consent: the individual has given clear consent for you to process their personal data for a specific purpose.<br>▪ Contract: the processing is necessary for a contract you have with the individual, or because they have asked you to take specific steps before entering into a contract.<br>▪ Legal obligation: the processing is necessary for you to comply with the law (not including contractual obligations).<br>▪ Vital interests: the processing is necessary to protect someone's life.<br>▪ Public task: the processing is necessary for you to perform a task in the public interest or for your official functions, and the task or function has a clear basis in law.<br>▪ Legitimate interests: the processing is necessary for your legitimate interests or the legitimate interests of a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the individual which require protection of personal data, in particular where the individual is a child. (This cannot apply if you are a public authority processing data to perform your official tasks.) | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Non-compliance | Data Protection / Privacy / Accountability | The creation or implementation of the generative AI system is not proportional to the intended goal. | Proportionality is a general principle of EU law. It requires you to strike a balance between the means used and the intended aim. In the context of fundamental rights, proportionality is key for any limitation on these rights. | Provider / Business user |
| Non-compliance | Data Protection / Privacy | You cannot comply with the purpose limitation principle. | ▪ Data repurposing is one of the biggest challenges. Can you use the data for a new purpose?<br>▪ Are the datasets that you are using originally collected for a different purpose?<br>▪ Did the original users give consent for only that specific purpose? | Provider / Business user |
| Non-compliance | Data Protection / Privacy | You cannot comply with all the applicable GDPR data subjects' rights. | ▪ Can you implement the right to withdraw consent, the right to object to the processing, and the right to be forgotten into the development of the AI system?<br>▪ Can you provide individuals with access and a way to rectify their data? | Provider / Business user |
| Non-compliance | Data Protection / Privacy | You might not have performed a (correct) Data Protection Impact Assessment (DPIA). | The use of AI is more likely to trigger the requirement for a DPIA, based on criteria in Art. 35 of the GDPR. The GDPR and the EDPB's Guidelines on DPIAs identify both "new technologies" and the type of automated decision-making that produce legal effects or similarly significantly affect persons as likely to result in a "high risk to the rights and freedoms of natural persons". | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Non-compliance | Liability / Accountability | You have not considered your liability risk. | ▪ Have you considered your legal accountability for damages caused by your AI system?<br>▪ Have you identified which parties could be potentially liable (also end-users and third parties)?<br>▪ Have you considered potential risk scenarios and determined their coverage?<br>▪ The use of black box systems makes it hard to attribute responsibility and determine liability, that is why it is important to ensure that AI systems are transparent, explainable, and auditable. This opacity also has a direct impact on the burden of proof when a claimant has to prove the damage caused and its causation.<br>▪ Liability regimes vary worldwide, and it is important to closely monitor the upcoming legislative proposals on liability.<br>▪ You can start managing liability by managing risks and implementing preventive measures to ensure compliance and transparency throughout the AI system lifecycle.<br>▪ In some cases, an insurance can also help you transfer part of the risk. | Provider / Business user |

| Threat Category | Subcategory | Potential Threat Identified | Description & Guiding Questions | Action Owner |
|---|---|---|---|---|
| Non-compliance | Data Protection / Privacy | You use third-party subcontractors that process data from children or other types of vulnerable people. | ▪ If you are processing data of children or other vulnerable groups, remember that all third parties you are dealing with could also be processing their data and in that case, they should comply with regulations.<br>▪ Your own system might be protecting the individuals, but remember to also check third-party libraries, SDKs, and any other third-party tooling you might be using. | Provider / Business user |
| Non-compliance | Copyright / IP | Your dataset has copyright or other legal restrictions. | Can you use the datasets that you need? Or are there any restrictions? This could also apply to libraries and any other proprietary elements you might want to use. | Provider / Business user / End-user |
| Non-compliance | Geolocation / Data Protection | You have geolocation restrictions to implement your generative AI system in other countries. | It could be that usage of your product would not be allowed in certain countries due to certain legal restrictions. | Provider / Business user |
| Non-compliance | Data Protection / Privacy | You cannot comply with the storage limitation principle. | ▪ Do you know how long you need to keep the data (training data, output data, queries/prompts, etc)?<br>▪ Do you need to comply with specific internal, local, national, and/or international retention rules for the storage of data? | Provider / Business user |

# APPENDIX I - HOW TO SELECT THE RIGHT STAKEHOLDERS FOR THE THREAT MODELING SESSION

You can find some suggestions and guidance in this framework which helps to create a meaningful engagement with stakeholders during impact and risk assessment sessions. It is a practical framework that was created to help anyone designing products or services using AI, machine learning, or algorithm-based data analytics to involve their stakeholders in the design and risk assessment process.

https://ecnl.org/publications/framework-meaningful-engagement-human-rights-impact-assessments-ai

# APPENDIX II – ASSESSING RISKS

The assessment of risks can be a complex process, but it is important to learn how to identify potential threats, identify risks and implement measures to mitigate them. This process can also serve as proof of accountability and when done properly, it also shows your responsibility and care towards individuals and society.

**HOW TO ASSESS RISKS**

If you have never assessed risks before, this brief guide can help you start with the process.

When you identify potential threats from the library, you are already identifying potential causes of risks. This is especially the case when you know you are vulnerable to those threats. Vulnerabilities could be caused by different reasons like, for instance, lack of compliance with certain rules. To classify and determine the level of your risks, you will assess what the probability or likelihood is of those threats happening, and what severity or impact they could have on individuals' privacy, their rights, and on society. This evaluation will result in a list of classified risks for you to act on.

**RISK MATRIX**

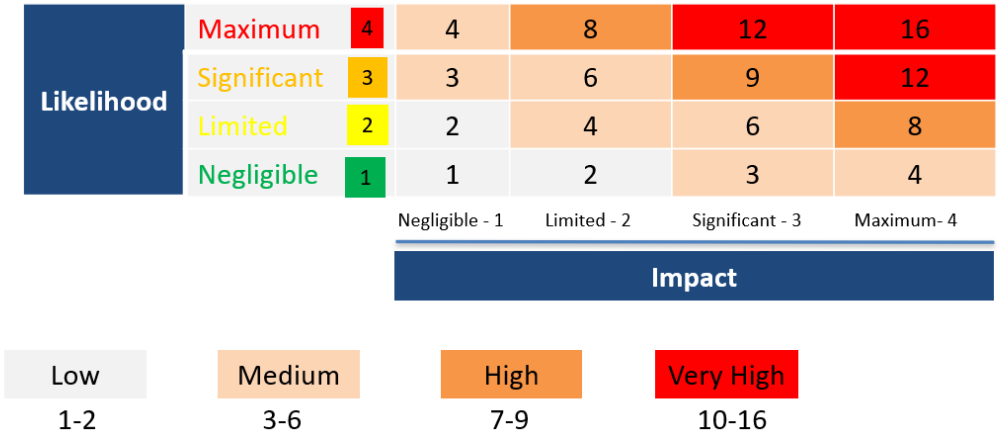
There are several methods and tools that can help you with the risk assessment process. A risk matrix is a tool to assess risks by evaluating the severity of a potential threat, as well as the likelihood of the threat happening.

Here are some examples of a risk matrix:

This is a simple matrix of 3x3; it has 3 levels of likelihood and 3 levels of severity (High / Medium / Low). Because it is simple, it is easy to use but it is also more open to errors in the classification. You can apply it to your identified threats and assign one of the three levels of classification to your risks.

| Severity | | | |
|---|---|---|---|
| | Medium - 3 | High - 6 | High - 9 |
| | Low - 2 | Medium - 4 | High - 6 |
| | Low -1 | Low - 2 | Medium - 3 |
| | | **Likelihood** | |

This is an example of a 4x4 matrix. It is still simple, but it allows for a more nuanced assessment of the risks due to the extra classification level.




These are just a couple of examples of simple practices that can help you start with the risk assessment process.

There are different ways to assess risks and make the classification of risks more accurate. You could for instance consider multiple threat actors and at-risk individuals[2][3], you could also add risk factors like for instance the number of individuals involved or the category of personal data[4], and you could even implement a (quantitative) privacy risk framework[5] or one of the standards[6] available for privacy risk management.

Whatever method you choose, just make sure risks are identified, classified and treated accordingly, and that they do not become part of a forgotten risk register 😉

[2] https://enterprivacy.com/wp-content/uploads/2022/01/Quantitative_Privacy_Risk_Analysis.pdf
[3] https://www.fairinstitute.org/blog/analyzing-privacy-risk-using-fair
[4] https://www.priv.gc.ca/en/privacy-topics/privacy-impact-assessments/gd_exp_202003/
[5] https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/risk-assessment/tools
[6] https://www.iso.org/obp/ui/#iso:std:iso-iec:27557:ed-1:v1:en

# REFERENCES

The information contained in this report is based on public sources that are included in the reference list of PLOT4ai.
You can consult the list of references here:

https://plot4.ai/references

Rhite
Rhite.tech
2023