



Trustworthy **AI**

7 Steps towards Trustworthy AI

Exercise



Co-funded by the
Erasmus+ Programme
of the European Union

This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.



Copyright holder: Stichting ALLAI Nederland



7 STEPS TOWARDS TRUSTWORTHY AI

Exercise

In this exercise you will learn how to assess the ethical implications of AI, based on real life AI cases - similar to the ones you may face later on in your career - in a practical, problem based and systematic manner. You will increase your problem-solving and analytical skills by analysing the AI use case following 7 steps that integrate the EU Ethics Guidelines for Trustworthy AI as part of the process. This will help you to approach ethical questions surrounding the use of AI in a holistic manner. By doing this you will learn to:

- ◆ Develop an ethical mindset for practical application: Identify, apply and balance ethical, moral and social elements and dilemmas.
- ◆ Apply the ethical mindset to the AI development and use process: Come up with mindful solutions for the problem and subject it to an ethical analysis.
- ◆ Justify your position: Test your solutions in order to convincingly communicate and defend your position.

HOW DOES IT WORK?

Your teacher/instructor has given you or your group a problem or case to analyse. This exercise lets you evaluate and assess the case from the perspective of the Ethics Guidelines for Trustworthy AI, by following 7 steps:

1. **State the problem**
2. **Identify relevant factors and stakeholders**
3. **List solutions**
4. **Assess the solutions against the 7 Requirements for Trustworthy AI**
5. **Test your solutions**
6. **Choose your solution**
7. **Review and Reflect**



STEP 1

State the problem

- ◆ Describe the case
- ◆ Identify and clearly define the problem
- ◆ Check the facts

Briefly summarise the case or problem and ask yourself: “what exactly is the problem that needs solving?” You can do this by asking questions like “Do I see a conflict of interest?” or “Are there parts of this case that make me uncomfortable?”. Some problems disappear when you take a closer look at the situation. Do some (online) research into details that are connected to the problem identified by you. Lastly, also try to look at the problem from a different perspective (for example from a consumer perspective instead of a company perspective, or from a worker perspective. Is it still the same problem? Did the problem get worse?

Example case: Tracing and preventing banking fraud

To help you understand the steps better, every step will have further explanations based on a **short** example case:

Companies like MasterCard or banks have to deal with people trying to commit fraud. Bank fraud entails illegally obtaining money, assets, or other property held by a financial institution. It also includes pretending to be a bank or other institution to con innocent people.

Problem: Banks suffer from people and organisations committing bank fraud. Bank fraud can sometimes be spotted by looking at transaction patterns. However, it is a lot of work to check all transactions and we are not always sure what we are looking for. Can AI help us fight bank fraud by recognising suspicious transaction patterns?

STEP 2

Identify relevant factors and stakeholders

When you identify the relevant stakeholders try to think of all the people involved, including possible organisations or institutional bodies. Other relevant factors that might be important to the problem and its solutions are professional code(s), laws and regulations and other practical constraints.

- ◆ Identify relevant stakeholders, both internally and externally
- ◆ Identify other factors that influence the problem



Example case: Tracing and preventing banking fraud

Relevant stakeholders include: the people or organisations that hold bank accounts, people within banks responsible for security and crime prevention including AI-experts and data-analysts, but also other important people in banks, such as the legal officers, client managers, line management, board members, etc. Also, people from outside of the bank such as banking/financial supervisors or organisations that deal with crime prevention, such as law enforcement.

Relevant factors could include: Regulation (e.g. financial laws and regulations, privacy regulation (GDPR), AI related regulation, fundamental rights, criminal law). Political or societal pressure. Competition. The financial 'bottom line', etc.

STEP 3

List of solutions

Think of 1-3 possible solutions or approaches to the problem. At least one solution should be AI-based (this can also be a pre-determined AI solution that you want to reflect upon). If you list more than one solution, at least one solution should be non-AI. Describe how the solutions would solve the problem and include how your solutions can be realised. In this step you can still be imaginative.

Example case: Tracing and preventing banking fraud

For the sake of this example we have limited the problem to recognising doubtful transaction patterns and pose one AI solution to discuss. This solution consists of a deep learning model that learns to recognize possible fraudulent transaction patterns, based on historical data. The data has to be a realistic mix of normal transactions and fraudulent transactions from both normal citizens and bigger organizations. The model will learn to detect fraudulent transactions, and can then be applied in real-time and automatically block accounts when supposed fraudulent transactions are spotted. We would need high quality data from banks to train this model.

STEP 4

Assess your AI solution(s) against the Ethics Guidelines for Trustworthy AI

In this step you will assess **each of your AI solutions** (do this one by one) against the Ethics Guidelines for Trustworthy AI¹ and **shape, adapt or replace** your solution accordingly. The Guidelines describe 7 requirements that are important to take into account when contemplating the development, deployment or use of AI.

¹ Ethics Guidelines for Trustworthy AI, High-Level Expert Group on AI, 2019: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>



The 7 Requirements for Trustworthy AI²

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability

Not all requirements will be equally relevant to your case and/or AI solution. However, do take a moment to think how a specific requirement can still be relevant, as some connections might not be as obvious as others. You are encouraged to rethink and rephrase your AI solution(s) according to issues that you might find or conclusions you might draw when executing this step.

The following sections contain a quick explanation of each requirement and some questions you could ask yourself in relation to your proposed solutions. You do not need to discuss every question in-depth. Choose the questions that you deem most relevant to your case and AI solution(s). We encourage you to read the which contain a thorough explanation of all requirements.

Requirement #1 - Human agency and oversight

AI systems should support human autonomy and allow them to make informed decisions. In order to achieve this, AI systems should act as enablers to a democratic, equitable society by supporting user's agency, fostering fundamental rights and allowing for human oversight. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation.

◆ Fundamental rights:

- ▶ Can you identify any negative impacts on fundamental rights your solution could have?
- ▶ Can you identify and document potential trade-offs made between the different principles and rights?

◆ Human agency:

- ▶ Does the AI system enhance or augment human capabilities?
- ▶ Is this AI system human-centric: does it leave meaningful opportunity for human choice?
- ▶ Does it enable individuals to have more control over their lives or does it limit their freedom and autonomy?

² <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>



◆ Human oversight:

- ▶ Can you describe the level of human control or involvement in your solution?
- ▶ Have you considered a 'human-in-the-loop' or a 'human-on-the-loop' or a 'human-in-command'?
- ▶ Does the human performing the oversight have the relevant skills, knowledge and authority?

Example case: Tracing and preventing banking fraud

This AI solution could complement human capabilities. It will learn patterns that a human might overlook and will be able to apply its knowledge automatically on a large-scale.

The AI solution however could also limit human autonomy, when for example it leads to 'confirmation bias' with the individual working with the system, or to the 'computer says no' effect.

Fundamental rights affected could involve: right to privacy (if there is unauthorised use of personal data or it affects our freedom to spend our money any which way we like); right to reasonable suspicion (if the fraud detection is based on characteristics someone happens to share with others, rather than an actual suspicion); right to defence (if the model is a black box and it turns out to be impossible to explain the rationale of the decision); right to non-discrimination (if the model turns out to produce biased results).

Our solution lacks human oversight or involvement for now. This might be solved by adding a control layer that manually checks the models decisions for mistakes and human rights impacts. It should also be made possible that certain people are qualified to undo decisions or stop the model from running/checking a certain bank account.

* Reflect: If your solution does not meet this requirement, is there a way to adapt it accordingly?

Requirement #2 - Technical robustness and safety

Technical robustness requires that AI systems are developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in the system's operating environment or the presence of other agents (human and artificial) that may interact with them in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured. This principle ultimately stresses the establishment of a balance of technical robustness and ethical constraints, but also being able to evaluate possible trade offs between the two.

◆ Resilience to attack and security:

- ▶ Can you identify any potential forms of attacks which the AI-system could be vulnerable to?



- ◆ Fallback plan and general safety:
 - ▶ Is there a probable chance that the AI-system may cause damage or harm to users or third parties?
- ◆ Accuracy:
 - ▶ How could the accuracy of the system be measured and ensured?
- * Reflect: If your solution does not meet this requirement, is there a way to adapt it accordingly?

Requirement #3 - Privacy and data governance

Privacy is a fundamental right particularly affected by AI systems, as data is the fuel of a successful AI-system. Oftentimes, this data is about very real people and not taking good care of data has very real effects on individuals. Prevention of harm to privacy must therefore be a priority. This requirement logically necessitates adequate data governance. This covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

- ◆ Respect for privacy and data protection:
 - ▶ Are there ways to develop the AI-system or train the model without or with minimal use of potentially sensitive or personal data?
- ◆ Quality and integrity of data:
 - ▶ Can you think of oversight mechanisms for data collection, storage, processing and use?
- ◆ Access to data:
 - ▶ What protocols, processes and procedures can you think of to manage and ensure proper data governance
 - ▶ Who should be allowed to access users' data and under what circumstances? (Think of qualifications and knowledge/competence to understand the details of data protection policy).

Example case: Tracing and preventing banking fraud

To train this model we have to use real data to find patterns that we can use on new transactions. However, we can take measures and oversight mechanisms that help respect privacy and keep the data safe. All data will be anonymised and encrypted. It will be stored safely and is only accessible upon request by a select group of people with valid reasons to look into it. It will not be immediately deleted after training, in case we need to find errors in the data if the system malfunctions. We however need to be critical of the aspect of 'proxy data', where data that is not personal or sensitive, anonymised personal data, can, in combination with other data, provide a proxy for sensitive insights about persons.

- * Reflect: If your solution does not meet this requirement, is there a way to adapt it accordingly?



Requirement #4 - Transparency

Questions for “Transparency” are very multifaceted. A first question involves recognising transparent systems and how they differ from opaque ones. Second, transparency is about being able to explain a system's decision or reasoning. It also encompasses transparency of elements relevant to AI systems. This includes the data collected, training and workings of the system, and the relevant business models. To realise this, data usage and other decisions taken in the design process should be properly documented and communicated about.

◆ Traceability:

- ▶ What mechanisms could you establish that facilitate the system’s auditability, such as ensuring traceability and logging of its processes and outcomes?
- ▶ Did you review the outcomes of or decisions taken by the system, as well as potential other decisions that would result from different cases (for example, for other subgroups of users)?

◆ Explainability:

- ▶ Can you explain why the system will make a certain choice in a way that is understandable for all users?

◆ Communication:

- ▶ What mechanisms can you put in place to inform (end-)users on the reasons and criteria behind the AI-system’s outcomes?
- ▶ What is the exact purpose of your AI-system and who or what may benefit from it?
- ▶ Can you specify usage scenarios for the system and clearly communicate them to ensure that the system is understandable and appropriate for the intended audience?

✱ Reflect: If your solution does not meet this requirement, is there a way to adapt it accordingly?

Requirement #5 - Diversity, non-discrimination and fairness

We tend to think that the AI-systems are objective and bias-free as they are based on data and logic. Nevertheless, there is no such a thing as unbiased data. The purpose for which the data is collected, who does the measuring, and the decision of what to be measured is always a human choice and thus the AI and data always carry our subjectivity. Hence, in order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system’s life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment.

◆ Unfair bias avoidance:

- ▶ Assess and acknowledge the possible limitations stemming from the composition of the used data sets.
- ▶ Assess whether there could be persons or groups who might be disproportionately affected by negative implications.



- ◆ Accessibility and universal design:
 - ▶ Assess whether the AI system is usable by those with special needs or disabilities or those at risk of exclusion. How can this be designed into the system and how can it be verified?
- ◆ Stakeholder participation:
 - ▶ Can you think of ideas to include the participation of different stakeholders in the AI system's development and use?

Example case: Tracing and preventing banking fraud

We need to make sure that the system does not produce biased outcomes. (Accidental) bias present in the dataset should not be repeated and amplified. To do this we must make sure that specific groups of people are not treated unfairly or discriminatorily against. We must carefully consider the datasets we use to train our model to make sure that past biases are not repeated and amplified. We also must make sure that we do not introduce biased features or indicators during the design process of the model. We must also check the system's outcomes throughout the system's entire lifecycle to make sure its decisions remain unbiased.

The development process should involve relevant stakeholders from inside as well as outside the organisation. These include management, legal officers, business department, front office and assessors, but also the banking authority, fraud experts and the consumer authority.

- * Reflect: If your solution does not meet this requirement, is there a way to adapt it accordingly?

Requirement #6 - Societal and environmental well-being

The well-being of the environment and society as a whole should be considered a "stakeholder" throughout the AI-system's life cycle. This requirement includes the impact on democracy and public discourse and the encouragement of sustainability and ecological responsibility. It involves both research into AI solutions addressing climate change or other societal concerns, as well as being mindful about the ecological footprint of training and deploying an AI system. Interdisciplinarity is an important factor to realise this requirement, as well as conducting impact assessments.

- ◆ Societal impact:
 - ▶ Did you ensure that the social impacts of the AI system are well understood? For example, did you assess whether there is a risk of job loss or deskilling of the workforce? What steps have to be taken to counteract such risks?
- ◆ Society and democracy:
 - ▶ Assess whether the logic of AI might simply and polarise public discourse.
 - ▶ Assess whether the AI-system could be used to manipulate or confuse people.



- ◆ Sustainable and environmental friendly AI:
 - ▶ What mechanisms could you establish to measure the environmental impact of the AI-system's development, deployment and use? (For example, think about the type of energy used by the data centers).
 - ▶ What measures can you think of that can reduce the environmental impact of your AI-system's life cycle?

* Reflect: If your solution does not meet this requirement, is there a way to adapt it accordingly?

Requirement #7 - Accountability

This requirement necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, before, during and after their development, deployment and use. This asks for proper auditing and record keeping, as well as legal frameworks for liability. Developers and deployers of AI-systems should be able to demonstrate the minimisation of negative effects.

- ◆ Auditability:
 - ▶ What mechanisms could you establish that facilitate the system's auditability, such as ensuring traceability and logging of its processes and outcomes?
- ◆ Minimising and reporting negative impact:
 - ▶ Carry out a risk or impact assessment of your AI system, taking into account different stakeholders that are (in)directly affected.
 - ▶ Think of processes that you can establish for third parties (e.g. suppliers, consumers, distributors) or workers to report potential vulnerabilities, risks or biases in the AI-system.
- ◆ Documenting trade-offs:
 - ▶ What are the relevant interests and values impacted by the AI-system?
 - ▶ What are the potential trade-offs between them?
 - ▶ How do you decide on such trade-offs? Document the trade of decision.
- ◆ Ability to seek redress:
 - ▶ What mechanisms can you establish to allow for redress in case of the occurrence of any harm or adverse impact?

Example case: Tracing and preventing banking fraud

We need to think of a mechanism that shows us why a certain transaction pattern is deemed doubtful. The system must have some form of explainability, this will help its auditability. We need to establish a process that integrates a risk assessment and reporting mechanism in case of adverse effects. We need to establish an accountability structure that documents the entire design, development and use process of the system. We need to ensure that people can seek explanation and redress when confronted with an outcome of the system they do not agree with.

STEP 5

Test your AI solution(s)

In this step you will test your AI-solution(s) using different small tests. You can answer the (relevant) questions provided by each test to help you decide whether your solution passes the test.

- ◆ Harm test - Does your solution do less harm than the other solutions?
 - ▶ Does one solution take some criteria better into account than others?
 - ▶ Is it possible to combine the best of different solutions in one solution?
 - ▶ Is your solution necessary to solve the problem and is it limited to solving the problem?
 - ▶ Does your solution pay specific attention to vulnerable groups and ensure that they are not treated with bias?

- ◆ Publicity test - Would I want my solution published in the newspaper?
 - ▶ What sort of questions and concerns from the public would your solution raise?
 - ▶ Does the solution take into account all the relevant stakeholders? How does it advantage or disadvantage some stakeholders over others?
 - ▶ Does the solution have some broader societal impact?
 - ▶ How democratic is the solution? Consider the effect on agency and the power of citizens.

- ◆ Defensibility test - Could you defend choosing this solution to a governmental committee, a committee of my peers or my parents?
 - ▶ Is the solution lawful (legally allowed)?
 - ▶ Does the solution undermine human rights (e.g. life, safety, privacy, non-discrimination, freedom of information, freedom of demonstration, a healthy and safe workplace, fair trial, etc.)?
 - ▶ What is the reasoning behind choosing this solution over others and can I defend that reasoning?

- ◆ Virtue test - How does your solution reflect you?
 - ▶ What kind of beliefs, assumptions, attitudes, and values does your solution reflect?
 - ▶ What kind of beliefs, assumptions, attitudes, and values does the process of selecting your solution reflect?
 - ▶ What kind of values and ideals do you want to promote with your solution?
 - ▶ Was the solution chosen independently or does it serve someone's interests?

- ◆ Professional test - What might an ethics committee say about your solution?
 - ▶ Does it promote the ethics of the field?
 - ▶ What would your peers, classmates, colleagues say about the ethical alignment of your solution?
 - ▶ What would your superior say when you describe the problem and suggest this solution?



Example case: Tracing and preventing banking fraud

For this example we carry out the publicity test: We would immediately be comfortable for our solution to be published in the newspaper. On the one hand it might help prevent crime and make banking safer, but on the other, it will likely concern the public about their privacy, whether they will land on some kind of 'blacklist' and what elements would inform the outcomes of my solution. For example, they might wonder whether they will be judged based on the type of purchases they make, or where they buy things, or at what times, or whether the details of their purchases stay private and how they can avoid or fight wrongful outcomes.

✱ Repeat step 4 and 5 for each of your AI solution.

STEP 6

Choose

In this step you make a tentative choice based on steps 4 and 5 and from your 3 (or more) solutions (both AI and non-AI) choose the solution that is - or has the potential to be - the most Trustworthy. In this step, consider the following:

- ◆ Which of the AI solutions adhere to the most requirements under step 4?
- ◆ Which AI solution passes the most tests of step 5?
- ◆ Which of the non-AI solutions solve the problem?
- ◆ Do the non-AI solutions have less ethical impact than the AI solutions?

STEP 7

Review and reflect

Review your process from steps 1-6. How do you look back at the earlier steps now that you have thought your solutions through? Do you view the problem in a different light? Are there different solutions that you could have thought of? Are there better solutions that do not involve AI-systems or other forms of technology?

Also, think about what could make it less likely that you would have to make such a decision in the future? Are there precautions you can take? Are there ways to gain more support for you when handling this problem? Are there organizational aspects that need to change?

Example case: Tracing and preventing banking fraud

Tracking transaction patterns is quite invasive and might not be necessary if we had better security. It might be more reasonable to find a security solution. We also might need to think of a way to test whether this solution will actually be effective in limiting bank fraud, before considering it. We might compare such testing to an actuarial approach where we accept a certain risk of fraud as part of our business risks.





Trustworthy AI



Co-funded by the
Erasmus+ Programme
of the European Union