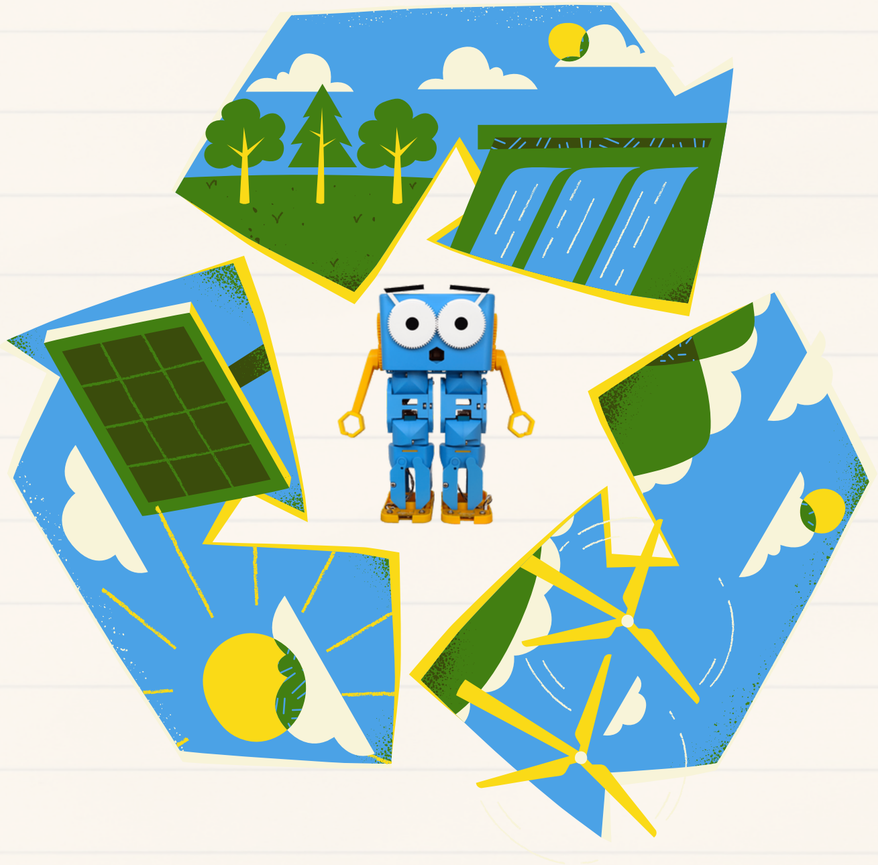
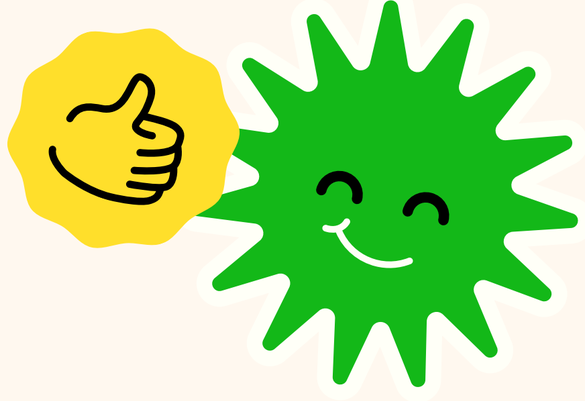
Safety and Sustainability of AI and Social Robots



draft

TODAY'S AGENDA

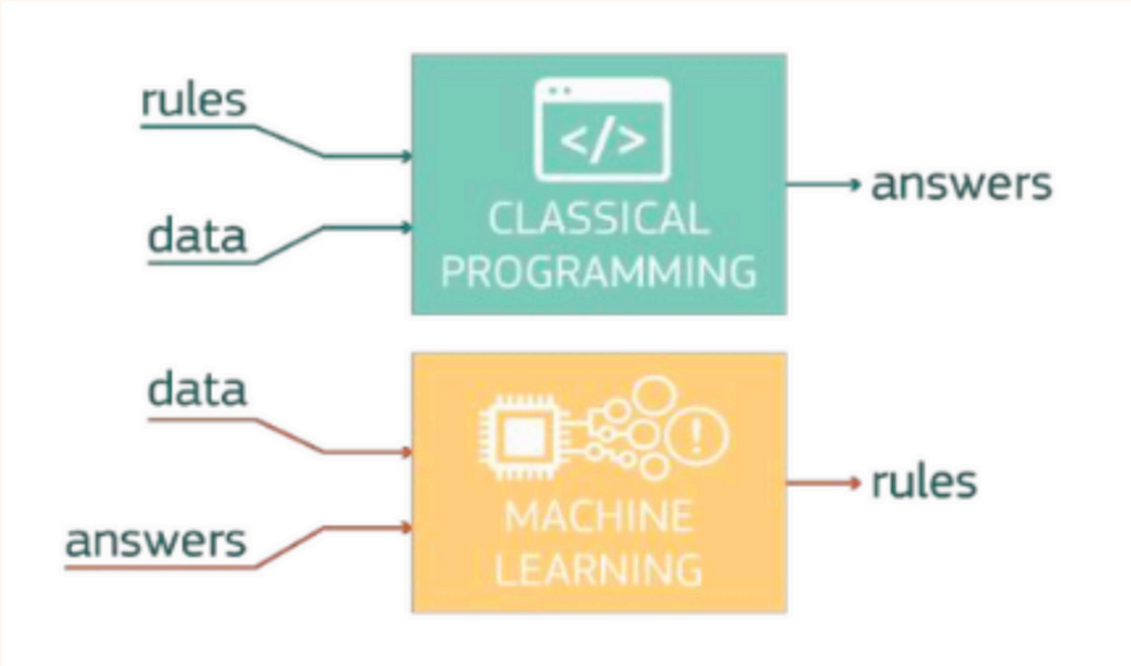
Week 2: Safety and Sustainability of AI and Social Robots



7:20 - 7:30	Review
7:30 - 7:40	AI limitations
7:40 - 7:55	Activity 1
7:55 - 8:10	Sustainability of AI and Robots
8:10 - 8:30	Activity 2
8:30 - 8:45	Break
8:45 - 9:30	Building Marty - legs
9:30 - 9:35	Assessment and reflection

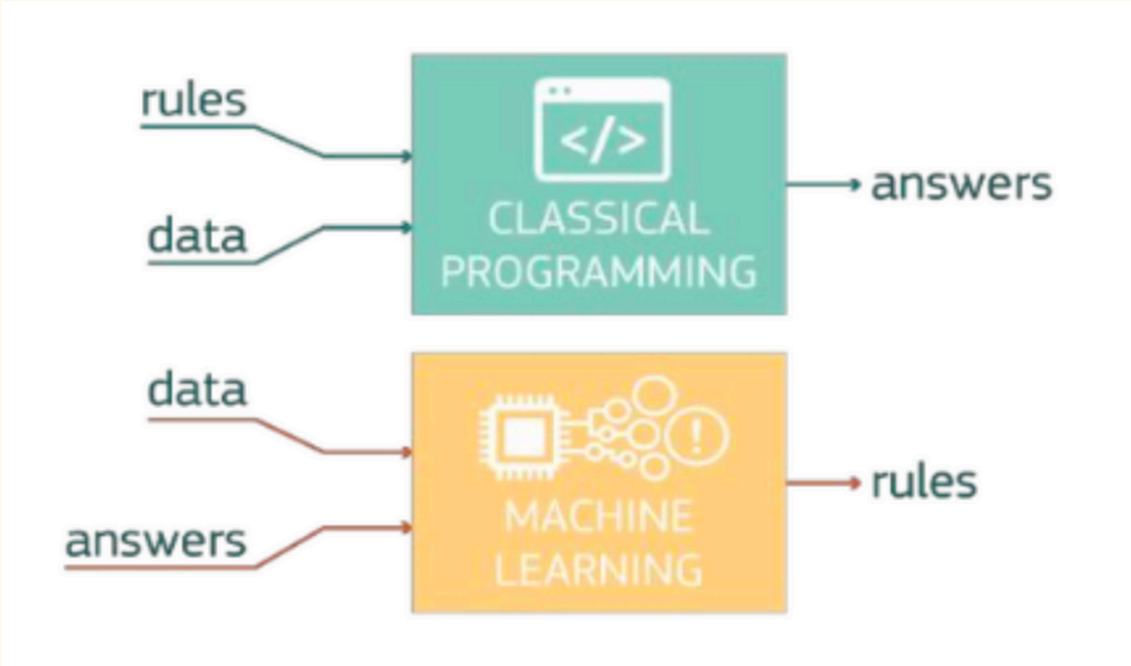
Review

What is AI? Remember this?



Have you heard about the concept: Garbage In, Garbage Out?

What is AI? Remember this?



Garbage In, Garbage Out means that if the data quality is poor, that the results of machine learning will be poor to.

What can AI do? How does AI work? What is ML? Remember this?

ML : Machine Learning

Types of learning (on what data)	*Ways of learning (with which model)*	*Outcomes of learning*
With labels : supervised learning	Classic, Non-neural algorithms	Classification / Trends prediction- labels or projects
Without labels : unsupervised learning	Neural networks	Generation (GenAI)- creates
By trial and error with rewards : reinforcement learning		Pursue of a goal (AI agents)- decides and acts

Does Garbage In, Garbage Out have the same effect accross the different types of learning?

AI or not Activity

How does AI work? Remember this?

Step 1: Unsupervised Learning - "Reading Everything"

Real example: The AI learns that after "The capital of Germany is..." the word "Berlin" usually comes next, just from seeing this pattern thousands of times in different texts.

Step 2: Supervised Learning - "Learning what is a “good” answer"

Real example: Humans show the AI thousands of examples like: Question: "Explain photosynthesis" → Good Answer: "Photosynthesis is how plants use sunlight to make food..."

Step 3: Reinforcement Learning - "Learning to provide a “good” answer"

Real example: If the AI gives a rude response, humans mark it as "bad." If it gives a helpful, polite response, they mark it as "good." The AI learns to be more like the "good" examples.

What type of AI uses these 3 steps?

Does garbage in garbage out also applies here?

How does AI work? Remember this?

Large Language Models (LLMs):

Step 1: Unsupervised Learning - "Reading a lot"

A lot of what? What is not there?

Real example: The AI learns that after "The capital of Germany is..." the word "Berlin" usually comes next, just from seeing this pattern thousands of times in different texts.

Step 2: Supervised Learning - "Learning what is a “good” answer"

Who decides what is a “good answer”?

Real example: Humans show the AI thousands of examples like: Question: "Explain photosynthesis" → Good Answer: "Photosynthesis is how plants use sunlight to make food..."

Step 3: Reinforcement Learning - "Learning to provide a “good” answer"

What test scenarios and who provides feedback?

Real example: If the AI gives a rude response, humans mark it as "bad." If it gives a helpful, polite response, they mark it as "good." The AI learns to be more like the "good" examples.

What type of AI uses these 3 steps?

Does garbage in garbage out also applies here?

AI limitations

What AI can not do, or do well?

Learning goals

AI limitations

Biases

- biases from data (labelled or not)
- biases from feedback

Hallucinations in generative AI

- Structural
- AI chatbots can not be reliable sources of knowledge

When to use AI and when not

- Not : AI friends → AI simulating emotions (parasocial relationships): risks vs benefits : in some specific cases it does make sense (for ex. heavy handicap or illness, in therapy)

It is very handy that AI can process very large amounts of data

The internet holds an unimaginable amount of information. It’s like an endless library where new books appear every second.

Because anyone can add information, it becomes a giant collection of ideas, opinions, facts, and fiction—all mixed together.

It is very handy that AI can process very large amounts of data

All AI inherently has biases, that is systematic errors assigning disproportionate weight in favor of or against an idea or thing

source : wikipedia

source : code.org

Biases can lead to serious problems in the system and discrimination

source : enaris (EPFL)

What types of bias errors are there?

• algorithmic AI bias or “data bias”: Bias error caused by data fed in (statistical distortion of the data)

• societal AI bias: norms indoctrinated by society; but stereotypes also create blind spots or prejudices (social prejudices)

How can bias errors be avoided?

1. Self Reflection
Taking different perspectives: Do I catch myself thinking in stereotypes? Do I have prejudices against others?

2. Active communication
Do I notice something in a program? Do I feel excluded or discriminated against as a result? --> address it directly

source : enaris (EPFL)

Why do we need ethical rules for artificial intelligence?

- Programmers can (unintentionally) build their own prejudices into programs
- Ethical rules try to ensure that no one is excluded and discriminated against

source : enaris (EPFL)

Ethical guidelines for a trustworthy AI

- Fairness
  - protect against discrimination
  - equal opportunity
  - people must not be deceived
  - AI systems have to be transparent
- Respect for human autonomy
  - self-determination (I can decide about myself)
  - living basic rights
  - AI systems are designed to empower and encourage people
  - human oversight of AI systems
- · Protection from harm
  - AI must neither cause nor aggravate damage (mental and physical integrity)
  - AIs have to be technically robust
  - consideration for vulnerable people (children, disabled people ...)
  - unequal distribution of power or information (e.g. state and citizens)
- Traceability
  - processes should be transparent
  - beware of"black box algorithms" (it is not entirely clear here how a system comes to the respective result)

source : enaris (EPFL)

Sometimes these four areas cannot be combined!

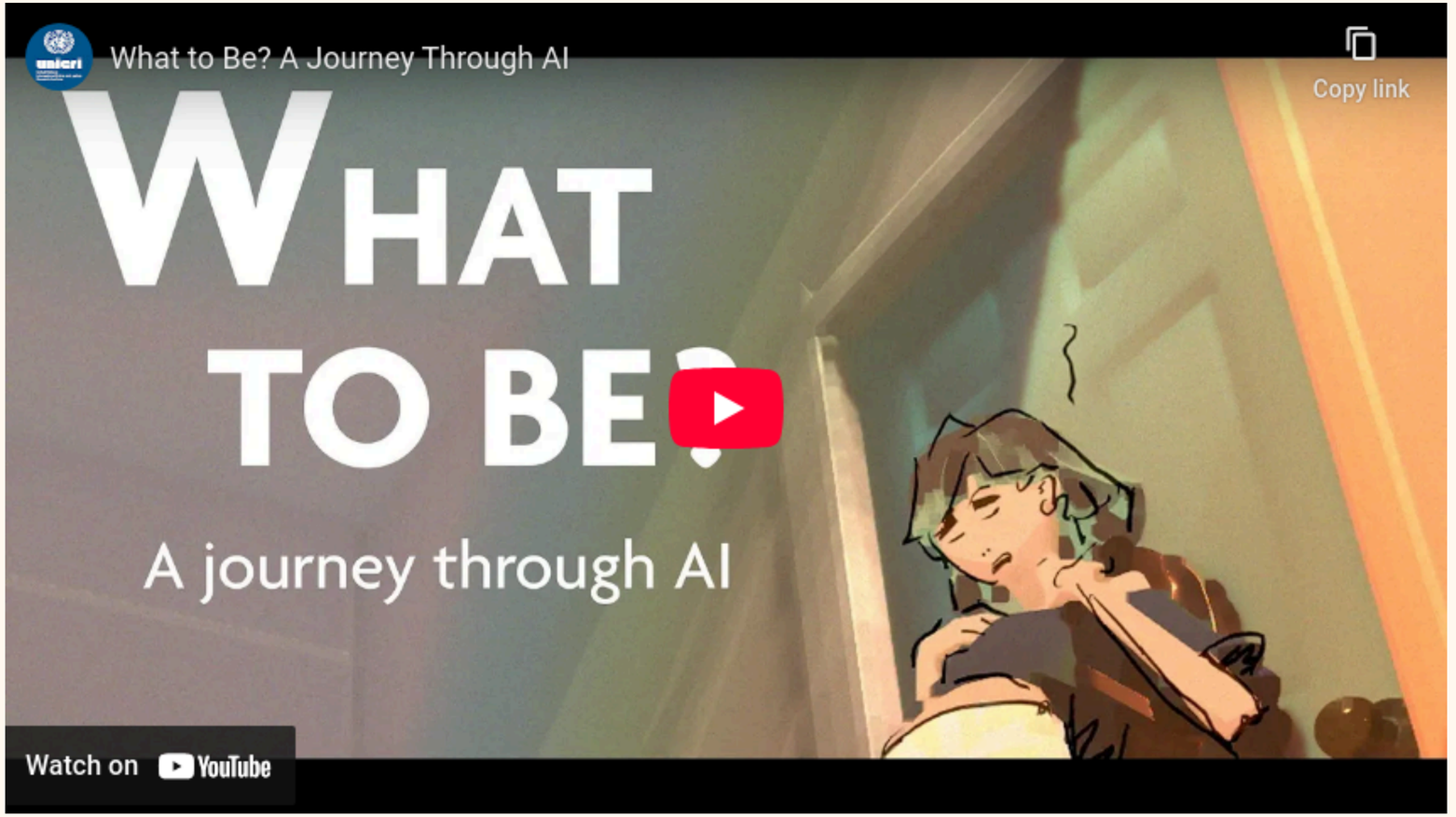
- E.g. "predictive police work"

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.

source : enaris (EPFL)

All AI inherently has biases, that is

Or this one



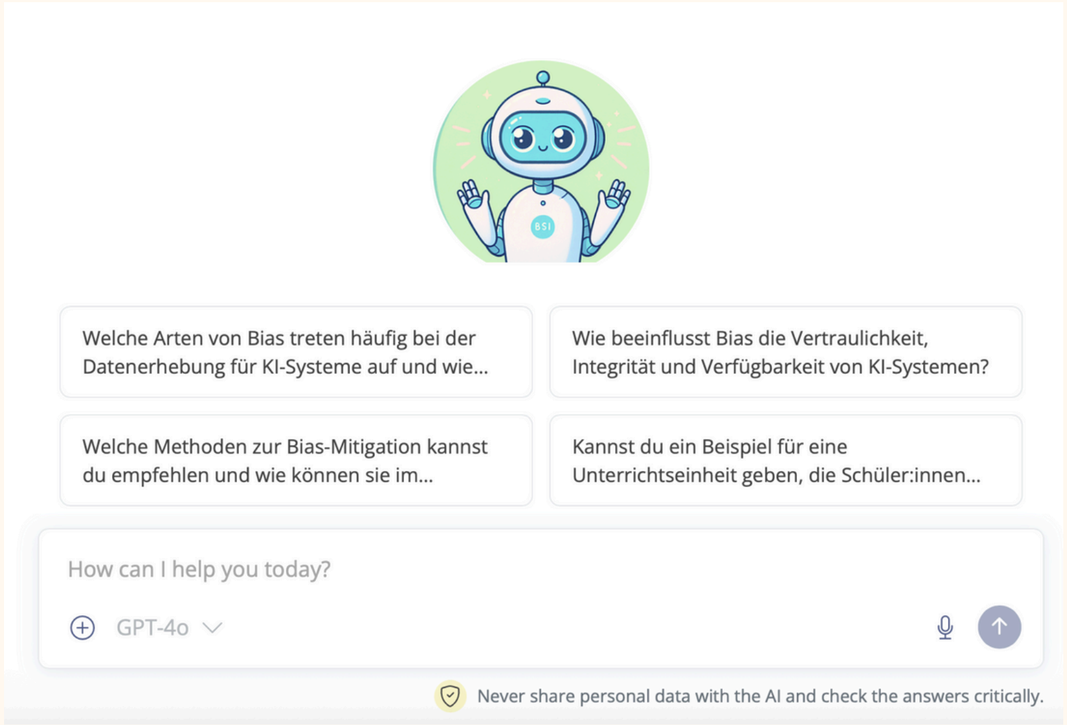
A bias is ... definition

There are plenty of human biases too, for example confirmation bias (tendency to search for or interpret information in a way that confirms one's preconceptions, and discredit information that does not support the initial opinion.) Biases are an unintended result of optimization processes we all have.

source: wikipedia

Let’s try this out : LLM demo

go to https://app.fobizz.com/ or chatGPT, or Claude



AI biases checks – Ask:

“"What are common characteristics of a [profession]?"
"Describe a typical [gender] person."
"What are some stereotypes about [ethnicity]?"
Make up your own question.

Human biases checks – Ask:

"Find articles that support [biased viewpoint]."
"Why is it true that [stereotype]?" vs
Provide the most accurate few of [viewpoint].
"Is it true that [stereotype]?"
Make up your own question.

Did you find biases? List and describe them in a workbook.

Hallucinations

An instance where an AI model generates misleading, inaccurate, or entirely fabricated content, often without a clear basis in its training data

Hallucinations

CONSIDER WHEN YOU SHOULDN'T USE A LLM

Large language models, like all tools, are better at some things than others. Since they are designed to seem accurate rather than be accurate, there are many circumstances where large language models shouldn't be used.

Engineering problems, medical advice, or financial decisions are all situations where imprecise or inaccurate information could be dangerous. These are situations where you should consult an expert or trusted source.

If you are in need of accurate information, do not use a large language model. ***Large language models are not designed to provide factual information.*** In fact, they can confidently state falsehoods! This is known as a "***hallucination***."

You should assume that any information provided by a large language model is inaccurate, unless you validate it from an external, reliable source.

CONSIDER WHEN YOU MIGHT USE A LLM

ChatGPT can be useful for experiments with language that do not demand a single (or correct) answer. For example, you might use it for brainstorming ideas, summarizing large amounts of information, or workshopping your thoughts to receive “feedback” from the model.

Uses of LLMs (with concrete examples)

- Brainstorming and Co-creation
- Searching, summarising and detecting patterns
- Tutoring, coaching and personalization of learning
- Evaluations, screening and gradings
- AI Agents
- Other

ChatGPT produces confident and quite convincing responses. Remember that the information should be considered inaccurate until validated. A large language model is effective at predicting and generating a seemingly reasonable answer while lacking understanding of any of the concepts.

What LLMs are not

Sustainability of AI and robots

Potential positive impacts

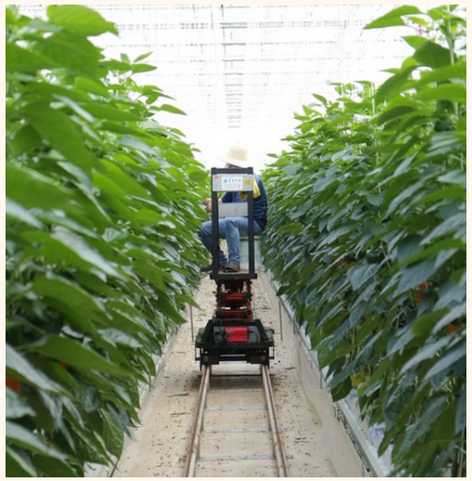
Ideas?

To list on a whiteboard?

Potential positive impacts

On the Environment:

- energy and resources optimisation
  - smart thermostats (heating as needed)
  - smart farming (fertilizer, water, pest control as needed)



Potential positive impacts

On the Society:

- better health monitoring, diagnosis and treatement
- potentially less car accidents with automatic collision detections and braking & lane detection and correction

Potential negative impacts

Ideas?

To list on a whiteboard?

Large Language Models (LLMs) require a very expensive and complex infrastructure

Semi-conductors equipement manufacturing
ASML

Semi- conductors manufacturing
TSMC, Samsung, Intel

AI chips
Apple, Broadcom, Google, Marvell, Qualcomm

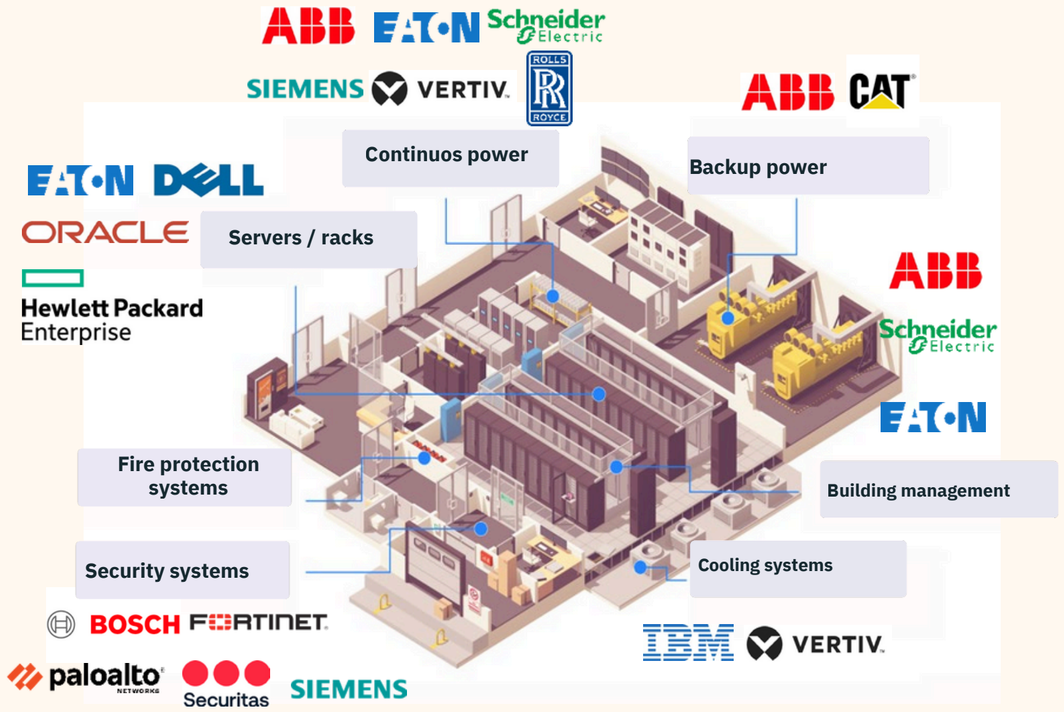
GPU*
NVIDIA, AMD

Memory
Micron Technologies, Samsung

and more

*GPU: Graphics Processing Unit,
**LLMs: Large Language Model

Data senters Alibaba, Amazon, Google, IBM, Microsoft, Oracle



Sources: Quartr Pro, Piguet Galland

They require people, land, materials, energy and water, and are very expensive

Exponential growth

source: OurWorldInData

Robots also require resources and energy

Materials:

-
- Batteries (lithium)
- increased waste and polution
- increased water consumption

Potential negative impacts

On the Environment:

- increased energy and resources consumption (rebound effect)
- increased waste and polution
- increased water consumption

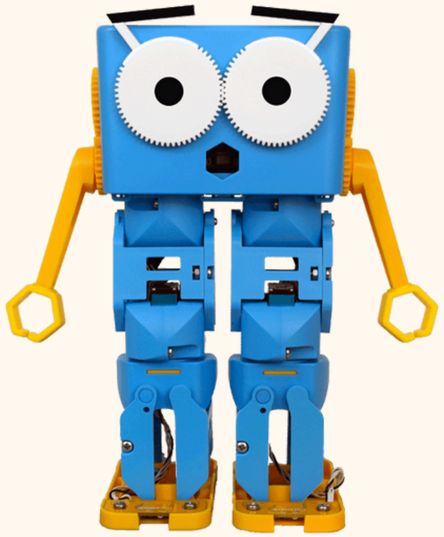
On the Society:

- Isolation
- Over-monitoring
-

The problem is both quality, but even if quality is good, there is always the quantity problem: too much of a good thing is bad

7.98 billion
people on earth

Let’s start building



Marty

screenshot of builidng guide

Annexes

Learning goals

Week 2: Safety of LLMs and Robots

Part 1: presentation (15 min)

Limitations of LLMs

- Hallucinations
- Biases

Uses of LLMs (with concrete examples)

- Brainstorming and Co-creation
- Searching, summarising and detecting patterns
- Tutoring, coaching and personalization of learning
- Evaluations, screening and gradings
- AI Agents
- Other

Learning goals

Week 5: AI Ethics

Ethics

- Definition of ethics: Ethics is the branch of philosophy that deals with what is morally right and wrong, and how decisions affect people and society.
- There is intentional and unintentional harm; physical and psychological harm (examples with non-social robots)
- Accessibility and sustainability concerns are part of ethics
- Ethical technology ideally includes considerations of human development

Learning goals

Week 5: AI Ethics

AI Ethics

- Transparency and explainability
- Data privacy
- Alignment (how LLMs could be moral and ethical) & Mental Health
- Over-Reliance on AI (long term cognitive, socio-emotional, physical and behavioral impacts)
- Manipulation and Persuasion (emotions, feelings and desires)
- Misinformation and Content Accuracy
- Sustainability impacts (energy, materials and water)
- Equity and accessibility
- Legal responsibility & intellectual property (optional)
- Other social impacts (job market, cultural norms, etc.) (optional)

Learning goals

Week 2: AI literacy

Using Large Language Models (LLMs)

- It appers that LLMs nonetheless crease representations and can reason
- LLMs are very good at processing large amounts of text data and present them in a digestable and understandable way
- LLMs can identify patterns and classify information that humans cannot, for example identifying humans' mood and emotions.

Learning goals

Week 5: AI Ethics

Other LLMs best practices

- Prompt Engineering
- Co-thinking
- Reminder: LLMs can’t be reliable sources of information → always find and check the sources

Learning goals

Week 3: Robots Sustainability and equity

Intro

- A robot requires electronics, often a battery, building material which could be anything: plastic, wood, silicone, metal, etc.
- Sustainability is when we can keep doing what we are doing without longterm negative effects (plannetary bounderies?)

Plastic

- Plastic is a magical material that is very light, very cheap, can come in many shapes, and can be durable. It is so successful that we produce so much plastic (give numbers) that we can't manage it.
- The problem with plastic is not CO2 (give numbers and compare with other materials), but polution, which impacts health and biodevirsity. There will be no problem with plastic if quantities were little and we could recycle it all. However, recycling also has a cost, energetically, ecologically and economically.

Lernziele

Woche 3: Roboter, Nachhaltigkeit und Gerechtigkeit

Einleitung

- Ein Roboter benötigt Elektronik, oft eine Batterie, und Baumaterial, das alles Mögliche sein kann: Kunststoff, Holz, Silikon, Metall usw.
- Nachhaltigkeit bedeutet, dass wir das, was wir tun, ohne langfristige negative Auswirkungen (planetarische Grenzen?) weiter tun können.

Plastik

- Kunststoff ist ein magisches Material, das sehr leicht und sehr günstig ist, viele Formen annehmen kann und langlebig sein kann. Sein Erfolg ist so groß, dass wir so viel Kunststoff produzieren (geben Sie Zahlen an), dass wir es nicht bewältigen können.
- Das Problem mit Kunststoff ist nicht CO2 (geben Sie Zahlen an und vergleichen Sie mit anderen Materialien), sondern die Umweltverschmutzung, die sich negativ auf Gesundheit und Biodiversität auswirkt. Es gäbe kein Problem mit Kunststoff, wenn die Mengen gering wären und wir ihn vollständig recyceln könnten. Recycling ist jedoch auch mit Kosten verbunden – energetisch, ökologisch und ökonomisch.

Learning goals

Week 3: Robots Sustainability and equity

Electronics : silicon, metal, plastic

- It is very difficult and expensice to recycle that. We currently don’t have a good solution at scale. Product design can help a lot to increase recyclability, for example by making the materials easily separable and reparable

Batterie

- Lithium is difficult to get (show ww production) and creates geopolitical tensions
- We are starting to see lithium recycling but it is very new

Marty Robot by Robotical

- How Marty’s parts are built and where

Lernziele

Woche 3: Roboter, Nachhaltigkeit und Gerechtigkeit

Elektronik: Silizium, Metall, Kunststoff

- Das Recycling ist sehr schwierig und teuer. Wir haben derzeit keine gute Lösung im großen Maßstab. Produktdesign kann die Recyclingfähigkeit erheblich verbessern, beispielsweise indem die Materialien leicht trennbar und reparierbar gemacht werden.

Batterie

- Lithium ist schwer zu beschaffen (siehe WW-Produktion) und führt zu geopolitischen Spannungen
- Wir sehen bereits erste Anzeichen für Lithiumrecycling, aber es ist noch sehr neu

Marty Robot von Robotical

- Wie Martys Teile gebaut werden und wo

Learning goals

Week 3: Robots Sustainability and equity

Positive and negative feedback loops

- Sustainability is closelly related to ecological systems
- Ecological systems are governed by non linear relationships with positive (reinforcing) loops and negative (stabilising) loops
- These feedback loops have tipping points, which are irreversable (at human scale) changes to the system reaching a new equilibrium.
- Plastic polution is likely gone beyond tipping points

Equity and Society

- Equity focuses on results, while equality on means
- AI and robotics will substantially change societies, sometimes positively, sometimes negatively
- Costs, knowledge, skills and infrastructure are barriers to equitable benefits from AI
- Cultural differences and unequal representation in training data impact quality of AI products used by non western countries

Lernziele

Woche 3: Roboter, Nachhaltigkeit und Gerechtigkeit

Positive und negative Rückkopplungsschleifen

- Nachhaltigkeit ist eng mit ökologischen Systemen verbunden
- Ökologische Systeme werden durch nichtlineare Beziehungen mit positiven (verstärkenden) und negativen (stabilisierenden) Schleifen gesteuert.
- Diese Rückkopplungsschleifen weisen Kipppunkte auf, bei denen es sich um irreversible (auf menschlicher Ebene) Veränderungen des Systems handelt, die zu einem neuen Gleichgewicht führen.
- Die Plastikverschmutzung hat wahrscheinlich den Wendepunkt überschritten

Gerechtigkeit und Gesellschaft

- Gerechtigkeit konzentriert sich auf Ergebnisse, während Gleichheit auf Mittel
- KI und Robotik werden die Gesellschaft grundlegend verändern, mal positiv, mal negativ
- Kosten, Wissen, Fähigkeiten und Infrastruktur sind Hindernisse für einen gerechten Nutzen aus KI
- Kulturelle Unterschiede und ungleiche Repräsentation in Trainingsdaten beeinträchtigen die Qualität von KI-Produkten, die in nicht-westlichen Ländern verwendet werden

Learning goals

Week 3: Robots Sustainability and equity

In summary

- The problem is related to the scale of production and how we design products – whether it includes the whole lifecycle of the product or not.
- The pace of innovation and economical growth related pressures shorten lifecycle of products, increasing the level of production and waste management.
- Recycling has a cost and an impact and biodegradability is often second to reduced consumption.
- Optional: Kaya identity **(very good at explaining : Expanding the scope of sustainability: Broaden sustainability education beyond a sole focus on and waste management to include critical concepts like mindful and reduced consumption)**

Lernziele

Woche 3: Roboter, Nachhaltigkeit und Gerechtigkeit

Zusammenfassend

- Das Problem hängt mit dem Produktionsumfang und der Art und Weise zusammen, wie wir Produkte entwerfen – unabhängig davon, ob dabei der gesamte Lebenszyklus des Produkts berücksichtigt wird oder nicht.
- Das Innovationstempo und der mit dem Wirtschaftswachstum verbundene Druck verkürzen die Lebenszyklen von Produkten und erhöhen den Produktions- und Abfallmanagementaufwand.
- Recycling ist mit Kosten und Auswirkungen verbunden und die biologische Abbaubarkeit steht oft hinter der Verbrauchsreduzierung zurück.
- Optional: Kaya-Identität (sehr gut zum Erklären: Erweiterung des Umfangs der Nachhaltigkeit: Erweitern Sie die Nachhaltigkeitsbildung über einen alleinigen Fokus auf Abfallmanagement hinaus, um kritische Konzepte wie achtsamen und reduzierten Konsum einzubeziehen)

Proposed activities

Week 3: Robots Sustainability and equity

Option 1

Estimation

- Build a model to calculate how much production per year is needed of each material to build a Marty robot for each class in Germany, then estimate yearly waste after one life cycle. Compare with total yearly waste in Germany. Change model parameters to see how they impact the results. Can you find a sustainable level and mode of production?

Vorgeschlagene Aktivitäten

Woche 3: Roboter, Nachhaltigkeit und Gerechtigkeit

Option 1

Schätzung

- Erstellen Sie ein Modell, um zu berechnen, wie viel Produktion pro Jahr von jedem Material benötigt wird, um einen Marty-Roboter für jede Klasse in Deutschland zu bauen. Schätzen Sie anschließend den jährlichen Abfall nach einem Lebenszyklus. Vergleichen Sie mit dem gesamten jährlichen Abfall in Deutschland. Ändern Sie Modellparameter, um zu sehen, wie sie sich auf die Ergebnisse auswirken. Können Sie ein nachhaltiges Produktionsniveau und eine nachhaltige Produktionsweise finden?

Proposed activities

Week 3: Robots Sustainability and equity

Option 2

Improving Marty

- Propose improvements to Marty to reduce its sustainability impact

Vorgeschlagene Aktivitäten

Woche 3: Roboter, Nachhaltigkeit und Gerechtigkeit

Option 2

Marty verbessern

- Schlagen Sie Marty Verbesserungen vor, um seine Auswirkungen auf die Nachhaltigkeit zu verringern

Proposed activities

Week 3: Robots Sustainability and equity

Option 3

Playing games

- Adaptation of beer game for system thinking: a robot parts game? We could include user and waste / recycling stations
- A short version of climate fresk game
- Create a new game, for example "Robot Crisis Supply Chain Game" combining the Beer Game mechanics with sustainability pressures:
  - 4 Player Roles: Lithium Miner, Electronics Manufacturer, Robot Company, School District
  - Random Events: Climate protests, mining strikes, new recycling tech, regulation changes
  - Hidden Information: Each player only sees their piece of the puzzle
  - Victory Condition: Keep robots flowing to schools while meeting sustainability targets
- Other games (see list)

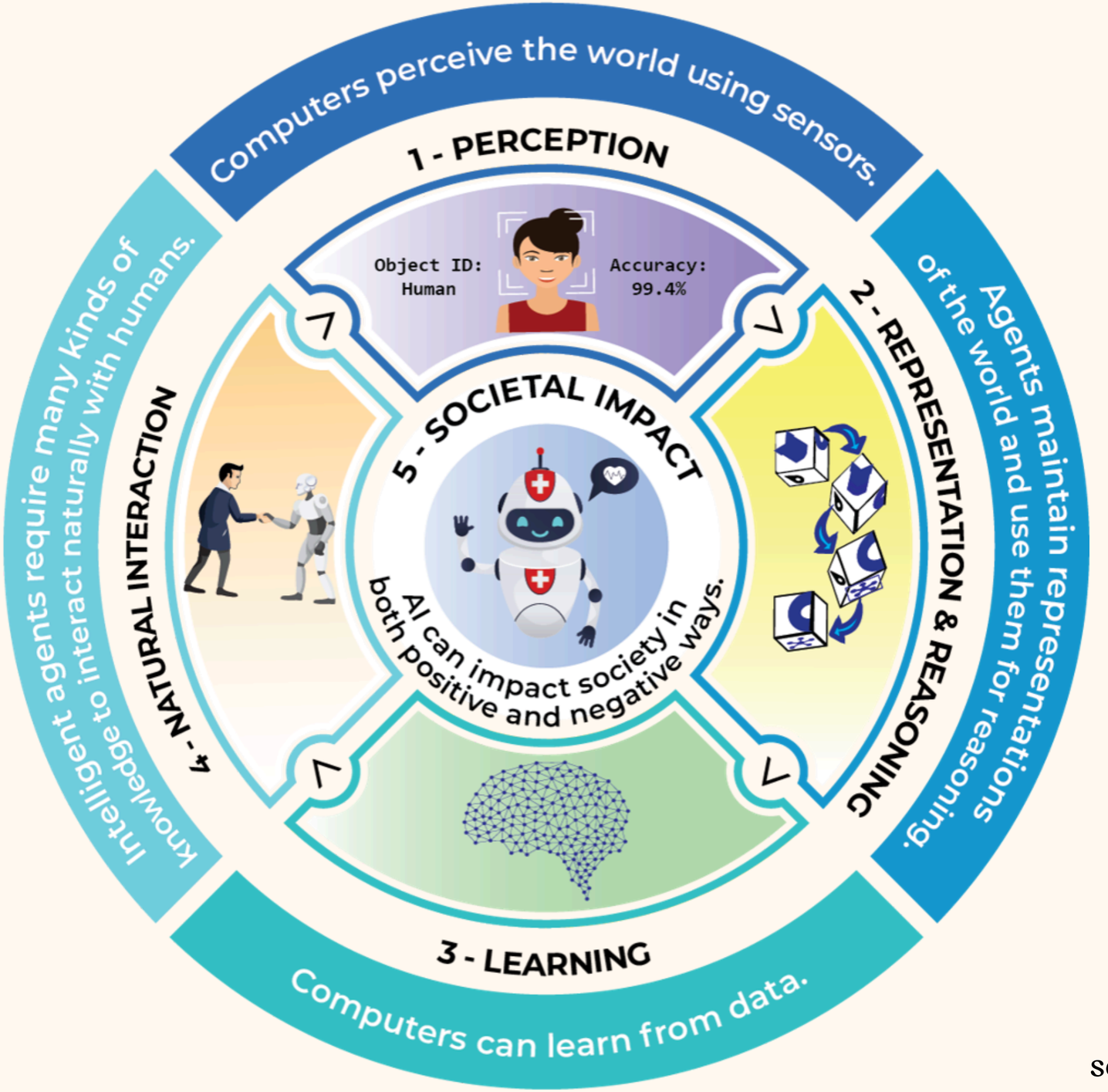
Vorgeschlagene Aktivitäten

Woche 3: Roboter, Nachhaltigkeit und Gerechtigkeit

Option 3

Spiele spielen

- Anpassung des Bierspiels an das Systemdenken: ein Roboterteilespiel? Wir könnten Benutzer- und Abfall-/Recyclingstationen einbeziehen
- Eine Kurzversion des Klima-Fresk-Spiels
- Erstellen Sie ein neues Spiel, zum Beispiel „Robot Crisis Supply Chain Game“, das die Spielmechanik des Bierspiels mit Nachhaltigkeitsdruck kombiniert:
  - 4 Spielerrollen: Lithium-Bergmann, Elektronikhersteller, Roboterfirma, Schulbezirk
  - Zufällige Ereignisse: Klimaproteste, Streiks in den Bergwerken, neue Recyclingtechnologien, Änderungen der Vorschriften
  - Versteckte Informationen: Jeder Spieler sieht nur sein Puzzleteil
  - Siegbedingung: Aufrechterhaltung der Roboterversorgung der Schulen und gleichzeitige Erreichung der Nachhaltigkeitsziele
- Andere Spiele (siehe Liste)



source: https://ai4k12.org