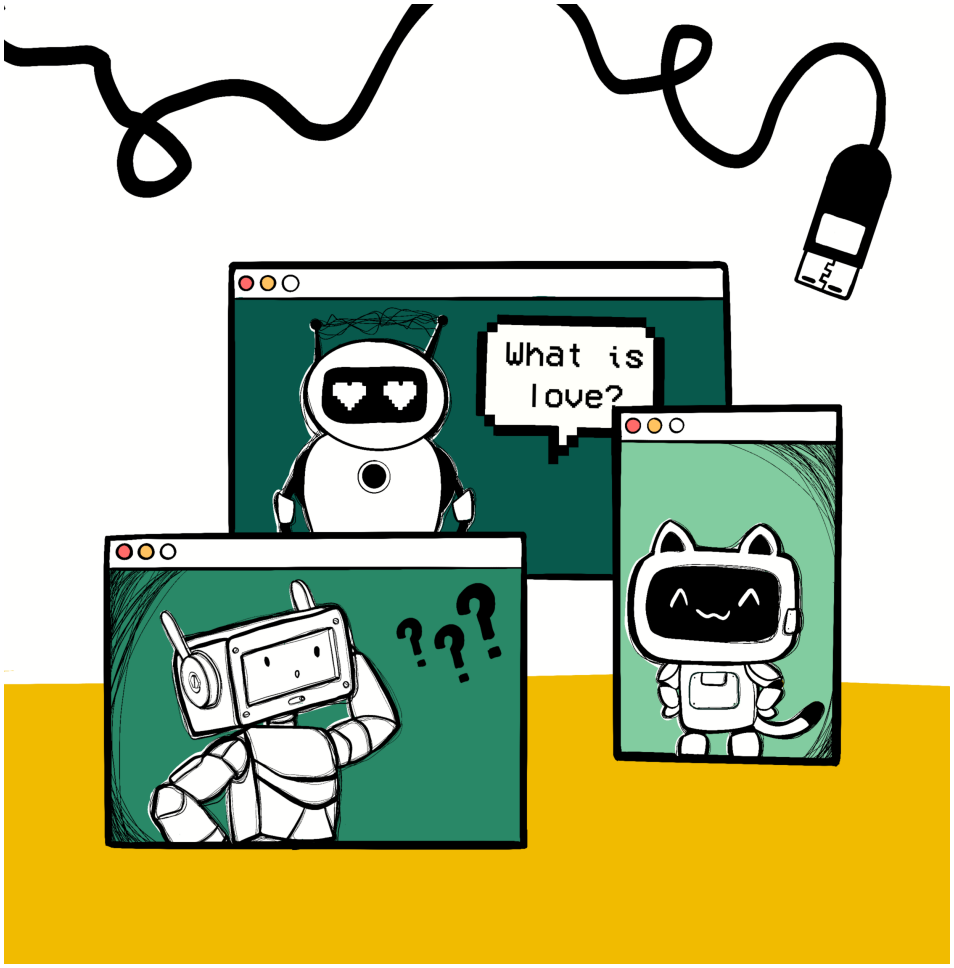
Module 2

Ethics in Artificial Intelligence

"***AI*** is very good at describing the **world** as it is today, with all of its Prejudice. But ***AI*** doesn't know **how the world should be.** "



About The Module

In this module, fundamental ethical aspects in the field of ***AI*** research are to be dealt with. In doing so, the students should, by independently develop and revise catalogs of rules, learn to independently ask critical questions and uncover the tension between individual basic ethical principles. In order to make this complex undertaking as appealing as possible, tasks with different approaches and degrees of complexity were put together, whereby the degree of difficulty can be adjusted by the teachers also. The tasks are thought experiments (like the well-known trolley dilemma, but also exercises for younger learners such as creating rules for a robot butler).

Objectives

The students are able to ...

... argue about the relevance of ethical guidelines in ***AI*** systems

... independently formulate moral rules and revise them with the help of the teacher's input

... find loopholes in self-established rules in group work

... describe the trolley problem and the tension between two contradicting philosophical directions

... form their own philosophical point of view and classify it correctly

... classify central points of the EU ethical guidelines and question them critically

... present personal viewpoints on ***AI*** systems in everyday life (using the example of "autonomous driving")

Agenda

Time	Content
20 min	Exercise - Create Robot Laws
10 min	Theory - Input: What are good ethical rules?
15 min	Exercise and Input: Bias-Errors
20 min	Exercise - MoralMachine
15 min	Theory - The Trolley Problem (Utilitarism vs. Virtue Ethics)
15 min	Theoretical Input Autonomous Driving
15 min	Exercise - Autonomous Driving

Introduction

About this module - Why an Ethical Approach in Artificial Intelligence?

When people think of artificial intelligence, many of them fear that it's about super-intelligent robots that want to subjugate humanity, as in Matrix or the Terminator movies. But ***AI*** research is miles away from such scenarios. However, research today is already dealing with basic ethical questions, because artificial intelligence makes mistakes. One-sided data sets cause bias errors and programs can discriminate against individuals and groups as a result. That is why it is necessary to develop an ethical sensitivity for difficult situations outside of ethics and religion classes, to get to know different perspectives and to acquire skills for decision-making.

Create Own Robot Laws

With the help of the first exercise, a playful, low-threshold approach should be created. In this exercise, a fictitious scenario is described in which a humanoid household robot is able to process all tasks just like a human. The students should think about what possibilities, but also what problems, appear possible when using ***AI*** systems in everyday life.

1. Imagine the following scenario ...

Imagine that you have a robot butler at home. It is a **humanoid (human-like) robot** that is mechanically capable of doing anything a human can do. You are the owner of the robot and can give it commands, e.g. answer questions or do housework. The robot's behavioral abilities are only limited by a number of rules that you have to define.

Slide 2 of the supporting material 1 can be used as supporting help and page 2 of the handout as a writing template.

2. The students invent "golden rules".

The student's first task is to find rules or "laws" that the robot must act in a way that it does not behave in a harmful or unintentional way. It can be advantageous for the following discussions if the students record their results in keywords or as a mind map / poster / slide. This exercise can be done well in small groups, so that the students can exchange ideas in these small groups. Alternatively, the "Think-pair-share" method can also be used. There should be enough time (around 10 to 15 minutes) so that students can really think about individual scenarios to which their rules apply.

On slide 3 in the accompanying material1 you will find some key questions in case the students need further suggestions before they start (or their work stagnates).

3. The students present their rules

Each team should have a few minutes to present their rules and clarify them with simple examples.

4. Look for weak points and loopholes

Take one of the rules presented and show how it can be circumvented / misinterpreted. Slide 4 of the Supplement1 gives some examples of questions.

5. The students should find and present ways in which they would undermine the other rules.

Here, too, the "think-pair-share" method is very suitable, as is working in small groups.

6. Material

- ✏ Ethics - Worksheet Robot Law.pdf

Input for Ethics in AI

How should a moral choice be made?

To help students, the general formulation of moral rules can be spoken of. What do you have to take into account in order for a moral principle to be firmly established?

A moral rule or decision should be:

- **Logical:** The goal should be to support a moral decision with evidence and reasoning rather than relying on feelings or social or personal tendencies. Making logical moral judgments also means ensuring that each of our particular moral judgments is also compatible with our other moral and non-moral beliefs. Discrepancies must be avoided in any case.

  Most philosophers agree that if we make a moral judgment, we must be willing to make the same judgment in similar circumstances. So if we see it as morally wrong, for example, that Ms. Miller has changed numbers in a research to her advantage, it is also morally wrong if our colleague, our spouse or our parents would change numbers. We cannot make an exception for ourselves by picking things for ourselves for which we would judge others for.

- **based on facts:** Before moral decisions can be made, as much relevant information as possible must be gathered. This information should be complete (or at least include all possible data available) and true.
- **Based on solid and defensible moral principles:** Our decisions should be based on generally formulated moral principles that stand up to critical scrutiny and rational criticism. What exactly makes a moral principle sound or acceptable is one of the most difficult questions in ethics.

If we don't want to discuss moral decisions in public, that's often a warning sign. If we don't want to be publicly accountable for our actions, there is a possibility that we do something that we cannot morally justify. Immanuel Kant's philosophical approach can be helpful here ,where we have to apply our moral judgments to each individual person that we can think of (also ourselves). You cannot advocate a moral principle if you don't want to apply it to the general public.

In addition, it can be helpful to look at a problem from another person's point of view in order to avoid moral short-sightedness (for example, one can ask the questions: "How would you like it if ...?")

Isaac Asimovs Robot Laws postulated in 1942 can serve as an aid to determining which rules the students could set up:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

However, there are similar approaches outside of science fiction literature. Many large corporations such as the BMW Group or Microsoft set up ethical principles for artificial intelligence.

There is still no mandatory standard. In view of the rapid development progress, it is not surprising that countless groups of experts are already trying to define general rules.

From the more general task of setting up ethical rules, this exercise will now focus on bias errors in ***AI*** systems, with a student-oriented text about an ***AI***-automated choice of secondary school serving as a basis for discussion.

Bias Errors

This **worksheet** can be discussed together in plenary to make sure that the students understand the problem behind it.

In these two tasks, students should find out **what bias errors are, how they can occur and how they can be avoided.**

Both task 1 and task 2 are suitable for group work, but can also be discussed in plenary. It is advisable to write what has been discussed in keywords so that you can follow up on it. Above all, it should become clear to the students that bias errors are only prejudices inserted into the data sets and not facts. In particular, the resulting discrimination should be highlighted as a serious problem.

The most important aspects of bias errors can be put down in writing at the end. There would also be the option to present the content as a teacher with a PowerPoint presentation or, depending on previous knowledge, working out the content in plenary together with the students.

Material

- ✎ Ethics - Worksheet Bias Error.pdf

Bias Errors and Ethical Guidelines

What are bias errors anyway?

A bias can be a distortion or a predisposition. These arise as soon as the necessary data are collected, which later lead to incorrect results and can lead to serious problems such as exclusion and discrimination.

For example, the Tay chatbot was presented on Twitter in 2016 and had to go offline again within a very short time when the bot began to publish countless discriminatory messages due to the existing database.

What types of bias errors are there?

- **algorithmic *AI* bias or "data bias":** A bias error is committed here due to the data fed in (statistical distortion of the data).
- **societal *AI* bias:** These are norms indoctrinated by society; but stereotypes also create blind spots or prejudices (social prejudices).

Societal bias often influences / creates algorithmic bias errors!

Artificial intelligence is only as good as its underlying database. Programmers are just people who are also stuck in their bubble with their own views and prejudices. This becomes problematic when these prejudices also flow into the programs in the form of selected data sets (**"Garbage in, garbage out."**). This results in new requirements in ***AI*** research. In addition to programming knowledge, a basic ethical understanding is also necessary in order to be able to weigh up whether their implementation helps or harms the common good.

For this reason, the creation of ethical and moral principles is necessary so that no person can be excluded by an artificial intelligence.

A large number of expert groups have already tried to create catalogs of rules: including those of the European Commission, which created the **ethical guidelines for trustworthy *AI*** . In the following, we are guided by four ethical principles, which are intended to serve as the foundation for trustworthy ***AI***s.

Four ethical principles:

Fairness:

If prejudices are built into the data that is fed in, a program does not act fairly. The term "fairness" is not so easy to define: Should all people in a group be treated exactly the same or should different social groups be treated differently in order to eradicate inequalities? Individuals and groups must be protected from **discrimination** and **equal opportunities** (on education, goods, Technology ...) must be guaranteed. In addition, users must not be misled and decisions should be presented transparently.

Respect for human autonomy:

People should still be able to exercise their **self-determination** fully and to insist on their basic rights. ***AI*** systems should empower and encourage people instead of manipulating or subordinating them. It is important that the work processes of the ***AI*** continue to be controlled by **human supervision** .

Harm prevention:

Artificial intelligence must neither cause nor aggravate damage (this includes both mental and physical integrity). The systems must be technically robust so that they are not susceptible to abuse.

It is particularly important to be considerate of those in need of protection. In addition, a main focus should be placed on the unequal distribution of power or information (e.g. government and citizens).

Traceability:

In order to create trust among the users of the ***AI*** system, the processes must be as transparent as possible. So-called **"black box algorithms"** occupy a special position, in which it is not entirely clear how a system comes to the respective result. Particular attention must be paid to traceability and transparent communication talking about the system.

Areas of conflict

Unfortunately, the principles just mentioned cannot always be combined. For example, "damage prevention" and "human autonomy" can conflict if one considers the use of ***AI*** systems in the area of "predictive police work". Special surveillance measures can then help in the fight against crime, but at the same time limit one's own freedom and data protection rights.

Different perspectives must also be taken into account when creating guidelines. Developers, clients and end users all have different points of view, which can even contradict one another.

Bias effects are an important current research topic. Many bias errors can only be mitigated or corrected once the effect is known. By actively dealing with and preventing bias errors, society's trust in artificial intelligence can be strengthened.

How can bias errors be avoided?

1. Self Reflection Do I catch myself thinking in stereotypes? Could it be because of a family or cultural background? Taking different perspectives can help identify distortions.

2. Active Communication Awareness is created through an active exchange and pointing out bias errors

Among other things, the following requirements can be taken into account when creating ethical guidelines in ***AI*** systems:

Fundamental rights, primacy of human action, resistance to attacks, reliability, respect for privacy, quality of data, security of data access, transparency / traceability, open communication, diversity / non-discrimination, participation of stakeholders, sustainability / environmental protection, social impact, verifiability, inclusion of legal Basics (national differences).

Material

- ✎ Ethics - Worksheet Bias Error.pdf

This exercise presents the well-known philosophical thought experiment, the trolley dilemma. This exercise can either be discussed offline as a print version in small groups or there is the possibility of **www.moralmachine.net** go through sceneries online. Also for this exercise, input is provided for discussion. In addition, it is worth making a kind of decision log for this exercise in order to find out how consistently the decisions were made in the group and which factors were decisive for a decision.

The Moral Machine

In this exercise you deal with a philosophical thought experiment in which you have to make serious decisions! In this thought experiment, a car's brakes fail and one has to decide where to hit the car. In most cases you have to choose between several human lives.

1. Choose someone in the group to take note of the decisions you make. Write down **why** you decide so. Do you actively intervene in the action (the moving car)?
2. Open the website **www.moralmachine.net** and click on "Start Judging"
3. Decide on one of the two options by clicking on the corresponding picture. You can get more detailed information about the people on the street by clicking on "Show descriptions".
4. Discuss in the group **why** you decide together on a possibility. Also write down when you were unable to agree immediately.
5. At the end your results will be displayed. What factors seem to be particularly important in your group? Which were less important?
6. Discuss your results in class! How did the others decide?

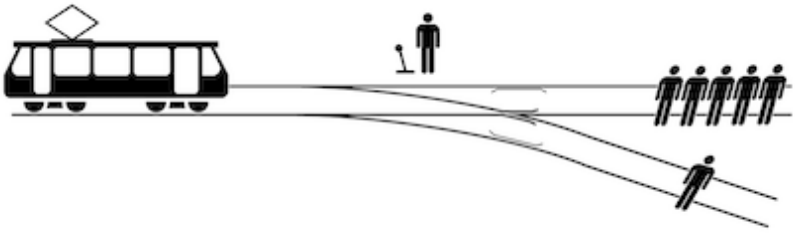
Additive: In Germany, Article 1 of the Basic Law reads: *"The dignity of life is inviolable".* In Austria, the General Civil Code applies: *"Everyone has innate rights, which are already evident through reason, and must therefore be regarded as a person. "*

What could these pieces of law mean in connection with this thought experiment?

Behind The Trolley-Problem

Driver or child? Police officer or homeless? Woman or man? One person has to die, one person can survive a fatal impact accident - just who?

The **Trolley Problem** is a moral-philosophical thought experiment that was first described in the 1930s. With this dilemma, by changing a switch, a tram that has gotten out of control and is rolling towards five people can be steered onto another track. Unfortunately, there is another person on the other platform. Now the moral question arises whether the active intervention of a switchman can lead to the death of one person in order to save the lives of the other five.



From this moral-philosophical dilemma, a multitude of questions can be derived that not only concern theoretical constructs, but also concern very concrete, current problems, such as autonomous driving. Because: How much decision-making power can a machine have? How are ethical decisions programmed for emergencies and who should have the right to decide about human lives (researchers, programmers, industrialists, etc.)? Who is ultimately liable in the event of an accident?

These questions can be discussed in class with the help of **moralmachine.net**. The Moral Machine is part of a research project of the Max Planck Institute (MPI) for Human Development, in which participants have to decide several times between two scenarios. In each scenario, a possible consequence of an inevitable accident is presented and one has to decide on one of the two results. The aim of the research project is to understand individual opinions regarding moral dilemmas that are a matter of life and death. Because of the visualization and presentation of the results, this project is also well suited for school.

While the students work on the individual scenarios in small groups, it is advisable to prepare a kind of decision protocol in which they write down which factors were

decisive in the respective decisions and whether the preferred choice of factors changes over the course of the rounds. Here, as a teacher, you can already pay attention to whether the students decide to actively intervene in the moving car in order to cause as little damage as possible (a more utilitarian approach) or whether they would rather abstain from the decision, because so or so people have to die.

Article 1 of the Basic Law of the Federal Republic of Germany or Section 16 of the General Civil Code from 1811 in Austria can serve as input. In both, human life is defined as inviolable. According to the law, human lives cannot be deviated from, but in human nature it looks quite different. It can be helpful here to suggest that the students put themselves in the shoes of another person. For example, role cards can also be written (elderly lady, a surgeon, a pregnant woman ...) with the help of which the students then have to make decisions similar to the trolley experiment. This makes it easier for the students to empathize with the fact that everyone has the right to life.

At the end of the online version of this thought experiment, the results are visually summarized again in a list on moralmachine.net. Here, too, it can be discussed again why people with certain characteristics are more likely to save than others.

Jeremy Bentham vs. Immanuel Kant

The Moral Machine research team based the two solutions provided on the philosophers Jeremy Bentham and Immanuel Kant. While according to Bentham the car should avoid the greatest possible calamity in a utilitarian manner and drive into the smaller crowd and kill them (including innocent people or driving people), according to Immanuel Kant the car should be based on virtue ethics and thus basic moral principles such as *"You shall not kill"* follow. According to Kant, one should not take an action that explicitly kills a person, but should choose the intended path, even if several people are killed as a result. In previous publications on the Moral Machine, participants tended to select options that were in line with Bentham's philosophy. So one might think that a utilitarian approach that avoids the greatest evil would be most sensible and community-friendly. However, as soon as the participants were asked whether they would buy such a car, they denied, as it could also endanger their own lives. You want cars that protect your own life by all means, but everyone else should buy cars that avoid the greatest possible damage.

Material

- 🌐 (DE) Maschinen im moralischen Dilemma (Wiener Zeitung)
- 🌐 (DE) Über autonomes Fahren nach Bentham und Kant
- 🌐 (DE) Ethische Regeln beim Autonomen Fahren in anderen Kulturen (Wiener Zeitung)
- ✏ Ethics - Worksheet Moral Machine.pdf

About Autonomous Driving

Dilemmas are characterized by the fact that the person concerned can only necessarily choose between two options and there is no other way out.

Solutions to these dilemma situations are repeatedly discussed in science and philosophy. The main question that arises is who is allowed to judge the lives of others or whether individuals are allowed to do so at all. Even if the programmers program in the fundamentally ethically correct decision, it still remains an external decision that does not capture a specific situation with all the given advantages and disadvantages, but can only be derived from an abstract basic idea. As a result, humans would no longer be self-determined but rather determined by others. This resulting consequence is problematic in many ways. On the one hand, the "correct" value images can be specified by the state or company, on the other hand, the human being as an individual would be completely disregarded.

The establishment of general rules such as "personal injury before property damage" may appear helpful in dilemma situations, but factors that could in turn cost the lives of many people are not taken into account in specific cases (e.g. the leakage of a tanker as a result of property damage or the Breakdown of the power grid in a big city). With the establishment of general rules, the problem arises that the diversity and complexity of the various conceivable situations are not covered. For this reason, the ethics committee decided to take solutions that offer the greatest potential for reducing accidents and that are technically feasible.

The protection of human life counts as the greatest good. Programming to minimize the number of victims can be justified if the programming reduces the risk of each individual road user equally. However, that does not mean that you can set off people's lives against those of others. One should not accept the death of one person in an emergency in order to save several, since every being has the right to life. However, the manufacturer is not liable if he does everything reasonable in an automated system to minimize personal injury as much as possible.

The same applies to the self-protection of the fellow travelers: the self-protection of the fellow travelers is not subordinate, however, bystanders may not be sacrificed in favor of the fellow travelers.

Further questions, e.g. regarding the consideration of animals in accidents and the utilization of private data, are answered in the report of the ethics committee for automated and networked driving.

The example of autonomous driving is intended to put what has been learned so far into a more concrete context. In this final exercise, existing proposals for solutions from the car industry and politics should be discussed. The students should see to what extent these correspond to the views of the students and also to the the basic discussed positions.

Autonomous Driving

This worksheet serves as a good basis for discussing the following questions with the students:

- Can a machine make the difference between life and death? Who programs these decisions into the machine (manufacturers, programmers, government ...)?
- Who do you think should have the most say in these decisions? Should consumers be included?
- What role will people still play when driving a car in the future?
- Would you want to drive a fully automated vehicle on the road? Why (not)? In your opinion, what does it take to be able to use autonomous cars without hesitation?
- Are the mentioned guidelines also practicable or could there be problems? In your opinion, what other aspects would be important to include in such guidelines? (see data protection, animals in accidents ...)

Depending on the class, these questions can be discussed in the **plenum** or in **small groups** . In order to present the questions in class later on, **posters or PowerPoint slides** can be created in small groups, depending on the time available.

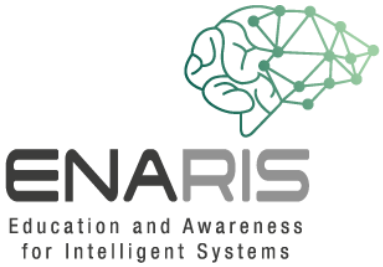
Material

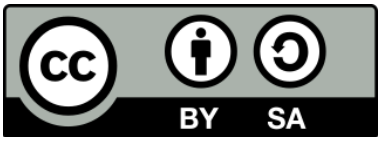
- ✎ Ethics - Worksheet Autonomous Driving.pdf

References

1. (DE) Ethik-Kommission - Automatisiertes und Vernetztes Fahren (ab Seite 14)
2. (DE) Der Mensch im automatisierten Fahrzeug - Digitale Ethik im Alltag

3. (DE) BMK - Automatisiertes Fahren


ENARIS
Education and Awareness
for Intelligent Systems


BY SA


EUROPEAN UNION


Interreg
Austria-Hungary 2014-2020
European Union – European Regional Development Fund


TU
Graz


mobilis
interaktív kiállítási központ


OESTERREICHISCHE
COMPUTER GESELLSCHAFT
AUSTRIAN
COMPUTER SOCIETY