Trustworthy AI

# Towards Trustworthy AI

## A Teacher's Guide

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

# Table of Contents

This project has been funded with support from the European Commission. This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use, which may be made of the information contained therein.

Section 1

# Introduction

# Teaching Trustworthy AI

This guide aims to help teachers familiarize themselves with Trustworthy Artificial Intelligence (TAI), based on the EU Ethics Guidelines for Trustworthy AI1 and implement Open Educational Resources on Trustworthy AI into their curricula. The guide explains AI as a technology, its ethical, legal and societal impact, the concept of Trustworthy AI and the importance of teaching students on Trustworthy AI, even if the educational field might seem unrelated at first.

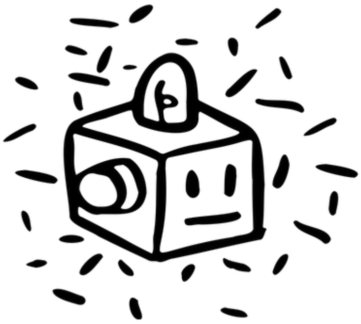
The guide includes three Open Educational Resources dedicated to Trustworthy AI. These resources are meant to help students familiarise themselves with the Ethics Guidelines for Trustworthy AI and develop the ability to understand, critically assess, discuss and apply Trustworthy AI in a practical manner.

The Open Educational Resources include: 8 short knowledge clips, a 42-part card deck, and a 7-step towards Trustworthy AI exercise. This guide will explain how to use these resources, their learning goals, and their importance for students from different fields of education.

In this guide you will come across:

◊ What is AI and what are its societal and ethical implications?
◊ What is Trustworthy AI and how can we realize it?
◊ Why should you teach your students about this and how is this topic relevant to your field?
◊ How to use the different educational resources?
◊ More about AI with extra background material and our glossary

This guide and the Open Educational Resources are aimed at supporting practical education of ethical and societal implications of AI based on the “7 requirements for Trustworthy AI” that are part of the Ethics Guidelines for Trustworthy AI. They go beyond general ethics education, but merely provide a practical approach to AI ethics. As such, the guide and Open Educational Resources aim to help educators become more confident in teaching about Trustworthy AI and stimulate meaningful and ethical discussion between students.



[1] High Level Expert Group on AI: Ethics Guidelines for Trustworthy AI, 2019

# What is AI?

There is no single accepted and rigid definition of AI. AI is a catch-all term for a large number of sub(fields) such as: cognitive computing, machine learning, deep learning, natural language processing, pattern recognition, etc. The central aim of AI research and development is, however, to automate intelligent behaviour such as image recognition, information gathering, planning, communicating, manipulating, detecting and predicting. In 2019, the High Level Expert Group of the European Commission have defined AI as:

“Systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals.”

Much progress has recently been made in the field of so-called narrow AI. Narrow AI is capable of carrying out single specific tasks, contrary to General AI that would be capable of any mental task that can be carried out by a human being. General AI does not exist and scientists disagree on whether it will ever be achieved. In particular with the growth of computer processing power, the availability of large volumes of data and the development of machine learning (ML). ML refers to AI that is based on the processing of "training data" on the basis of which an algorithm learns to recognise patterns and devise rules.

Algorithms can be understood as a collection of instructions that reach a specific outcome. You can compare an algorithm to a recipe for an apple pie: a list of instructions on how to bake an apple pie. Without ingredients and someone following the instructions the recipe will never become a pie. The same goes for algorithms. Without data (ingredients) and a model (someone who follows the recipe), the algorithm will not produce any outcome. And as with a pie the type of ingredients and the skills of the cook can make all the difference in the end result, with AI, the type of data and the decisions made around the purpose of the AI system make all the difference in the end result.

Many AI-systems are software-based, acting in the virtual world as, for example, voice assistants, recommender systems or image analysis software. AI can also be embedded into hardware devices. Think about advanced robots, autonomous cars, drones or the Internet of Things (EU Commission, 2019).

Although AI involves many different techniques, its main focus and strength nowadays lies in techniques surrounding pattern recognition and data analysis. By training the system using large datasets, it gains “knowledge” of patterns that can be found in the data. This knowledge can then be applied to new data that the system has not seen before and a prediction about the unknown data can be made.

# Technical, legal, ethical and societal concerns

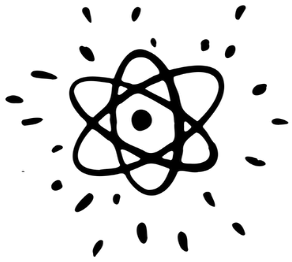
AI's ability to perform tasks in an automated manner, and its potential to be applied in all parts of society makes the technology very attractive, but also brings challenges. Technically AI still has many weaknesses. As one can tell by this method, AI does not have any common sense like humans do. It does not understand the context or the meaning of the data it processes but uses statistical techniques to make its predictions. One needs to train an AI-system with many examples before it can recognize something as simple as a cat. If you proceed to show the system a picture where the cat is upside down, it will likely not recognize it as a cat. The general concept of a cat is not something AI can learn. Further, the performance of an AI system still heavily depends on humans. Humans need to process and label data such that AI can learn from it, resulting in outcomes that reflect how we interpreted the data.

The potential of AI is double sided: it can benefit society as much as it has the potential to negatively affect it. When developed and deployed in a responsible manner, AI can take over dangerous, heavy or monotonous work and increase human efficiency in the work they still do. It can also support us when taking on big issues such as climate change or social injustice. Further, there is a lot of information hidden in all sorts of data that AI can help discover. This will enable us to make informed decisions and keep on innovating. On the downside AI brings complex challenges regarding safety, fairness, privacy, autonomy, accountability and legislation that bring potential harm. With every AI application we design or deploy we must make sure it is safe to use (even in unpredictable situations), inclusive, secure, ethical, guarantees our fundamental rights and much more.

Of course, AI can cause harm when being used for illegal practices. It can be seen as a new tool for committing crimes. However, it also presents risks when used for seemingly innocent purposes. An example of a risk that we likely all experience is the rise of fake news and people being sorted into filter bubbles. Media platforms, such as Facebook, use recommender systems designed to maximise someone's time spent on the platform. As a result of this we are only being shown articles similar to posts that have previously grasped our attention, and which confirm or try to change our beliefs about certain subjects. Because of this, individuals are exposed to a narrower stream of information which will distort our reality and increase the social division of our society even further. These AI systems influence what people see online, which then influences their choices in life. This way, AI poses a threat to our goal of maintaining an open and fair democratic society.

Filter bubbles are the result of commercial organizations wanting to maximize their profit. But even when we think we are using AI for good there are significant risks that we need to be aware of. For example, AI is being deployed more and more for surveillance and monitoring purposes: to prevent crimes and make the world a safer, fairer place. However, (disproportionate) surveillance results in a so-called "chilling effect", where civilians feel like they are constantly being watched and adapt their behaviour accordingly. This undermines, amongst other things, our right to freedom and respect to human dignity.

Another example of accidental harm when using AI to support human results from biased or inaccurate systems. Flaws in a system or bias in a dataset used for training and testing can lead to unfair, discriminating outcomes. Even if the developers did not intend to discriminate, a system might enlarge underlying bias towards certain groups of people present in the data they use. Other potential problematic AI uses are for example biometric recognition, social scoring or the use of autonomous weapons. These AI applications all pose high risks to our safety and fundamental rights.



Section 2

# Trustworthy AI

# Trustworthy AI

AI has the potential to strengthen society and support humans on different fronts. However, we need to carefully implement AI-systems to guard our fundamental rights, health and safety. Developing, deploying and using AI in a human-centric manner can bring us AI that is safe, secure, fair and inclusive. To achieve this, the High-Level Expert Group on Artificial Intelligence developed the Ethics Guidelines for Trustworthy AI[2]. As a basis for these Guidelines the group chose 3 pillars that should underpin AI for it to be trustworthy:

1. it should be lawful, meaning that it should comply with all applicable laws and regulations
2. it should adhere to ethical principles and values, and
3. it should be both technically and socially robust and not cause unintentional harm

For pillars 2 and 3, the group developed **7 requirements for Trustworthy AI** that should guide the development, deployment and use of AI in Europe:

1. Human Agency & 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. Students should be taught about appropriate levels of human agency and autonomy, human control and overall human dignity.

2. Technical Robustness & Safety: This principle requires that AI systems should be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimizing unintentional and unexpected harm and preventing unacceptable harm. Additionally, the physical and mental integrity of humans should be ensured. Students should be taught about how to recognize and ensure accuracy and reliability of AI systems. Further, students should know how to balance technical robustness and ethical constraints.

3. Privacy & Data governance: Privacy is a fundamental right particularly affected by AI systems. Prevention of harm to privacy must 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. Students should be taught about how to collect and recognize high quality data, how to handle it sensitively, maintain privacy and prevent biases in the data and models built from it.

4. Transparency: This requirement has two parts. It encompasses transparency of elements relevant to AI systems. This includes the data collected, training and working of the system, explanations of its outcomes and the relevant business models. It also encompasses the obligation to be transparent about the use of AI systems and not use them covertly. Students should know how to recognize transparent systems, and also be provided with the skills to develop explainable AI. This involves teaching students on how to properly document and communicate about data usage, as well as decisions taken in the design process.

[2] Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence, 2019

5. Diversity, Non-Discrimination & Fairness: The outcomes of AI-systems should be non-discriminatory and free from unacceptable bias. Another important part of this requirement is equal access through inclusive design processes. Inclusion and diversity must be enabled throughout the entire AI-system's life cycle. This includes the consideration and involvement of all affected stakeholders throughout the process Students should be taught about the importance and added value of interdisciplinary expertise when developing, deploying and using AI systems. They should also be taught about the potential discriminatory effects of choices made throughout the development process.

6. Environmental & Societal Well-Being: The environment and society as a whole should be considered a "stakeholder" throughout the AI-system's life cycle. This requirement includes 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. It also involves understanding and mitigating the larger societal, democratic or systemic effects AI can bring.

7. Accountability: This requirement necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use. Students should be taught about auditing and record keeping, as well as legal frameworks for liability and being able to demonstrate minimization of negative effects.

These requirements must be addressed and evaluated continuously throughout the AI-system's entire life cycle - from the design and development phase to the end of usage - considering both technical and non-technical methods to ensure that they are met. By doing this, ethical and robust AI can be fostered and secured.

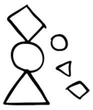
## Why teach Trustworthy AI?

Many students will likely be involved in AI development, deployment, procurement or use in their future occupations. Therefore, it is important that they acquire the necessary skills to develop and implement AI in a responsible and trustworthy manner. Thus, we have transformed the 7 requirements for Trustworthy AI (the "7 Requirements") into specific exercises and educational resources that help teach specific skills including the ability to:

1. Identify the applicability of the 7 Requirements in different contexts and its different dimensions for different stakeholders
2. Deliberate about possible implementations of the 7 requirements
3. Select and implement a course of action in response to ethical analysis regarding the requirements

By acquiring these skills and by developing the ability to understand and act according to ethical and social values, students can ensure a "human-in-command" approach to AI, where it remains up to them to decide if, when and how AI systems should be developed, deployed and used. Since the responsibility for Trustworthy AI does not only fall upon the developers, but on different stakeholders and experts in other fields, teaching students from different domains is of great importance.

## Trustworthy AI for STEM students



It is easy to imagine how students studying STEM subjects (mathematics, biology, physics, chemistry etc.) can easily end up being involved in any stage of AI development. Often discussions arise around AI's unintended consequences, algorithmic bias, data collection and protection etc. Many of these problems arise at the very first stages of the AI life cycle, its development. It is important that STEM students, the 'future AI developers', know how to appreciate ethical values early on and keep apply these values in different stages of AI development. This will enable them to identify unforeseen consequences or correlations or causal relationships between certain development choices and ethical problems, at an early stage. For example, they can ensure that data collection is done properly and responsibly, minimizing the risk of biased or harmful outcomes.

Every AI developer or AI researcher will study one or more STEM subjects at some point in their lives. Thus, as with any scientific researcher, it is important that ethics is well embedded in their educational career. Not only to prevent unforeseen consequences of a system's development, but to acquire reliable knowledge for making informed decisions.

## Trustworthy AI for Business Administration students

Teaching Trustworthy AI in the field of Business Administration is important given the exponential growth of AI in business settings or for commercial use and the fast growth of tech companies. Consumers are easily exposed to AI through e.g. customer service chatbots, predictive/recommending systems make decisions about them or IoT or IoB (Internet of Bodies) devices that use AI for their functioning. Within organizations, recruiting, hiring and worker assessment is done or supported by AI more often. Furthermore, businesses can benefit from the data analytics opportunities that AI can offer in any industry or use AI to manage business functions. For example, AI can be of great benefit for logistics as it can help determine the best organizational system for flights. Likewise, it can help in the healthcare sector by reviewing medical records or treatment approaches and supplement the knowledge of medical practitioners. The increased use of AI in an individual's daily life also makes it attractive for businesses to add AI to their product lineup. Meaning, by integrating and enhancing their product or services with AI features or technology.

Due to AI's wide applicability, it is important that businesses make sure that AI systems they develop, deploy or use is trustworthy. Although AI clearly brings many opportunities, companies should be cautious of risks, ranging from unsafe or unsecure AI to biased or unfair AI outcomes, but also to mass job displacement, or the potential to negatively impact society or democracy. One should not overlook these risks just because of the economic possibilities that AI can offer.

## Trustworthy AI for Political Science students

As we have discussed so far, it is not enough that AI is developed in a trustworthy manner, but it is also crucial that it is deployed responsibly. Unfortunately, we cannot put all of our trust and faith into businesses and developers for this to happen. Policies and legislative frameworks provide the boundaries within which AI can be acceptably developed and used. This is why it is important that political sciences students are educated about Trustworthy AI as well. Assuming that many of these students will become policy makers, or will have a role in governance, it is essential that they understand the true capabilities and limitations of AI and how policy and legislation can set the right boundaries for AI.

Section 3

# Educational Resources

# Educational Resources

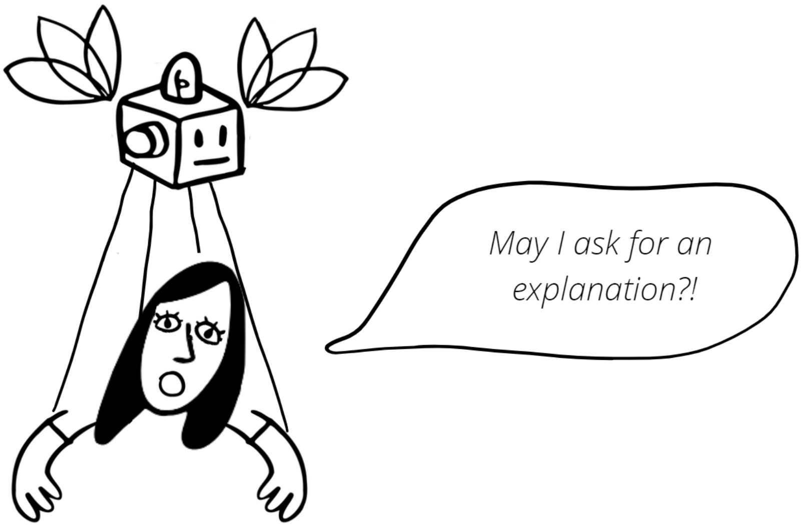
Ethical dilemmas often lead to confusion about how AI must be developed and used in a trustworthy manner. To help students gain knowledge about this, the following open educational resources have been developed.

- ◊ Knowledge clips introducing Trustworthy Ai and the 7 Requirements for Trustworthy AI
- ◊ Trustworthy AI Card Deck
- ◊ 7 Steps towards Trustworthy AI Exercise

# Knowledge Clips

The purpose of providing knowledge clips is to give a brief overview of the Ethical Guidelines of Trustworthy AI and its seven requirements. The clips can be used in a lecture, as an independent education resource, or integrated into an existing course.

The clips consist of 8 videos in total (one introductory video and 7 requirement videos) lasting around 2-4 minutes each. Each requirement video includes an explanation of the requirement and the principles it entails, a real-life example of applying such a requirement, and ends with relevant questions concerning the requirement.



# Trustworthy AI Card Deck

The purpose of using this card deck is to create meaningful ethical discussion in class, and to understand the diversity in which AI can be applied to different domains. Using this card deck, students will understand the complexity of some ethical dilemmas, and its relation to different stakeholders' points of view. The cards provide a general explanation of AI techniques that can be mixed and matched onto different domains to create a use case. A lack of familiarity with AI techniques can lead to certain difficulties with some exercises, which is why we have also created a deck of pre-selected use case cards for ease of use. Furthermore, we are aware that not all AI techniques are present in this deck. AI-techniques and approaches (and their names for that matter) are constantly evolving and new techniques are being developed as we speak. Thus, for now, we simply want to present a set of techniques that is AI-beginner-friendly so it can be used by people with different backgrounds.

Before using the Card Deck, we advise to have students watch the short knowledge clips to gain familiarity with the concept of Trustworthy AI and the 7 Requirements for Trustworthy AI. Nevertheless, any extra knowledge about AI will be useful to come up with use cases required for some of the exercises. You can get a view of general AI techniques in the following link: What is Artificial Intelligence? In 5 minutes

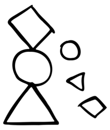
## *Composition of the Card Deck*

### The deck consists of 5 different sets of cards:

With these different sets of cards students can get inspired by use case examples to think of how to use different AI techniques in different domains. By using the requirement cards, they can critically analyze techniques and use cases from the viewpoint of different stakeholders. This can all be done through a series of class exercises/games as explained below.

- ◊ **AI technique:** A generic AI technique that can be used in many domains in various ways
- ◊ **Domain:** A sector in which an AI technique can be applied containing multiple subdomains falling under the same sector
- ◊ **Requirement:** An ethical requirement that evaluates the trustworthiness of an AI technique applied in a specific domain. These cards can also be used to evaluate Use Case cards.
- ◊ **Stakeholder:** A person's role in the development or deployment of an AI system having their own "competing interest" e.g., money, efficiency, safety, fairness, privacy, autonomy, etc.
- ◊ **Use case:** An example of the workings and goals of an AI technique applied to a domain where several stakeholders would be involved.

# Discussion starters



## Exercise 1 - Defend your use case

Materials: domain cards, technique cards, stakeholder cards, requirement cards

*6 players: 1 judge, 5 developers*

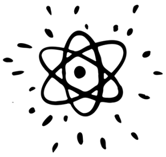
1. A card from the domain deck is drawn. This will be the domain in that round.
2. Each player draws and shows a stakeholder card.
3. Each player draws one technique card without showing.
4. Each player designs a use case for AI (on paper) within the domain from the point of view of their stakeholder role.
5. **NB: for an easier version: let each player picks up a use case card and design the AI system described on that use case**
6. Once all players designed their use case, the judge will draw one requirement card per student.
7. The judge asks the first developer whether their use case complies with the requirement and if so, how.
8. If the answer is acceptable, the developer receives 2 points.
9. If it is not acceptable, or if the student is unable to answer, any of the other developers can argue why their solution complies with the requirement.
10. If this answer is acceptable, the other developer receives 2 points.
11. The judge repeats steps 7-10 for each of the developers.
12. The design with the most points is discussed by all developers to evaluate the results.

## Exercise 2 – Group design and discussion

Materials: Technique card, Domain card, Stakeholder card, Requirement card

*5 players*

1. As a group pick a domain card and a technique card.
2. Together, come up with a use case for that domain and technique.
3. Each player picks up a stakeholder card.
4. Pick up one requirement card at a time and start discussing that requirement from the perspective of your stakeholder.
5. Repeat step 4 for each requirement.
6. Add on: Rank the requirement cards in order of priority from your own stakeholder's point of view and discuss why you chose that rank.



### Exercise 3 - Discuss a use case card

Materials: Use case card, Stakeholder card, Requirement card

*5 players*

1. In small groups pick up a use case card.
2. Each player picks up a stakeholder card.
3. Pick up one requirement card at a time (5 in total) and start discussing that requirement from the perspective of your stakeholder.
4. Add on: Rank the requirement cards in order of priority from you own stakeholder's point of view.

## Stakeholder's views

A Stakeholder is a person's role in the development or deployment of an AI system having their own "competing interest" e.g., money, efficiency, safety, fairness, privacy, autonomy, *etc*. We understand that role-playing some of these stakeholders may be difficult for some, so we provide a few examples for inspiration!

**Governance:** Different governmental officers aim to ensure that AI is transparent, sustainable, and compliant with different ethical and legal requirements. Competing interest: legal compliance

**Authority/supervisor:** Supervisory authorities are independent public authorities that ensure that uniform and consistent application and enforcement of the rules and laws (within their specific field of expertise). Competing interest: safety

**Deployer:** The deployer of an AI system aims to ensure that the system operates at a profit and meets its goals Competing interest: efficiency, money

**Affectee:** Affectees care about their own wellbeing (physical or mental) and the protection of their fundamental human rights. Competing interest: fairness, privacy, autonomy

**Domain expert:** Domain experts aim to inform developers, deployers, and any other stakeholders with correct and reliable information about their field of expertise to ensure that a system works correctly both at a technical and social level. Competing interest: fairness, safety, correct information

**Developer:** Developers aim to develop systems that work correctly with the goals of the deployer in mind. Competing interest: efficiency, safety

# 7-Steps towards Trustworthy AI

The “7 Steps towards Trustworthy AI” functions as a thinking framework that teaches students how to approach case studies in a practical, problem-based and systematic manner. In the exercise students will analyze a problem or case using 7 steps, with the requirements for Trustworthy AI integrated as part of the process. They will learn to identify, apply and balance ethical, moral and social elements and dilemmas relating to AI in general. In addition, the exercise should teach students how to convincingly communicate and defend their positions and design ideas for technical solutions and/or organizational processes.

This exercise is a flexible teaching method that can be applied to any case study and can be integrated in different ways in class; in the form of group-work, as an individual written assignment, class activity or presentation. Analyzing a case like this provides students with a practical way to be “hands-on” in analyzing situations they may face in the workforce. The exercise is meant to enhance students’ problem-solving and analytical skills, as well as improve their abilities to make decisions and deal with ambiguities. Whilst doing this, they will also improve their knowledge of ethics and develop an ethical mindset for practical application. Furthermore, analyzing cases in class or discussion groups can deepen students’ understanding that various, valid perspectives on an issue are possible.

In short, the steps your students will go through are:

1. State the problem
2. Identify relevant factors and stakeholders
3. List AI solutions
4. Teste solutions against the Requirements for Trustworthy AI
5. “Harm-test” the solutions
6. Choose
7. Review and reflect

## How to use the exercise

- ◊ **Choose a form of exercise:** this exercise can be done in multiple forms. Students can take on the cases in the form of an **individual** assessment, but you can also choose to make it a **group** or **class** exercise. Further, you can decide whether you want the students to just discuss the case by going through the steps, making it a “*thinking exercise*”, or whether you want to make it into a *written assessment* or even a *presentation*.

- ◊ **Selecting the case:** Since the exercise is already quite a process in itself, it is preferred that the teacher presents a case to the class (or multiple cases for the students to choose from). The case could be connected to this class’ topic or to the field of study by presenting them with an existing case (for example from the news) or a realistic, simulated one. As an alternative, students could choose their own case. However, this prevents you from steering the exercise. To choose a case that is applicable to the exercise make sure the case:

- Is realistic and relatable enough for the students to be engaged. It should be interesting to your specific students and provide a challenge that they want to tackle.
- Matches the students' skill level or topic knowledge.
- Tells a story with one or more aspects of conflict.
- Contains different relevant stakeholders.
- Does not have an obvious solution and has multiple possible solutions. Even if the case is simple, it should be complex enough to generate group discussion and alternative solutions to the problem posed.
- Is applicable to the requirements.

The cases can be either simple or complex. When looking for a more complex case, think about giving the students lots of information including data charts and graphs about the case or a complex story where the problem is not immediately evident and lots of different factors need to be taken into consideration.

Below you can find some simpler **example cases**. However, we like to encourage you to come up with your own cases as you know your students and their interests best. You can also look at the use cases cards of the Card Deck for inspiration.

**Let them do the exercise.** Depending on the form you chose, students can now work on the exercise individually or, if you choose to make it a class exercise, you can guide them through the steps.