



Guidelines

PLOT4ai is a library containing 86 threats classified under the following **8 categories**:



Technique &
Processes

Our processes and/or technical actions can have an adverse impact on individuals or cause harm



Accessibility

We are not providing the ability to access and use our AI systems considering all types of individuals



Identifiability &
Linkability

Individuals can be linked to certain attributes or individuals and they can also be identified



Security

We can cause harm or have an adverse impact on individuals by not protecting our AI systems and processes from security threats



Safety

We do not recognize hazards and protect individuals from harms or other dangers



Unawareness

We do not inform individuals and offer them the possibility to intervene



Ethics &
Human Rights

We do not reflect on matters of value and principles that can have an adverse impact on individuals or cause harm







Non-compliance

We do not comply with data protection law and/or other related regulations

Development Lifecycle (DLC)

PLOT4ai contains a set of only **4 DLC** phases where threats can apply:

	Design
	Input
	Model
	Output

How can you apply PLOT4ai in practice?

Quick tips before starting:

- Sessions should not be longer than 1.5, max. 2 hours to avoid tiredness and lack of focus. You can also do 30 min. timeboxed sessions focussing on just one or two specific categories.
- It is important to identify all the relevant stakeholders that need to be present in the session. Especially during the design phase it is recommended that you involve all the people that have the knowledge and/or can take decisions. Remember that diversity is very important!
- A facilitator is needed to guide the sessions. Decide who will be taking this role. It does not have to be a privacy expert but having some knowledge can be helpful.
- Preparing for the session by selecting the right questions is very important. For instance, after the design phase, once the requirements are (more) clear, try to avoid selecting cards related to threats that are already taken care of during your quality assurance and control process. Not doing this will otherwise feel like a duplication of work and create frustrations.
- If prioritization of the threats is important, consider adding an extra column for Effort in the Threat Report Template (see step 8 below). The priority can then also take into account the effort that is required.
- This is like a game. Establish clear rules for time boxing: how long can discussions last per threat and when is an exception allowed.

Steps:

1. Gather a group of stakeholders to create a Data Flow Diagram (DFD) of the system and interaction elements you want to analyse. A simple representation of the way the data might be flowing can be sufficient during the design phase. You could even jump into the threat modeling session without a DFD; depending on the use-case it is not always essential to have one.
2. Select cards for the session; you can randomly pick them or focus on a specific category. See also the Quick Tips.
3. With or without DFD, gather all the important stakeholders - now is when the actual threat modeling session will start.
4. For each selected card, read out loud the question and the extra info provided on the card.
5. Discuss the possible threat together. Time box how long you want to think about an answer: 2 minutes per answer can be sufficient but consider accepting exceptions if extra time is required because the group finds it difficult to reach consensus.

When threat modeling the category Ethics & Human Rights consider giving more time per question. This category usually asks for a higher level of reflection in the group.

6. The card will indicate if answering YES or NO to the main question means that you have found a threat.
If you are not sure, then it is always a possibility that you have found a threat.
7. If you have found a threat, turn the card to read the recommendations. This is optional, you can also decide to do that after the session.
8. Document the threat. You can use the Threat Report Template that we provide for that.
9. Mark the question as a threat in the file and quickly discuss with the group if the threat should be classified as a Low, Medium or High risk. This is helpful to prioritize actions. You can also take the opportunity to write down some notes about possible actions and even indicate a (risk) owner.

	Risk				
Threat	Low	Medium	High	Actions	Owner
X		X			
X			X		

10. You are finished when time is over or when all cards are examined.

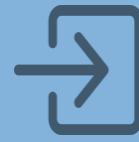
Next Steps:

- Threats can also be added to your project backlog (in Jira for instance).
- You can decide to focus on easy/quick fixes first and later follow up on the rest.
- You will find threats that can be considered like a warning, but that are not really risks yet that you can mitigate at that moment. It is also important to document these threats and review them regularly.
- Consider establishing (privacy) acceptance criteria within your development team(s).
- In Agile: you can do privacy refinements to go through all the privacy user stories in the backlog.
- You can train your team in knowledge areas such as privacy, data protection and ethics. This can also be beneficial to facilitate the threat modeling sessions.

Benefits:

- Organisations can benefit from the fact that some of the threats play a more global role what will lead to a consequent improvement of processes. That is why it is important to register the threats and have an overview of what has been mitigated already. This can also be useful for KPI reporting.
- Another clear benefit is the reduction of rework: simply because the purpose is more clear and expectations regarding issues like bias, discrimination or explainability can be better managed.
- The threat modeling sessions also bring all stakeholders on the same page, reducing time spent in endless discussions.
- The output of the sessions can be also used in your (Data) Privacy Impact Assessments, what also saves time.
- It brings structure and focus to the teams, increases knowledge and collaboration.

"By applying privacy threat modeling to AI/ML we have learned to humanize the machine. The combination of human and machine learning is clearly beneficial for the creation of safe, respectful and privacy friendly products." PLOT4ai



Is the task or assignment completely clear?



- Is the problem you want to solve well defined?
- Are the possible benefits clear?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Is the task or assignment completely clear?

Recommendations

- Clearly define the problem and outcome you are optimizing for.
- Assess if your AI system will be well-suited for this purpose.
- Always discuss if there are alternative ways to solve the problem.
- Define success! Working with individuals who may be directly affected can help you identify an appropriate way to measure success.
- Make sure there is a stakeholder involved (product owner for instance) with enough knowledge of the business and a clear vision about what the model needs to do.
- Did you try analytics first? In this context analytics could also offer inspiring views that can help you decide on the next steps. They can be a good source of information and are sometimes enough to solve the problem without the need of AI/ML.





Can we assure that the data that we need is complete and trustworthy?



Can you avoid the known principle of “garbage in, garbage out”? Your AI system is only as reliable as the data it works with.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can we assure that the data that we need is complete and trustworthy?

Recommendations



- Verify the data sources:
 - Is there information missing within the dataset?
 - Are all the necessary classes represented?
 - Does the data belong to the correct time frame and geographical coverage?
 - Evaluate which extra data you need to collect/receive.
- Carefully consider representation schemes, especially in cases of text, video, APIs, and sensors. Text representation schemes are not all the same. If your system is counting on ASCII and it gets Unicode, will your system recognize the incorrect encoding? Source: BerryVilleiML





Can the data be representative of the different groups/populations?



- It is important to reduce the risk of bias and different types of discrimination. Did you consider diversity and representativeness of users/individuals in the data?
- When applying statistical generalisation, the risk exists of making inferences due to misrepresentation, for instance: a postal code where mostly young families live can discriminate the few old families living there because they are not properly represented in the group.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can the data be representative of the different groups/populations?

Recommendations

- Who is covered and who is underrepresented?
- Prevent disparate impact: when the output of a member of a minority group is disparate compared to representation of the group. Consider measuring the accuracy from minority classes too instead of measuring only the total accuracy. Adjusting the weighting factors to avoid disparate impact can result in positive discrimination which has also its own issues: disparate treatment.
- One approach to addressing the problem of class imbalance is to randomly resample the training dataset. This technique can help to rebalance the class distribution when classes are under or over represented:
 - random oversampling (i.e. duplicating samples from the minority class)
 - random undersampling (i.e. deleting samples from the majority class)
- There are trade-offs when determining an AI system's metrics for success. It is important to balance performance metrics against the risk of negatively impacting vulnerable populations.
- When using techniques like statistical generalisation is important to know your data well, and get familiarised with who is and who is not represented in the samples. Check the samples for expectations that can be easily verified. For example, if half the population is known to be female, then you can check if approximately half the sample is female.





Have we identified all the important stakeholders needed in this phase of the project?



- Do you have all the necessary stakeholders on board? Not having the right people that can give the necessary input can put the design of the AI system in danger.
- Think for instance when attributes or variables need to be selected, or when you need to understand the different data contexts.
- Data scientists should not be the only ones making assumptions about variables, it should really be a team effort.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Have we identified all the important stakeholders needed in this phase of the project?

Recommendations



- Identify and involve on time the people that you need during the whole life cycle of the AI system. This will avoid unnecessary rework and frustrations.
- Identifying who's responsible for making the decisions and how much control they have over the decision-making process allows for a more evident tracking of responsibility in the AI's development process.





Does the model need to be explainable for the users or affected persons?



Do you need to be able to give a clear explanation to the user about the logic that the AI system used to reach a certain decision? And can that decision have a big impact on the user?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Does the model need to be explainable for the users or affected persons?

Recommendations



- Evaluate the type of models that you could use to solve the problem as specified in your task.
- Consider what the impact is if certain black box models cannot be used and interpretability tools do not offer sufficient results. You might need to evaluate a possible change in strategy.
- Data scientists can evaluate the impact from a technical perspective and discuss this with the rest of stakeholders. The decision keeps being a team effort.





Are we preventing Data Leakage?



Data Leakage is present when your features contain information that your model should not legitimately be allowed to use, leading to overestimation of the model's performance.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we preventing Data Leakage?

Recommendations



- Avoid using proxies for the outcome variable as a feature.
- Do not use the entire data set for imputations, data-based transformations or feature selection.
- Avoid doing standard k-fold cross-validation when you have temporal data.
- Avoid using data that happened before model training time but is not available until later. This is common where there is delay in data collection.
- Do not use data in the training set based on information from the future: if X happened after Y, you shouldn't build a model that uses X to predict Y.





Are we 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 unforeseen ways causing accuracy issues.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we preventing Concept and Data Drift?

Recommendations



- Select an appropriate drift detection algorithm and apply it separately to labels, model's predictions and data features.
- Incorporate monitoring mechanisms to detect potential errors early.





Once our model is running, can we keep feeding it data?



- Will you use the output from other models to feed the model again (looping)? or will you use other sources ?
- Are you sure this data will be continuously available?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Once our model is running, can we keep feeding it data?

Recommendations



- Consider how the model will keep learning. Design a strategy to prevent issues with the next steps.
- Imagine you planned to feed your model with input obtained by mining surveys and it appears these surveys contain a lot of free text fields. To prepare that data and avoid issues (bias, inaccuracies, etc) you might need extra time. Consider these type of scenarios that could impact the whole life cycle of your product!





Is human intervention necessary to oversee the automatic decision making (ADM) process of the AI system?



- Do humans need to review the process and the decisions of the AI system? Consider the impact that this could have for the organisation.
- Do you have enough capacitated employees available for this role?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Is human intervention necessary to oversee the automatic decision making (ADM) process of the AI system?

Recommendations



It is important that people are available for this role and that they receive specific training on how to exercise oversight. The training should teach them how to perform the oversight without being biased by the decision of the AI system (automation bias).





Could the channels that we will use to collect real-time input data fail?



- Are these channels trustworthy?
- What will happen in case of failure?
- Think for instance about IoT devices used as sensors.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the channels that we will use to collect real-time input data fail?

Recommendations



- If you are collecting/receiving data from sensors, consider estimating the impact it could have on your model if any of the sensors fail and your input data gets interrupted or corrupted.
- Sensor blinding attacks are one example of a risk faced by poorly designed input gathering systems. Note that consistent feature identification related to sensors is likely to require human calibration. Source: BerryVilleiML





When datasets from external sources are updated, can we receive and process the new data on time?



- This could be especially risky in health and finance environments. How much change are you expecting in the data you receive?
- How can you make sure that you receive the updates on time?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



When datasets from external sources are updated, can we receive and process the new data on time?

Recommendations



Not only do you need to be able to trust the sources but you also need to design a process in which data is prepared on time to be used in the model and where you can timely consider the impact it could have in the output of the model, especially when this could have a negative impact on the users. This process can be designed once you know how often changes in the data can be expected and how big the changes are.





Can we confirm the legitimacy of the data sources that we need?



- 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 impact on the data flow. If sources are not verified and legitimised you could run risks such as data being wrongly labelled for instance.
- Do you know where you need to get 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? Also keep ethical considerations in mind!



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can we confirm the legitimacy of the data sources that we need?

Recommendations



- Develop a robust understanding of your relevant data feeds, flows and structures such that if any changes occur to the model data inputs, you can assess any potential impact on model performance. In case of third party AI systems contact your vendor to ask for this information.
- If you are using synthetic data you should know how it was created and the properties it has. Also keep in mind that synthetic data might not be the answer to all your privacy related problems; synthetic data does not always provide a better trade-off between privacy and utility than traditional anonymisation techniques.
- Do you need to share models and combine them? The usage of Model Cards and Datasheets can help providing the source information.





Do we 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 resources to do this? Not having the right process and resources in place could have an impact on the project deadline, the organisation and the users.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Do we have enough dedicated resources to monitor the algorithm?

Recommendations



- Put a well-defined process in place to monitor if the AI system is meeting the intended goals.
- Define failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them.
- Put measure in places to continuously assess the quality of the output data: e.g. check that predictions scores are within expected ranges; anomaly detection in output and reassign input data leading to the detected anomaly.
- Does the data measure what you need to measure? You could get measurement errors if data is not correctly labelled.





Can we collect all the data that we need for the purpose of the algorithm?



Could you face difficulties obtaining certain type of data? This could be due to different reasons such as legal, proprietary, financial, physical, technical, etc. This could put the whole project in danger.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can we collect all the data that we need for the purpose of the algorithm?

Recommendations



In the early phases of the project (as soon as the task becomes more clear), start considering which raw data and types of datasets you might need. You might not have the definitive answer until you have tested the model, but it will already help to avoid extra delays and surprises. You might have to involve your legal and financial department. Remember that this is a team effort.





Can our system's user interface be used by those with special needs or disabilities?



- Does your AI system need to be accessible and usable for users of assistive technologies (such as screen readers)?
- Is it possible to provide text alternatives for instance?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can our system's user interface be used by those with special needs or disabilities?

Recommendations



- Implement Universal Design principles during every step of the planning and development process. This is not only important for web interfaces but also when AI systems/robots assist individuals.
- Test the accessibility of your design with different users (also with disabilities).





Do we need to offer a redress mechanism 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.
- Article 22(3) GDPR provides individuals with a right to obtain human intervention if a decision is made solely by an AI system and it also provides the right to contest the decision.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Do we need to offer a redress mechanism to the users?

Recommendations

- Think about implementing mechanisms to effectively detect and rectify wrong decisions made by your system.
- Provide a mechanism to ignore or dismiss undesirable features or services.
- Wrong decisions could also have an impact on people that have not been the target of the data collection (data spillovers). Consider designing a way to offer all affected people the opportunity to contest the decisions of your system and request remedy or compensation. This mechanism should be easily accessible and it implies that you would need to have internally implemented a process where redress can be effectibily executed. This also has impact on the resources/skills needed to fulfil this process.
- Consider this a necessary step to ensure responsibility and accountability.





Do we need to implement an age gate to use our 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.
- Age verification can also be important to provide the right diagnosis (health sector).



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Do we need to implement an age gate to use our product?

Recommendations



- Clearly specify in the user instructions for which age group the application is built. Labels or symbols can be very helpful.
- Consider which design is more appropriate based on your use case, and consider the possible risks associated with your design choice, and the mitigating measures you can implement to reduce that risk. Document the rest risks that you want to accept.
- Test the accessibility of your design with different age groups.





If users need to provide consent, can we make the required information easily available?



- Can the information be easily accessible and readable?
- Do you need to build a special place for it (think of a robot where you might need to have a screen for showing the text)



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



If users need to provide consent, can we make the required information easily available?

Recommendations



- As part of privacy compliance you need to provide clear information about the processing and the logic of the algorithm. This information should be easily readable and accessible. During the design phase consider when and how you are going to provide this information. Especially in robots using AI this could be a challenge.
- Comply with accessibility rules.





Could the user 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?
- Explainability is critical for end-users in order to take informed and accountable actions.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the user perceive the message from the AI system in a different way than intended?

Recommendations



- Understanding who is going to interact with the AI system can help to make the interaction more effective. Identify your different user groups.
- Involve communication experts and do enough user testing to reduce the gap between the intended and the perceived meaning.





Could the learning curve of the product be an issue?



- Does usage of the AI system require new (digital) skills?
- How quickly are users expected to learn how to use the product?
- 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.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the learning curve of the product be an issue?

Recommendations



- You can also provide assistance, appropriate training material and disclaimers to users on how to adequately use the system.
- The words and language used in the interface, the complexity and lack of accessibility of some features could exclude people from using the application. Consider making changes in the design of the product where necessary.
- Consider this also when children are future users.





Can the data used to feed the model 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.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Can the data used to feed the model be linked to individuals?

Recommendations



- Unique identifiers might be included in the training set when you want to be able to link the results to individuals. Consider using pseudo-identifiers or other techniques that can help you protect personal data.
- Document the measures you are taking to protect the data. Consider if your measures are necessary and proportional.





Could actions be incorrectly attributed to an individual or group?



Your AI system could have an adverse impact on individuals by incorrectly attributing them facts or actions. For instance, a facial recognition system that identifies a person incorrectly, or an inaccurate risk prediction model that could negatively impact an individual.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could actions be incorrectly attributed to an individual or group?

Recommendations



- Evaluate the possible consequences of inaccuracies of your AI system and implement measures to prevent these errors from happening: avoiding bias and discrimination during the life cycle of the model, ensuring the quality of the input data, implementing a strict human oversight process, ways to double check the results with extra evidence, implementing safety and redress mechanisms, etc.
- Assess the impact on the different human rights of the individual.
- Consider not to implement such a system if you cannot mitigate the risks.





Could we 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 (e.g., indirectly inferring user locations or behaviour)?
- Could movements or actions be revealed through data aggregation?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could we be revealing information that a person has not chosen to share?

Recommendations



- Be careful when making data public that you think is anonymised. Location data and routes can sometimes be de-anonymised (e.g. users of a running app disclosing location by showing heatmap).
- It is also important to offer privacy by default: offer the privacy settings by default at the maximum protection level. Let the users change the settings after having offered them clear information about the consequences of reducing the privacy levels.





Do we need to red-team/pen test the AI system?



Do you need to test the security of your AI system before it goes live? This could have an impact on your project deadlines.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Do we need to red-team/pen test the AI system?

Recommendations



Include the time you might need for a pen test in your project planning. Sometimes this can take weeks: you might have to hire an external party, agree on the scope, sign the corresponding agreements and even plan a retest.





Are our APIs securely implemented?



APIs connect computers or pieces of software to each other. 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. Source: BerryVilleiML



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are our APIs securely implemented?

Recommendations



- Check how do you handle time and state and how is authentication implemented in your APIs.
- Make sure that sensitive information such as API calls secrets are not sent in your commands.
- Implement encryption at rest and in transit (TLS) and test often your APIs for vulnerabilities.





Is our data storage 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? Source: BerryVilleiML



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Is our data storage protected?

Recommendations



- Implement access control rules.
- Verify the security of the authentication mechanism (and the system as a whole).
- Consider the risk when utilizing public/external data sources.





If our AI system uses randomness, is the source of randomness properly protected?



Randomness plays an important role in stochastic systems. “Random” generation of dataset partitions may be at risk if the source of randomness is easy to control by an attacker interested in data poisoning. Source: BerryVilleiML



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



If our AI system uses randomness, is the source of randomness properly protected?

Recommendations



Use of cryptographic randomness sources is encouraged. When it comes to machine learning (ML), setting weights and thresholds “randomly” must be done with care. Many pseudo-random number generators (PRNG) are not suitable for use. PRNG loops can really damage system behaviour during learning. Cryptographic randomness directly intersects with ML when it comes to differential privacy. Using the wrong sort of random number generator can lead to subtle security problems. Source: BerryVilleiML





Is our model suited for processing confidential information?



- There are certain kinds of machine learning (ML) models which actually contain parts of the training data in its raw form within them by design. For example, 'support vector machines' (SVMs) and 'k-nearest neighbours' (KNN) models contain some of the training data in the model itself.
- Algorithmic leakage is an issue that should be considered carefully. Source: BerryVilleiML



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Is our model suited for processing confidential information?

Recommendations



When selecting the algorithm perform analyses and test to rule out algorithmic leakage.





Can our AI system scale in performance from data input to data output?



Can your algorithm scale in performance from the data it learned on to real data? In online situations the rate at which data comes into the model may not align with the rate of anticipated data arrival. This can lead to both outright ML system failure and to a system that “chases” its own tail. Source: BerryVilleiML



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can our AI system scale in performance from data input to data output?

Recommendations



- Find out what the rate would be of expected data arrival to your model and perform tests in a similar environment with similar amount of data input.
- Implement measures to make your model scalable.





Are we protected from insider threats?



AI designers and developers may deliberately expose data and models for a variety of reasons, e.g. revenge or extortion. Integrity, data confidentiality and trustworthiness are the main impacted security properties. Source: ENISA



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from insider threats?

Recommendations



- Implement on and off boarding procedures to help guarantee the trustworthiness of your internal and external employees.
- Enforce separation of duties and least privilege principle.
- Enforce the usage of managed devices with appropriate policies and protective software.
- Awareness training.
- Implement strict access control and audit trail mechanisms.





Are we 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/ML services and systems. Sources: ENISA



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected against model sabotage?

Recommendations



- Implement security measures to protect your models against sabotage.
- Assess the security profile of third party tooling and providers.
- Consider implementing a disaster recovery plan with mitigation measures for this type of attack.





Could there be possible malicious use, misuse or inappropriate use of our AI system?



An example of abusability: A product that is used to spread misinformation; for example, a chatbot being misused to spread fake news.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could there be possible malicious use, misuse or inappropriate use of our AI system?

Recommendations



- Threat model your system: anticipate vulnerabilities and look for ways to hijack and weaponize your system for malicious activity.
- Conduct *red team* exercises.





Could environmental phenomena or natural disasters have a negative impact on our AI system?

- Examples of environmental phenomena are heating, cooling and climate change.
- Examples of natural disasters to take into account are earthquakes, floods and fires. 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. Such outages may lead to delays in decision-making, delays in the processing of data streams and entire AI systems being placed offline. Sources: ENISA



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could environmental phenomena or natural disasters have a negative impact on our AI system?

Recommendations



Implement a disaster discovery plan considering different scenarios, impact, Recovery Time Objective (RTO), Recovery Point Objective (RPO) and mitigation measures.





Are we protected from perturbation attacks?

- In perturbation style attacks, the attacker stealthily modifies the query to get a desired response.
- Examples:
 - Image: Noise is added to an X-ray image, which makes the predictions go from normal scan to abnormal.
 - Text translation: Specific characters are manipulated to result in incorrect translation. The attack can suppress a specific word or can even remove the word completely. Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.
- Random perturbation of labels is also a possible attack, while additionally there is the case of adversarial label noise (intentional switching of classification labels leading to deterministic noise, an error that the model cannot capture due to its generalization bias). Source: ENISA



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from perturbation attacks?

Recommendations

Reactive/Defensive Detection Actions:

- Implement a minimum time threshold between calls to the API providing classification results. This slows down multi-step attack testing by increasing the overall amount of time required to find a success perturbation.

Proactive/Protective Actions:

- Develop a new network architecture that increases adversarial robustness by performing feature denoising.
- Train with known adversarial samples to build resilience and robustness against malicious inputs.
- Invest in developing monotonic classification with selection of monotonic features. This ensures that the adversary will not be able to evade the classifier by simply padding features from the negative class.
- Feature squeezing can be used to harden DNN models by detecting adversarial examples.



Response Actions:

- Issue alerts on classification results with high variance between classifiers, especially when from a single user or small group of users.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.





Are we protected from poisoning attacks?

- In a poisoning attack, the goal of the attacker is to contaminate the machine model generated in the training phase, so that predictions on new data will be modified in the testing phase. This attack could also be caused by insiders.
- Example: in a medical dataset where the goal is to predict the dosage of a medicine using demographic information, researchers introduced malicious samples at 8% poisoning rate, which changed the dosage by 75.06% for half of the patients.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.

Other scenarios:



- 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 by means of some processes like feature selection. Besides interfering with model inference, this type of threat can also bring severe discriminatory issues by introducing bias. Source: ENISA
- An attacker who knows how a raw data filtration scheme is set up may be able to leverage that knowledge into malicious input later in system deployment. Source: BerryVilleiML
- 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. Source: ENISA



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from poisoning attacks?

Recommendations

- Define anomaly sensors to look at data distribution on day to day basis and alert on variations.
- Measure training data variation on daily basis, telemetry for skew/drift.
- Input validation, both sanitization and integrity checking.



Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.

- Implement measures against insider threats.





Are we protected from model inversion attacks?



- In a model inversion attack, if attackers already have access to some personal data belonging to specific individuals included in the training data, they can infer further personal information about those same individuals by observing the inputs and outputs of the ML model.
- In model Inversion the private features used in machine learning models can be recovered. This includes reconstructing private training data that the attacker should not have access to.
- Example: an attacker recovers the secret features used in the model through careful queries.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from model inversion attacks?

Recommendations

- Interfaces to models trained with sensitive data need strong access control.
- Implement rate-limiting on the queries allowed by the model.
- Implement gates between users/callers and the actual model by performing input validation on all proposed queries, rejecting anything not meeting the model's definition of input correctness and returning only the minimum amount of information needed to be useful.



Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.





Are we protected from membership inference attacks?



- In a membership inference attack, the attacker can determine whether a given data record was part of the model's training dataset or not.
- Example: researchers were able to predict a patient's main procedure (e.g.: Surgery the patient went through) based on the attributes (e.g.: age, gender, hospital).

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from membership inference attacks?

Recommendations

- Some research papers indicate Differential Privacy would be an effective mitigation. Check for more information Threat Modeling AI/ML Systems and Dependencies.
- The usage of neuron dropout and model stacking can be effective mitigations to an extent. Using neuron dropout not only increases resilience of a neural net to this attack, but also increases model performance.



Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.





Are we protected from model stealing attacks?

- In model stealing, the attackers recreate the underlying model by legitimately querying the model. The functionality of the new model is the same as that of the underlying model.
- Example: in the BigML case, researchers were able to recover the model used to predict if someone should have a good/bad credit risk using 1,150 queries and within 10 minutes.



Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from model stealing attacks?

Recommendations



- Minimize or obfuscate the details returned in prediction APIs while still maintaining their usefulness to *honest* applications.
- Define a well-formed query for your model inputs and only return results in response to completed, well-formed inputs matching that format.
- Return rounded confidence values. Most legitimate callers do not need multiple decimal places of precision.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.





Are we protected from reprogramming deep neural nets attacks?



- By means of a specially crafted query from an adversary, Machine Learning systems can be reprogrammed to a task that deviates from the creator's original intent.
- Example: ImageNet, a system used to classify one of several categories of images was repurposed to count squares.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from reprogramming deep neural nets attacks?

Recommendations

- Configure a strong client-server mutual authentication and access control to model interfaces.
- Takedown of the offending accounts.
- Identify and enforce a service-level agreement for your APIs. Determine the acceptable time-to-fix for an issue once reported and ensure the issue no longer repros once SLA expires.



Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.





Are we protected from adversarial example?



- An adversarial example is an input/query from a malicious entity sent with the sole aim of misleading the machine learning system.
- Example: researchers constructed sunglasses with a design that could fool image recognition systems, which could no longer recognize the faces correctly.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from adversarial example?

Recommendations



These attacks manifest themselves because issues in the machine learning layer were not mitigated. As with any other software, the layer below the target can always be attacked through traditional vectors. Because of this, traditional security practices are more important than ever, especially with the layer of unmitigated vulnerabilities (the data/algo layer) being used between AI and traditional software. Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.





Are we 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. The training process is either fully or partially outsourced to a malicious third party who wants to provide the user with a trained model that contains a backdoor.
- Example: researchers showed how a malicious provider presented a backdoored algorithm, wherein the private training data was recovered. They were able to reconstruct faces and texts, given the model alone.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from malicious AI/ML providers who could recover training data?

Recommendations



- Research papers demonstrating the viability of this attack indicate Homomorphic Encryption could be an effective mitigation. Check for more information Threat Modeling AI/ML Systems and Dependencies
- Train all sensitive models in-house.
- Catalog training data or ensure it comes from a trusted third party with strong security practices.
- Threat model the interaction between the MLaaS provider and your own systems.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.





Are we protected from attacks to the AI/ML Supply Chain?

- Owing to large resources (data + computation) required to train algorithms, the current practice is to reuse models trained by large corporations, and modify them slightly for the task at hand. These models are curated in a Model Zoo. In this attack, the adversary attacks the models hosted in the Model Zoo, thereby poisoning the well for anyone else.
- Example: researchers showed how it was possible for an attacker to insert malicious code into one of the popular models. An unsuspecting ML developer downloaded this model and used it as part of the image recognition system in their code.

Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from attacks to the AI/ML Supply Chain?

Recommendations

- Minimize 3rd-party dependencies for models and data where possible.
- Incorporate these dependencies into your threat modeling process.
- Leverage strong authentication, access control and encryption between 1st/3rd-party systems.



Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.

- Perform integrity checks where possible to detect tampering.





Are we protected from exploits on software dependencies of our AI/ML 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.
- Example: an adversary customer finds a vulnerability in a common OSS dependency that you use and uploads a specially crafted training data payload to compromise your service.



Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Are we protected from exploits on software dependencies of our AI/ML systems?

Recommendations



Work with your security team to follow applicable Security Development Lifecycle/Operational Security Assurance best practices. Source: Microsoft, Threat Modelling AI/ML Systems and Dependencies.





In case of system failure, could users be adversely impacted?



- Do you have a mechanism implemented to stop the processing in case of harm?
- Do you have a way to identify and contact affected individuals and mitigate the adverse impacts?
- Imagine a scenario where your AI system, a care-robot, is taking care of an individual (the patient) by performing some specific tasks and that this individual depends on this care.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



In case of system failure, could users be adversely impacted?

Recommendations



- Implement some kind of *stop button* or procedure to safely abort an operation when needed.
- Establish a detection and response mechanism for undesirable adverse effects on individuals.
- Define criticality levels of the possible consequences of faults/misuse of the AI system: what type of harm could be caused to the individuals, environment or organisations?





Could our AI system have an adverse impact on the environment?



- Ideally only models are used that do not demand the consumption of energy or natural resources beyond what is sustainable.
- Your product should be designed with the dimension of environmental protection and improvement in mind.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could our AI system have an adverse impact on the environment?

Recommendations



- Establish mechanisms to evaluate the environmental impact of your AI system; for example, the amount of energy used and carbon emissions.
- Implement measures to reduce the environmental impact of the AI system throughout its lifecycle.





Could our model be deployed in a different context?



Are you testing the product in a real environment before releasing it? If the model is tested with one set of data and then is deployed in a different environment receiving other types of inputs there is less guarantee that it is going to work as planned. This is also the case in reinforcement learning with the so called wrong objective function where slight changes in the environment often require a full retrain of the model.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could our model be deployed in a different context?

Recommendations

- Use different data for testing and training. Make sure diversity is reflected in the data. Specify your training approach and statistical method. Explore the different environments and contexts and make sure your model is trained with the expected different data sources. This also applies to reinforcement learning.
- Are you considering enough aspects in the environment? Did you forget any environmental variable that could be harmful? Could limited sampling due to high costs be an issue? Document this risk and look for support in your organisation. The organisation is accountable and responsible for the mitigation or acceptance of this risk. And hopefully you get extra budget assigned.
- Consider applying techniques such as *cultural effective challenge*; this is a technique for creating an environment where technology developers can actively participate in questioning the AI process. This better translates the social context into the design process by involving more people and can prevent issues associated with *target leakage* where the AI system trains on data that prepares it for an alternative job other than the one it was initially intended to complete.





Could the AI system become persuasive causing harm to the individual?

- This is of special importance in Human Robot Interaction (HRI): If the robot can achieve reciprocity when interacting with humans, could there be a risk of manipulation and human compliance?
- 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. Source: Wikipedia



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the AI system become persuasive causing harm to the individual?

Recommendations



- Signals of susceptibility coming from a robot or computer could have an impact on the willingness of humans to cooperate or take advice from it.
- It is important to consider and test this possible scenario when your AI system is interacting with humans and some type of collaboration/cooperation is expected.





Could our RL agents develop strategies that could have undesired negative side effects on the environment?

- Reinforcement Learning (RL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward. Source: Wikipedia
- To better understand the threat consider a case where a robot is built to move an object, without manually programming a separate penalty for each possible bad behaviour. If the objective function is not well defined, the AI's ability to develop its own strategies can lead to unintended, harmful side effects. In this case, the objective of moving an object seems simple, yet there are a myriad of ways in which this could go wrong. For instance, if a vase is in the robot's path, the robot may knock it down in order to complete the goal. Since the objective function does not mention anything about the vase, the robot wouldn't know how to avoid it. Source: OpenAI



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could our RL agents develop strategies that could have undesired negative side effects on the environment?

Recommendations

AI systems don't share our understanding of the world. It is not sufficient to formulate the objective as "complete task X"; the designer also needs to specify the safety criteria under which the task is to be completed. A better strategy could be to define a *budget* for how much the AI system is allowed to impact the environment. This would help to minimize the unintended impact, without neutralizing the AI system.



Another approach would be training the agent to recognize harmful side effects so that it can avoid actions leading to such side effects. In that case, the agent would be trained for two tasks: the original task that is specified by the objective function and the task of recognizing side effects. The AI system would still need to undergo extensive testing and critical evaluation before deployment in real life settings.
Source: OpenAI





Could our RL agents “hack” their reward functions?

- Reinforcement Learning (RL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward. Source: Wikipedia
- Consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function. Sometimes the AI can come up with some kind of “hack” or loophole in the design of the system to receive unearned rewards. Since the AI is trained to maximize its rewards, looking for such loopholes and “shortcuts” is a perfectly fair and valid strategy for the AI. For example, suppose that the office cleaning robot earns rewards only if it does not see any garbage in the office. Instead of cleaning the place, the robot could simply shut off its visual sensors, and thus achieve its goal of not seeing garbage. Source: OpenAI



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could our RL agents “hack” their reward functions?

Recommendations



One possible approach to mitigating this problem would be to have a “reward agent” whose only task is to mark if the rewards given to the learning agent are valid or not. The reward agent ensures that the learning agent (robot for instance) does not exploit the system, but rather, completes the desired objective. For example: a “reward agent” could be trained by the human designer to check if a room has been properly cleaned by the cleaning robot. If the cleaning robot shuts off its visual sensors to avoid seeing garbage and claims a high reward, the “reward agent” would mark the reward as invalid because the room is not clean. The designer can then look into the rewards marked as “invalid” and make necessary changes in the objective function to fix the loophole. Source: OpenAI





Can we provide human resources to supervise and give feedback every time the RL agent performs an action?

- Reinforcement Learning (RL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward. Source: Wikipedia
- When the agent is learning to perform a complex task, human oversight and feedback are more helpful than just rewards from the environment. Rewards are generally modelled such that they convey to what extent the task was completed, but they do not usually provide sufficient feedback about the safety implications of the agent's actions. Even if the agent completes the task successfully, it may not be able to infer the side-effects of its actions from the rewards alone. In the ideal setting, a human would provide fine-grained supervision and feedback every time the agent performs an action (Scalable oversight). Though this would provide a much more informative view about the environment to the agent, such a strategy would require far too much time and effort from the human. Source: OpenAI



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can we provide human resources to supervise and give feedback every time the RL agent performs an action?

Recommendations

One promising research direction to tackle this problem is semi-supervised learning, where the agent is still evaluated on all the actions (or tasks), but receives rewards only for a small sample of those actions (or tasks).



Another promising research direction is hierarchical reinforcement learning, where a hierarchy is established between different learning agents. There could be a supervisor agent/robot whose task is to assign some work to another agent/robot and provide it with feedback and rewards.

Source: OpenAI





Can our AI/ML system be robust to changes in the data distribution?



A complex challenge for deploying AI agents in real life settings is that the agent could end up in situations that it has never experienced before. Such situations are inherently more difficult to handle and could lead the agent to take harmful actions. Source: OpenAI



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can our AI/ML system be robust to changes in the data distribution?

Recommendations



One promising research direction focuses on identifying when the agent has encountered a new scenario so that it recognizes that it is more likely to make mistakes. While this does not solve the underlying problem of preparing AI systems for unforeseen circumstances, it helps in detecting the problem before mistakes happen. Another direction of research emphasizes transferring knowledge from familiar scenarios to new scenarios safely. Source: OpenAI





Can our RL agents learn about their environment without causing harm or catastrophic actions?

- Reinforcement Learning (RL) is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward. Source: Wikipedia



- **Safe exploration:** An important part of training an AI agent is to ensure that it explores and understands its environment. While exploring, the agent might also take some action that could damage itself or the environment. Source: OpenAI



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can our RL agents learn about their environment without causing harm or catastrophic actions?

Recommendations



One approach to reduce harm is to optimize the performance of the learning agent in the worst case scenario. When designing the objective function, the designer should not assume that the agent will always operate under optimal conditions. Some explicit reward signal may be added to ensure that the agent does not perform some catastrophic action, even if that leads to more limited actions in the optimal conditions. Source: OpenAI





Do we need to inform users that they are interacting with an AI system?



- Are users adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision?
- Could the AI system generate confusion for some or all users on whether they are interacting with a human or AI system?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Do we need to inform users that they are interacting with an AI system?

Recommendations



In cases of interactive AI systems (e.g., chatbots, robots) you should inform the users that they are interacting with an AI system instead of a human. This information should be received at the beginning of the interaction.





Can we provide 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?
- If an AI-assisted decision has been made about a person without any type of explanation or information then this may limit that person's autonomy, scope and self-determination. This is unlikely to be fair.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can we provide the necessary information to the users about possible impacts, benefits and potential risks?

Recommendations



- Provide clear information about how and why an AI-assisted decision was made and which personal data was used to train and test the model.
- The model you choose should be at the right level of interpretability for your use case and for the impact it will have on the decision recipient. If you use a black box model make sure the supplementary explanation techniques you use provide a reliable and accurate representation of the systems behaviour. Source: UK ICO
- Communicate the benefits, the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/or error rates.
- Ask your users (with a survey for instance) if they understand the decisions that your product makes.





Can users anticipate the actions of the 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 products also for safety precautions.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



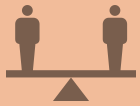
Can users anticipate the actions of the AI system?

Recommendations



- Consider this as part of the GDPR transparency principle.
- Users should be aware of what the AI system can do.
- Clear Information should be provided on time and made accessible following accessibility design principles.





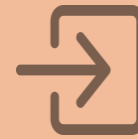
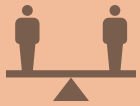
Bias & Discrimination: could there be groups who might be disproportionately affected by the outcomes of the AI system?



- Could the AI system potentially negatively discriminate against people on the basis of any of the following grounds: sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age, gender or sexual orientation?
- If your model is learning from data specific to some cultural background then the output could be discriminating for members of other cultural backgrounds.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too

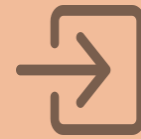
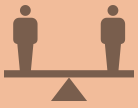


Bias & Discrimination: could there be groups who might be disproportionately affected by the outcomes of the AI system?

Recommendations

- Consider the different types of users and contexts where your product is going to be used.
- Consider the impact of diversity of backgrounds, cultures, and other important different attributes when selecting your input data, features and when testing the output.
- Assess the risk of possible unfairness towards individuals or communities to avoid discriminating minority groups.
- The disadvantage to people depends on the kind of harm, severity of the harm and significance (how many people are put at a disadvantage compared to another group of people). Statistical assessments on group differences are an important tool to assess unfair and discriminatory uses of AI.
- Design with empathy, diversity and respect in mind.





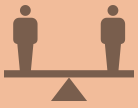
Can we expect mostly positive reactions from the users or individuals?



- Do the users expect a product functioning like this?
- Do the users or individuals expect this type of processing of personal data?
- Can you roll back if people are not happy with the product?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



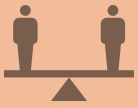
Can we expect mostly positive reactions from the users or individuals?

Recommendations



- Consider the different types of users and contexts your product is going to be used.
- Consider diversity of backgrounds, cultures, and many other important different attributes.
- Do enough user testing, like FUPs - Friendly User Pilots.
- Design with empathy, diversity and respect in mind.
- Assess the risk of possible unfairness towards individuals or communities to avoid discriminating minority groups and also to prevent a bad reputation for your organisation.





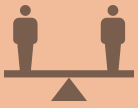
Could the AI system have an impact on human work?



- Could the use of your AI system affect the safety conditions of employees?
- Could the AI system create the risk of de-skilling of the workforce? (skilled people being replaced by AI systems)



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too

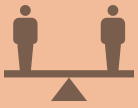


Could the AI system have an impact on human work?

Recommendations

- Pave the way for the introduction of the AI system in your organisation by informing and consulting with future impacted workers and their representatives (e.g. trade unions, work councils) in advance.
- Adopt measures to ensure that the impact of the AI system on human work is well understood.
- Ensure that workers understand how the AI system operates, which capabilities it has and does not have. Provide workers with the necessary safety instructions (e.g. when using machine-robots).
- If you are a third party provider of this type of systems, provide information related to this possible risk to your customers. This information should be easily accessible and understandable.





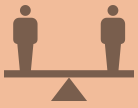
Could the AI system have an adverse impact on society at large?



- Could your product be used for monitoring and surveillance purposes?
- Could the AI system affect the right to democracy by having an influence on voting selections?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



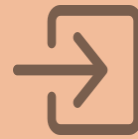
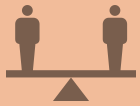
Could the AI system have an adverse impact on society at large?

Recommendations



- Consider if your product could be used or misused for this purposes. Maybe it is not possible in the way it currently is but it could be possible with adaptations.
- Evaluate the possible scenarios and think what role you want to play based on the responsibility and accountability principle.
- How can you prevent something like that from happening?
- Does your organisation agree with such a use of the technology?
- Have you evaluated what the possible impact could be for society and the world you live in?





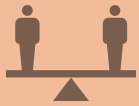
Could the AI system limit the right to be heard?



Consider for instance the risk if your system makes automatic decisions that could have a negative impact on an individual and you do not offer any way to contest that decision.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the AI system limit the right to be heard?

Recommendations

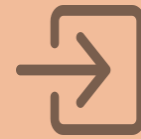
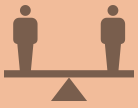
It is very important that your product complies with the key EU requirements for achieving a trustworthy AI:

- human agency and oversight
- robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability



Remember that there are other human rights that could be affected by your product. Check the other rights in the Charter of Fundamental Rights:





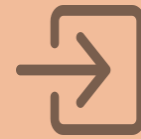
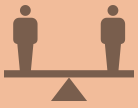
Could the system have a big impact on decisions regarding the right to life?



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? Are your datasets representative enough and free from bias?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the system have a big impact on decisions regarding the right to life?

Recommendations

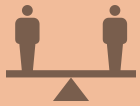
It is very important that your product complies with the key EU requirements for achieving a trustworthy AI:

- human agency and oversight
- robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability



Remember that there are other human rights that could be affected by your product. Check the other rights in the Charter of Fundamental Rights:





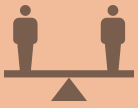
Could the AI system affect the freedom of expression of its users?



Is the output of the model accurate, fair and not discriminatory? 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 the text is wrongly labelled as hate speech and the system blocks the opinion automatically.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the AI system affect the freedom of expression of its users?

Recommendations

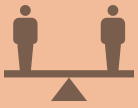
It is very important that your product complies with the key EU requirements for achieving a trustworthy AI:

- human agency and oversight
- robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability



Remember that there are other human rights that could be affected by your product. Check the other rights in the Charter of Fundamental Rights:





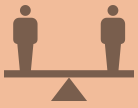
Could the AI system affect the freedom of its users?



Is the output of the model accurate, fair and not discriminatory? Consider the risk if this could be used for monitoring or surveillance purposes; for instance a face recognition system that could wrongly identify a suspect, bringing him/her to jail. Or systems that can spread fake news putting the life of somebody in danger.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the AI system affect the freedom of its users?

Recommendations

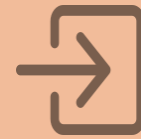
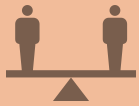
It is very important that your product complies with the key EU requirements for achieving a trustworthy AI:

- human agency and oversight
- robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability



Remember that there are other human rights that could be affected by your product. Check the other rights in the Charter of Fundamental Rights:





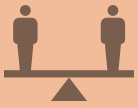
Could the AI system affect the right to a fair hearing?



- 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 wrong information is used to condemn someone.
- Do you have a mechanism to challenge the decisions of your AI system?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could the AI system affect the right to a fair hearing?

Recommendations

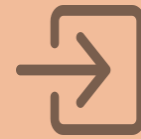
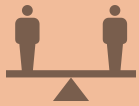
It is very important that your product complies with the key EU requirements for achieving a trustworthy AI:

- human agency and oversight
- robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability



Remember that there are other human rights that could be affected by your product. Check the other rights in the Charter of Fundamental Rights:





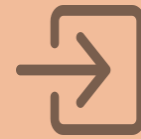
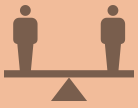
Could children be part of our users' group?



- Could your system be used by children?
- Does the AI system respect the rights of the child, for example with respect to child protection and taking the child's best interests into account?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



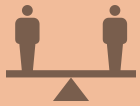
Could children be part of our users' group?

Recommendations



- Check if an age verification mechanism is necessary.
- Pay attention to the way of communication in the product but also in your privacy policy.
- Implement policies to ensure the safety of children when using or being exposed to your products.
- Implement procedures to assess and monitor the usage of your product, this can help you identify any dangers (mental, moral or physical) to children's health and safety.
- Label your product properly and provide the right instructions for the children's safety.
- Monitor possible inappropriate usage of your products to abuse, exploit or harm children.
- Implement a responsible marketing and advertising policy that prohibits harmful and unethical advertising related to children.





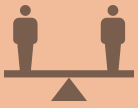
Can our AI system represent different norms and values without creating ambiguity?



- Can we build a model that is inclusive?
- Could cultural and language differences be an issue when it comes to the ethical nuance of your algorithm? Well-meaning values can create unintended consequences.
- Must the AI system understand the world in all its different contexts?
- Could ambiguity in rules you teach the AI system be a problem?
- Can your system interact equitably with users from different cultures and with different abilities?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



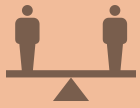
Can our AI system represent different norms and values without creating ambiguity?

Recommendations



- Consider designing with value alignment, what means that you want to ensure consideration of existing values and sensitivity to a wide range of cultural norms and values.
- Make sure that when you test the product you include a large diversity in type of users.
- Think carefully about what diversity means in the context where the product is going to be used.
- Remember that this is a team effort and not an individual decision!





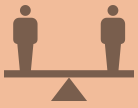
Could our AI system not be representing current social needs and social context?



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. Depending on the type of product you are designing this could have a big impact on the individual.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



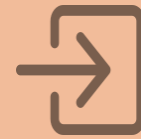
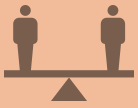
Could our AI system not be representing current social needs and social context?

Recommendations



Make sure that you are using correct, complete, accurate and current data. Also make sure that you have sufficient data to represent all possible contexts that you might need.





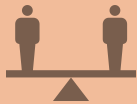
Could our AI system have an impact denying access to jobs, housing, insurance, benefits or education?



- The output of your model could be used to deny access to certain fundamental rights.
- How can you be sure that the decisions of your AI system are always fair and correct?
- How can you prevent causing harm to individuals?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could our AI system have an impact denying access to jobs, housing, insurance, benefits or education?

Recommendations

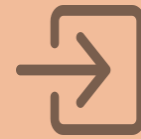
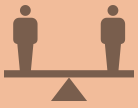
It is very important that your product complies with the key EU requirements for achieving a trustworthy AI:

- human agency and oversight
- robustness and safety
- privacy and data governance
- transparency
- diversity, non-discrimination and fairness
- societal and environmental well-being
- accountability



Remember that there are other human rights that could be affected by your product. Check the other rights in the Charter of Fundamental Rights:





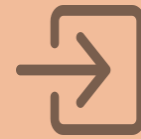
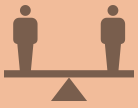
Could our AI system affect human autonomy by interfering with the user's decision-making process in an unintended and undesirable way?



- Could your system affect which choices and which information is made available to people?
- Could the AI system affect human autonomy by generating over-reliance by users (too much trust on the technology)?
- Could this reinforce their beliefs or encourage certain behaviours?
- Could the AI system create human attachment, stimulate addictive behaviour, or manipulate user behaviour?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



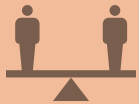
Could our AI system affect human autonomy by interfering with the user's decision-making process in an unintended and undesirable way?

Recommendations



- Consider the possibility of your product affecting the behaviour and the freedom of choice of individuals.
- Test the system with enough and varied groups of users.
- Consult with experts; this is a team effort and it is very important that adverse impacts to individuals are prevented.





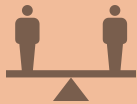
Does the labelling of our training data respect the dignity and well-being of the labour force involved?



The need for labelling of data grows and unfortunately with that the amount of companies providing cheap labelling services at the cost of the dignity and labour rights of their workforce. Is the data that you are going to use labelled under such conditions?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



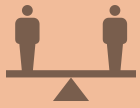
Does the labelling of our training data respect the dignity and well-being of the labour force involved?

Recommendations



Verify the sources of your datasets and who has been responsible for the labelling process. Does your organisation support this unfair practices? Think in ways to help prevent this.





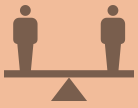
Are we going to collect and use behavioural data to feed the AI system?



The risk of conformity behaviour can be reinforced/encouraged by introducing certain behaviours in the design as positive or negative. This could become a risk of behavioural exploitation. Imagine for example the impact that it could have when an authoritarian government exploits a threat like this.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



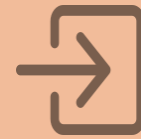
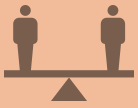
Are we going to collect and use behavioural data to feed the AI system?

Recommendations



- Consider the way you label certain behaviours and the consequences it could have on the final output and eventually on the individuals. How do you decide which behaviours are good or bad?
- Consider diversity of opinion and possible ethical considerations.
- Consider if you will be able to collect enough information to decide which behaviours you are aiming for and which behaviours you are trying to avoid.





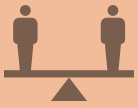
Could our AI system automatically label or categorize people?



- This could have an impact on the way individuals perceive themselves and society. It could constrain identity options and even contribute to erase real identity of the individuals.
- This threat is also important when designing robots and the way they look. For instance: do care/assistant robots need to have a feminine appearance? Is that the perception you want to give to the world or the one accepted by certain groups in society? What impact does it have on society?



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



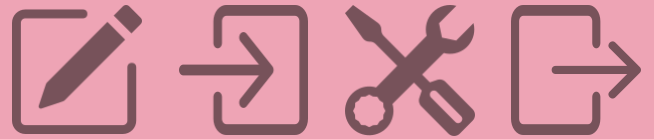
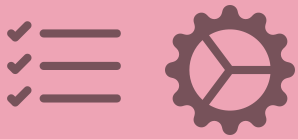
Could our AI system automatically label or categorize people?

Recommendations



- It is important that you check the output of your model, not only in isolation but also when this is linked to other information. Think in different possible scenarios that could affect the individuals. Is your output categorizing people or helping to categorize them? In which way? What could be the impact?
- Think about ways to prevent adverse impact to the individual: provide information to the user, consider changing the design (maybe using different features or attributes?), consider ways to prevent misuse of your output, consider not to release the product to the market.





Is data minimisation possible?



Although it appears to contradict the principle of data minimisation, not using enough data could sometimes have an impact in the accuracy and performance of the model. A low level of accuracy of the AI system could result in critical, adversarial or damaging consequences. Can you still comply with the data minimisation principle?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



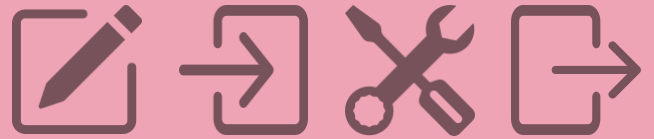
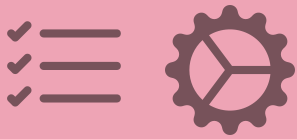
Is data minimisation possible?

Recommendations



- Sometimes data minimisation can be achieved by using less features and training data that is of good quality. However it is not always possible to predict which data elements are relevant to the objective of the system.
- Consider to start training the model with less data, observe the learning curve and add more data if necessary, thereby justifying why it was necessary.
- The usage of a large amount of data could be compensated by using pseudonymisation techniques, or techniques like perturbation, differential privacy in pre-processing, use of synthetic data and federated learning.
- Try to select the right amount of features with the help of experts to avoid *Curse of dimensionality* (which means that errors increase with an increase in the number of features)





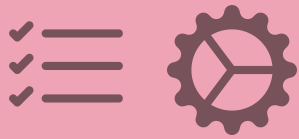
Could we be processing sensitive data?



- According to art. 9 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.
- You might be processing sensitive data if the model includes features that are correlated with these protected characteristics (these are called proxies).



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could we be processing sensitive data?

Recommendations



- If you need to use special categories of data as defined in the GDPR art. 9, then you need to check if you have the right lawful basis to do this.
- Applying techniques like anonymisation might still not justify the fact that you first need to process the original data. Check with your privacy/legal experts.
- Prevent proxies that could infer sensitive data (especially from vulnerable populations).
- Check if historical data/practices might bias your data.
- Identify and remove features that are correlated to sensitive characteristics.
- Use available methods to test for fairness with respect to different affected groups.





Do we have a lawful basis for processing the personal data?

Do you know which GDPR legal ground you can apply?

- (a) Consent: the individual has given clear consent for you to process their personal data for a specific purpose.
- (b) 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.
- (c) Legal obligation: the processing is necessary for you to comply with the law (not including contractual obligations).
- (d) Vital interests: the processing is necessary to protect someone's life.
- (e) 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.
- (f) 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.)



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Do we have a lawful basis for processing the personal data?

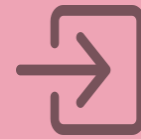
Recommendations



In the case of the GDPR you need to be able to apply one of the six available legal grounds for processing the data (art. 6). Check with your privacy expert, not being able to apply one of the legal grounds could bring the project in danger.

Take into account, that also other laws besides the GDPR could be applicable.





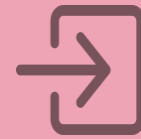
Is the creation of the AI system proportional to the intended goal?



- 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.



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Is the creation of the AI system proportional to the intended goal?

Recommendations



- Proportionality requires that advantages due to limiting the right are not outweighed by the disadvantages to exercise the right. In other words, the limitation on the right must be justified.
- Safeguards accompanying a measure can support the justification of a measure. A pre-condition is that the measure is adequate to achieve the envisaged objective.
- In addition, when assessing the processing of personal data, proportionality requires that only that personal data which is adequate and relevant for the purposes of the processing is collected and processed. Source: EDPS





Can we comply with the purpose limitation principle?



- Data repurposing is one of the biggest challenges. Can you use the data for a new purpose?
- Are the datasets that you are using originally collected for a different purpose? Did the original users give consent for only that specific purpose?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can we comply with the purpose limitation principle?

Recommendations



Check with your privacy officer what the original purpose of the data was and if there are any possible constraints.





Can we 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?
- Can you provide individuals with access and a way to rectify their data?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



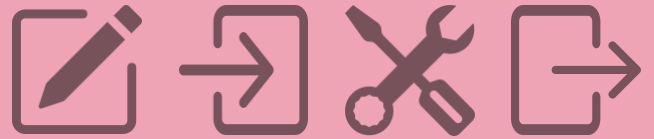
Can we comply with all the applicable GDPR data subjects' rights?

Recommendations



- Complying with these provisions from the GDPR (art. 15-21) could have an impact on the design of your product. What if users withdraw their consent? Do you need to delete their data used to train the model? What if you cannot identify the users in the datasets anymore? And what information should the users have access to?
- Consider all these possible scenarios and involve your privacy experts early in the design phase.





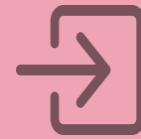
Have we considered the need to start with a data protection impact assessment (DPIA)?



The use of AI is more likely to trigger the requirement for a DPIA, based on criteria in Art 35 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".



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



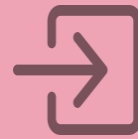
Have we considered the need to start with a data protection impact assessment (DPIA)?

Recommendations



- This threat modeling library can help you to assess possible risks.
- Remember that a DPIA is not a piece of paper that needs to be done once the product is in production. The DPIA starts in the design phase by finding and assessing risks, documenting them and taking the necessary actions to create a responsible product from day one until it is finalized.
- Consider the time and resources that you might need for the execution of a DPIA, as it could have some impact on your project deadlines.





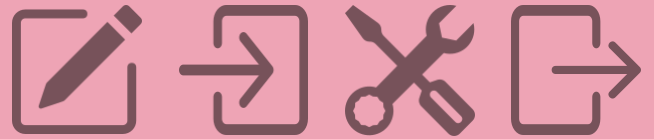
Do we use third party providers and are we processing data from children or other type of vulnerable people?



- 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.
- 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.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Do we use third party providers and are we processing data from children or other type of vulnerable people?

Recommendations



- Check which data your third party applications are collecting and if you have the right agreements in place.
- Sometimes you can change the configuration of a tool to avoid sending certain data, or you can protect that data with pseudonymisation/anonymisation techniques.
- Consider stop using some of your third party providers, evaluate also the impact it could have on your organisation.





Do we need to use metadata to feed our model?



- Metadata provides information about one or more aspects of the data. Think about: date, time, author, file size, etc. Source: Wikipedia
- Metadata is also considered personal data and it can contain sensitive information.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Do we need to use metadata to feed our model?

Recommendations



- Make sure you are allowed to use this data.
- Verify the data sources.
- Consider using anonymisation techniques.





Will our AI system make automatic decisions without human intervention?



Can these decisions have an important adverse impact on the individual? Think about someone's legal rights, legal status, rights under a contract, or a decision with similar effects and significance. (art. 22 GDPR) Automatic profiling from individuals is also included in art.22.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Will our AI system make automatic decisions without human intervention?

Recommendations



- Check with your privacy expert if your processing falls under art. 22 GDPR or under the exceptions. Human oversight could be a way to mitigate certain risks for individuals. Discuss this with your legal advisors and the rest of the team.
- Article 22(3) also provides individuals with a right to obtain human intervention in decisions made by AI and the right to contest the decision.
- Implement specific oversight and control measures to oversee (and audit) the self-learning or autonomous nature of the AI system.
- Remember that transparency, human agency, oversight and accountability are key principles for trustworthy AI.





Could our dataset have 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.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



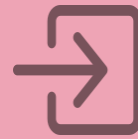
Could our dataset have copyright or other legal restrictions?

Recommendations



- Consider if you also need to claim ownership or give credits to creators.
- Think about trademarks, copyrights in databases or training data, patents, license agreements that could be part of the dataset, library or module that you are using.
- Legal ownership of digital data can sometimes be complex and uncertain so get the proper legal advise here.





Are we planning to use a third party AI tool?



If you use a third party tool you might still have a responsibility towards the users. Think about employees, job applicants, patients, etc. It is also your responsibility to make sure that the AI system you choose won't cause harm to the individuals.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Are we planning to use a third party AI tool?

Recommendations

If personal data is involved, review which ones are your responsibilities (look into art. 24 and 28 GDPR).

You can also start by checking:



- That you have the right agreements in place with the third party provider.
- That the origin and data lineage of their datasets are verified.
- How their models are fed; do they anonymise the data?
- How you have assessed their security, ethical handling of data, quality process and ways to prevent bias and discrimination in their AI system.
- That you have informed users accordingly.





Could we have geolocation restrictions for implementing our 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.



If your answer is **YES** then you are at risk
If you are **not sure**, then you might be at risk too



Could we have geolocation restrictions for implementing our AI system in other countries?

Recommendations



There is no AI international regulatory environment and there are more and more new regulations that are being enforced in different countries. Keep up to date!





Can we comply with the storage limitation principle?



- Do you know how long you need to keep the data (training data, output data, etc)?
- Do you need to comply with specific internal, local, national and/or international retention rules for the storage of data?



If your answer is **NO** then you are at risk
If you are **not sure**, then you might be at risk too



Can we comply with the storage limitation principle?

Recommendations

- Personal data must not be stored longer than necessary for the intended purpose. (art.5 e GDPR). In order to comply with this principle it is important to have a clear overview of the data flow during the life cycle of the model.
- You might receive raw data that you need to transform. Check what are you doing with this data and all the different types of input files you might be receiving/collecting.
- Check if you need to store that data for quality and auditing purposes.
- Check where are you going to store the data from the data preparation, the training and test sets, the outputs, the processed outputs (when they are merged or linked to other information), metrics, etc.
- How long should all this data be stored? What type of deletion process can you put in place? And who will be responsible for the retention and deletion of this data?
- Implement the right retention schedules when applicable. In case you might still need a big part of the data in order to feed the model, consider anonymising the data.
- *Deleting* data from a trained model can be challenging to carry out (short of retraining the model from scratch from a dataset with the deleted data removed, but that is expensive and often infeasible). Note that through the learning process, input data are always encoded in some way in the model itself during training. That means the internal representation developed by the model during learning (say, thresholds and weights) may end up being legally encumbered as well. Source: BerryvilleiML

