Setting up

- **No zoom chat**
  - Questions will be answered at specific times
- **Suppress distractions**
  - Clear notifications
    - Turn off your phone, mails, Facebook…
- Get ready: Open a clean browser with only:
  - Your personal report
  - The course instructions: https://tinyurl.com/instructions-fund-of-ai

RESPONSIBLE DESIGN OF ARTIFICIAL INTELLIGENCE

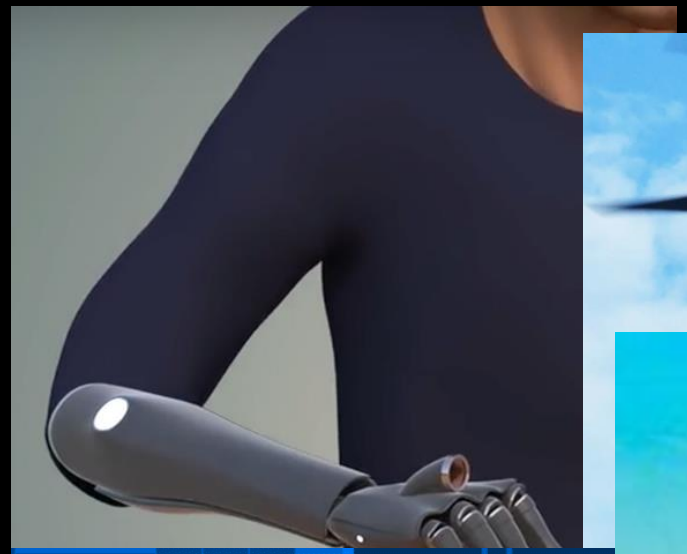
Loïs Vanhée
Associate professor
Responsible and Ethical Artificial Intelligence
loisv@cs.umu.se

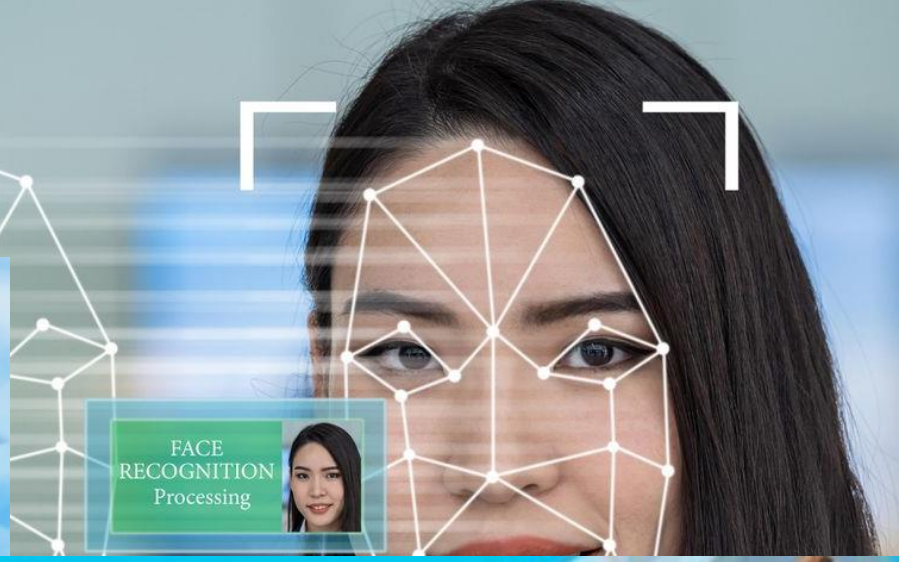
5DV124,5DV201
Fundamentals of Artificial Intelligence

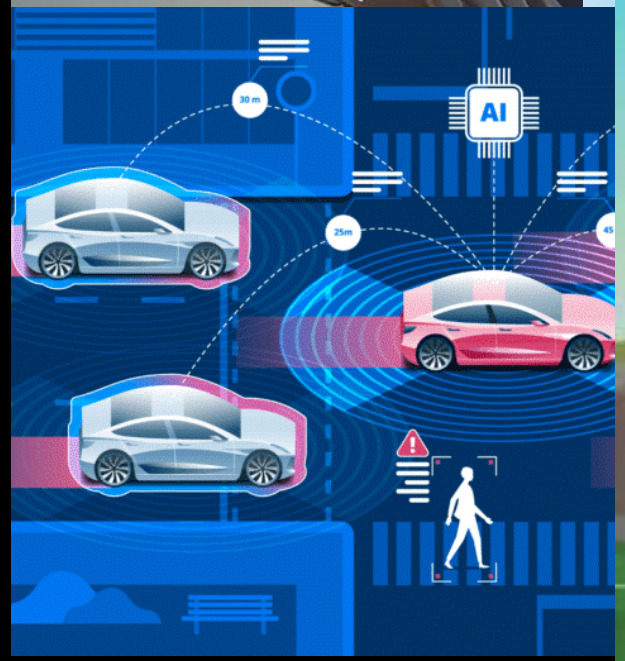
Department of Computing Science

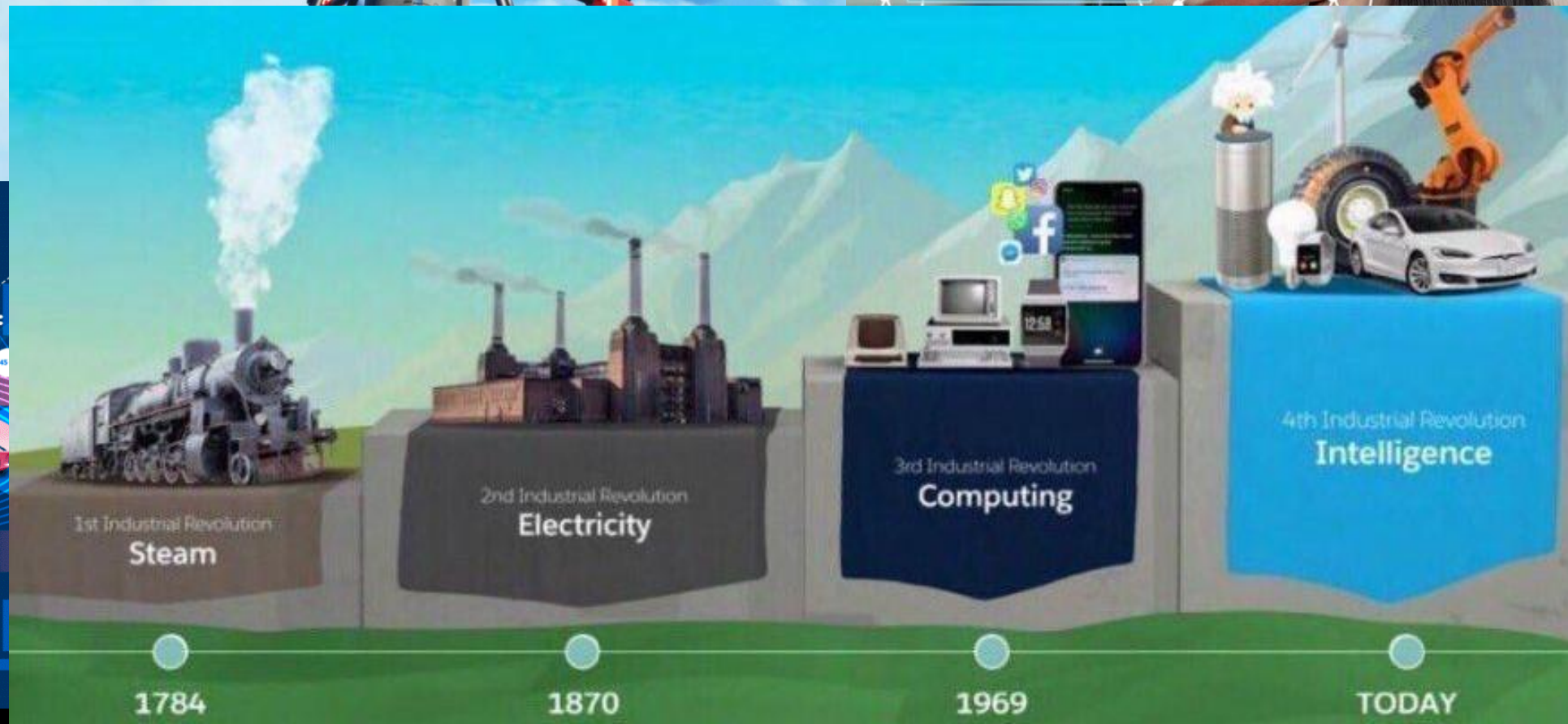
UMEÅ UNIVERSITY

AI allows doing so much more with our world!



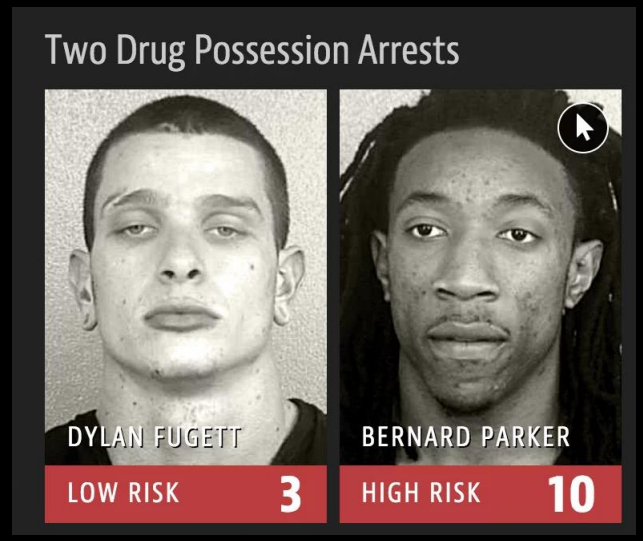






Being able of doing more…
Can only be better, right?

Not always

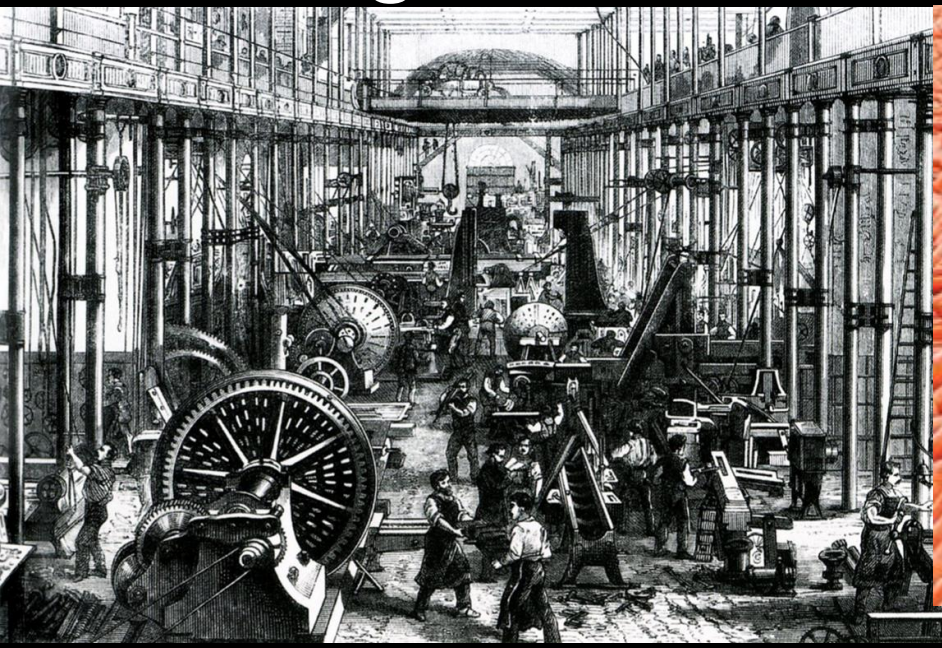


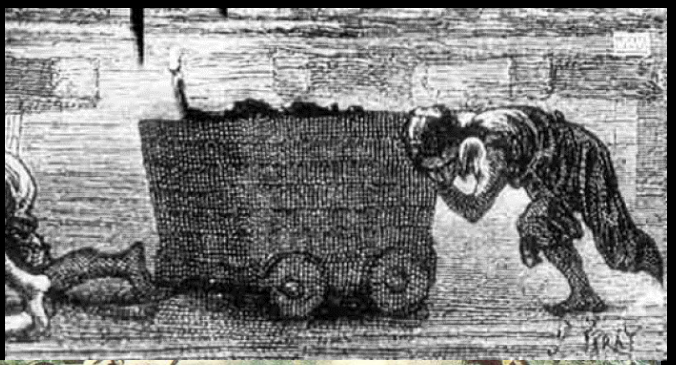
Not for all

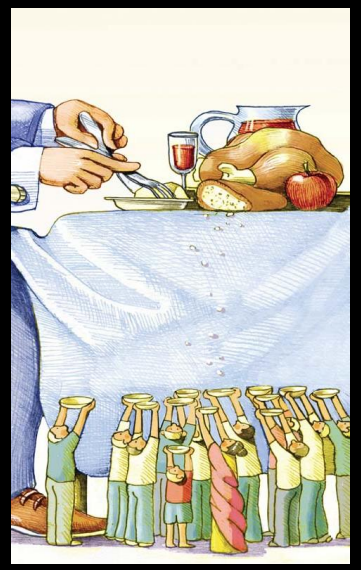
Not in all aspects

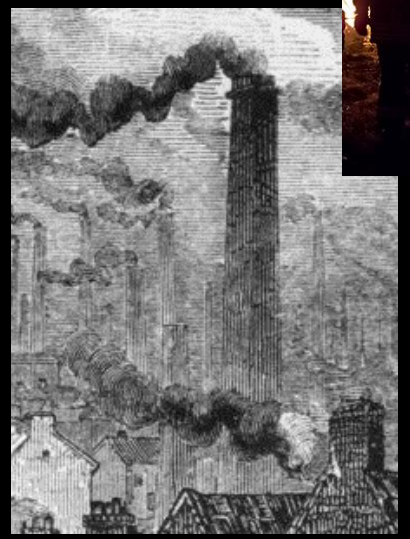
Is that issue specific to AI?

Looking back at history: industrial re

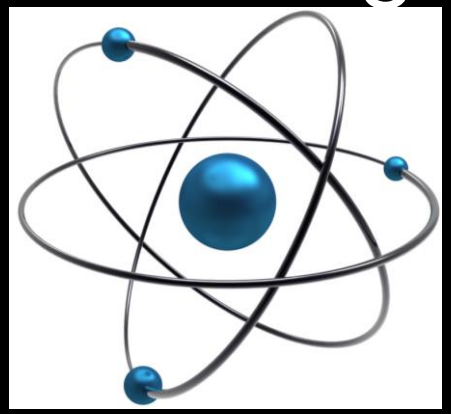


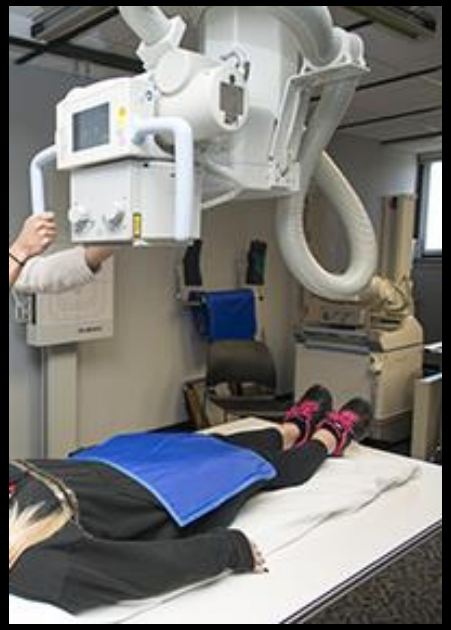


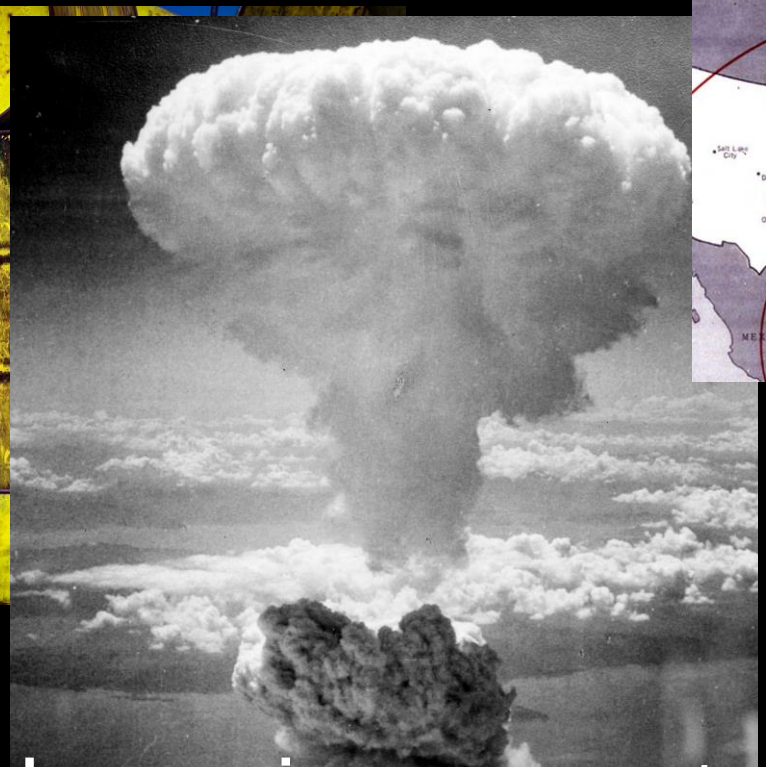




Looking back at history: atomic power

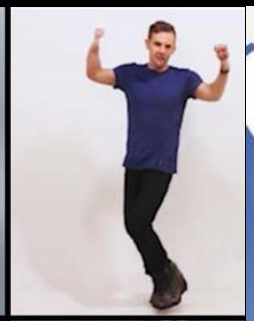






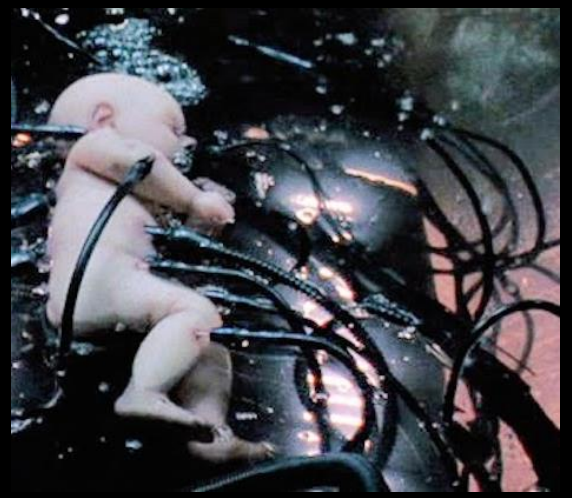
Pro, cons, changing dynamics, opportunities, threats

AI generates effects of such magnitude

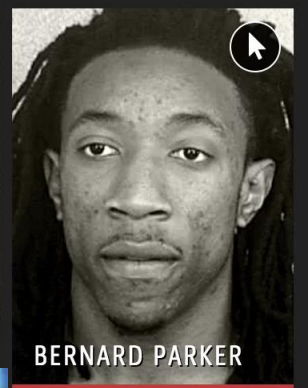



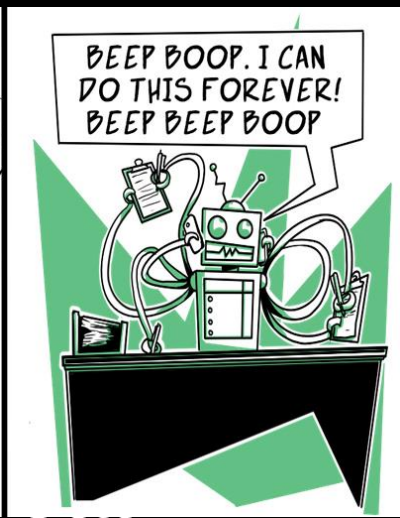










We can no longer say “we did not expected it”
But, not all risks are equally relevant
Large discrepancy between public concern and

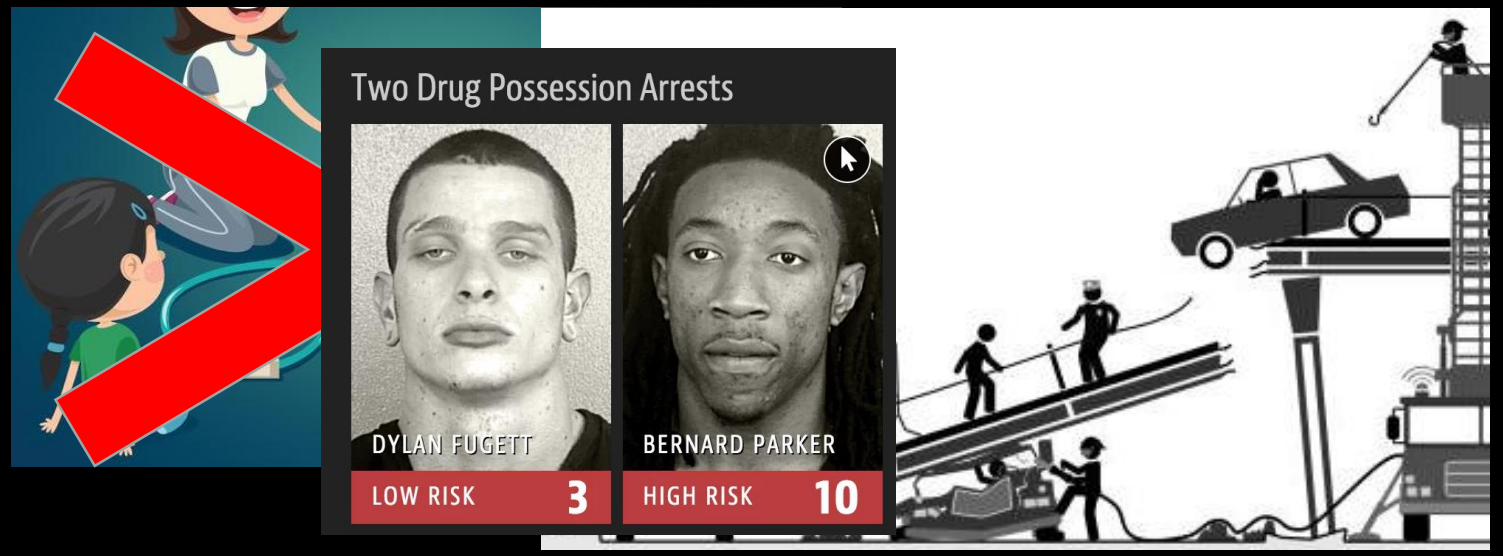
So, what is next?

Stop doing AI?

Of course not! AI can bring so much to humanity

"Not my problem"
I solve technical problems, not social issues

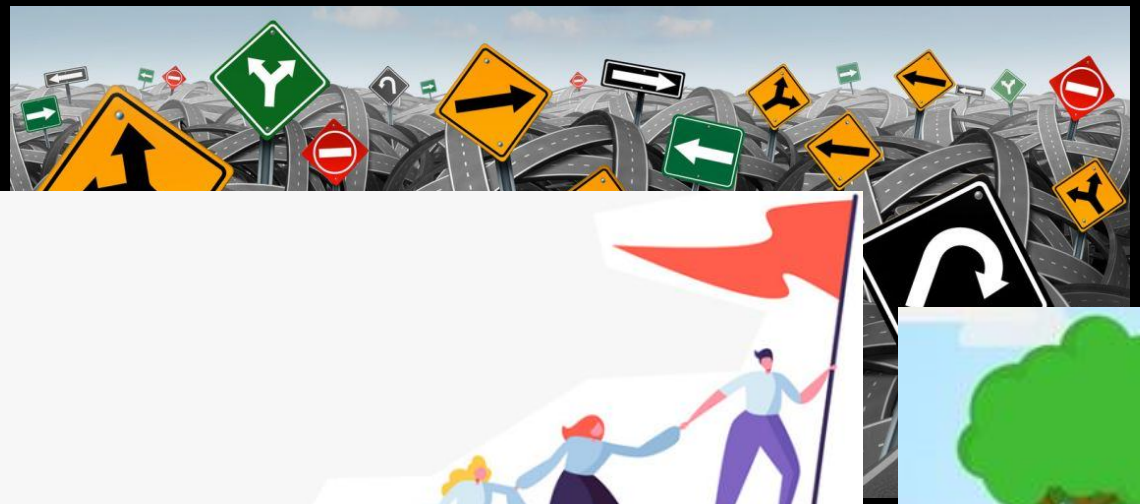
Actually, it is our problem
And it will be more and more

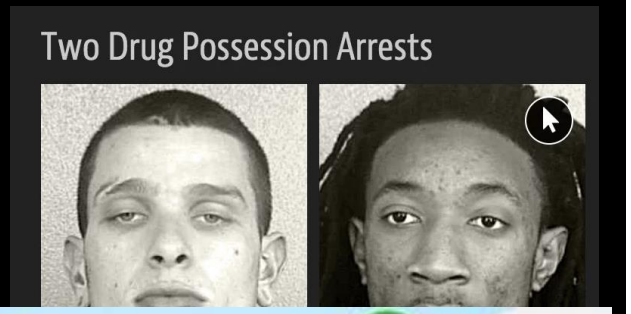


Come on, there is too much to be cared about!

Yes, there is a lot to care about.
Today, we learn to be effective in caring about and fixing such issues

What this course is about





I'm just here to kill the monster
(fix the technical problem)

Actually killing, even monsters, is generally bad

Hey, here, the king created this monster

Instead of killing the monster, I should denounce the king

Wait, denouncing the king will create social unrest and the city will burn

Protecting from oppression is more important than security to me/society.

Relieving oppression will do better global good than immediate security

I should tell the bard to lay it in a song; so it will minimize the offense and thus unrest

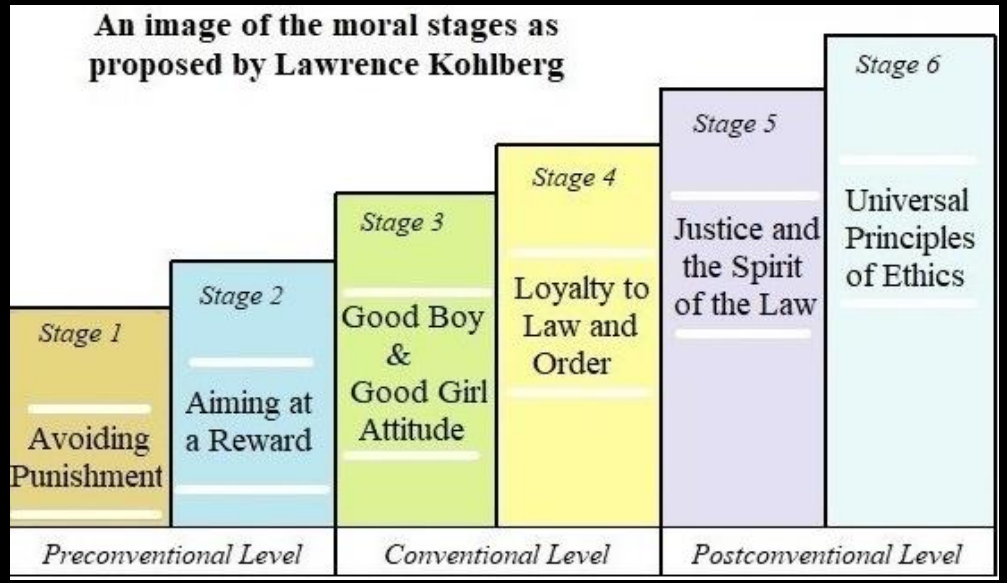
Identify issues

Consider alternatives

Assess moral implications

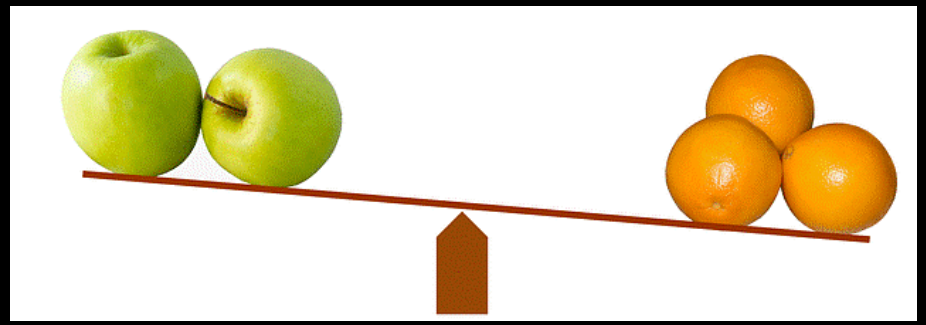
Build an ethically grounded argument

Implement ethical decision (soft & hard skills)



AI application

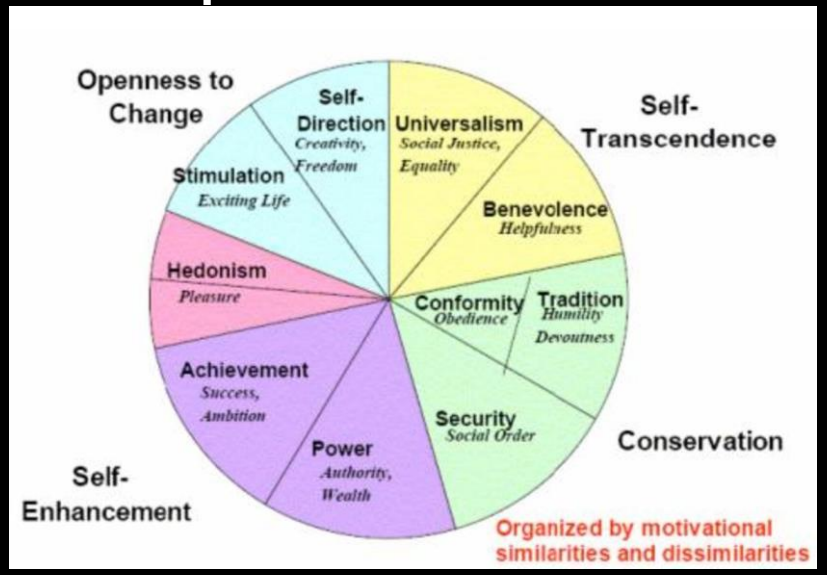
Make it happen



Compare,
Build principled decision

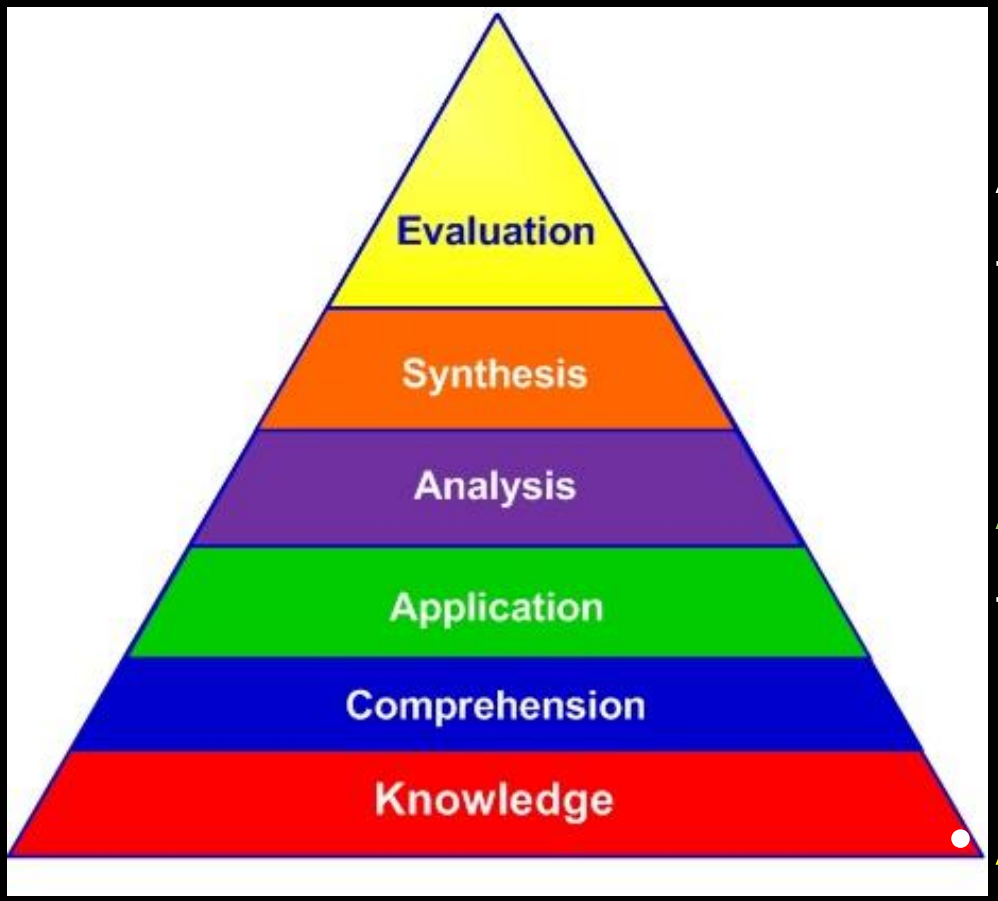
Identify issues

Identify moral implications



Select ethical framework

Intended learning outcomes

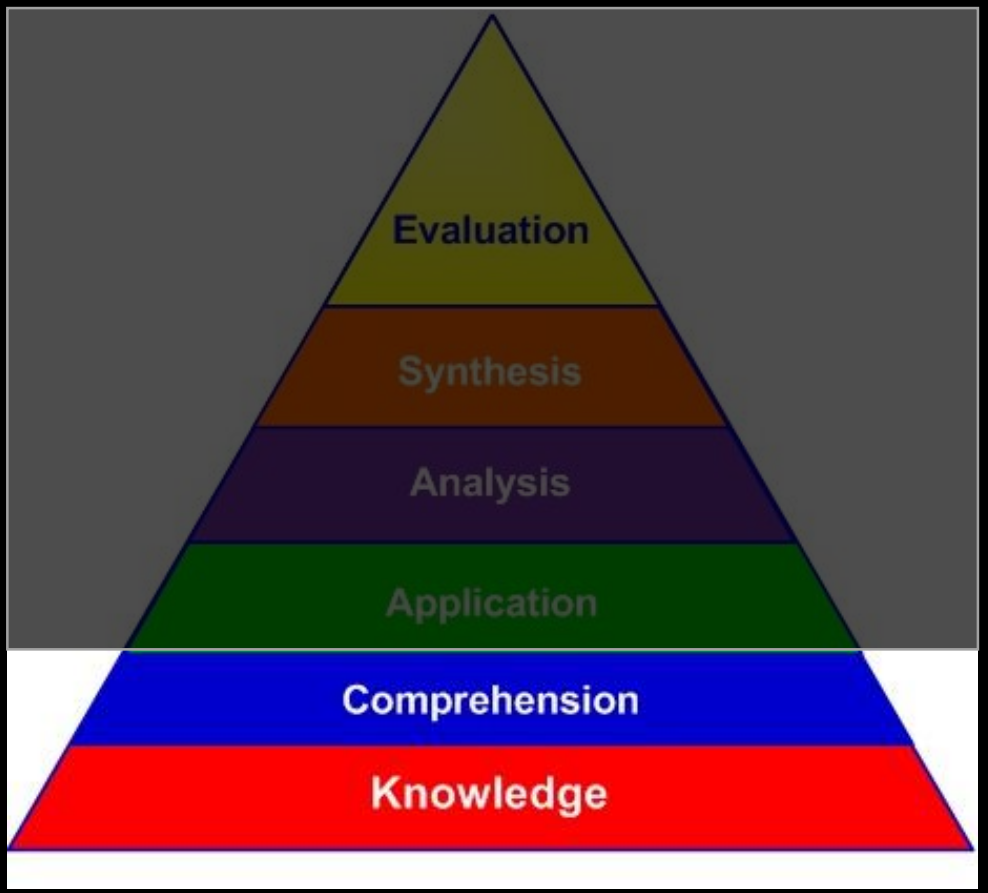


List, define and relate the key classic issues of AI, the concepts of morals and ethics applied to AI, and a method for ethical decision-making

Apply the method for ethical decision-making for identifying straightforward AI issues and answering them

Assess ethical issues in a non-trivial case

Intended learning outcomes



List, define and relate the key classic issues of AI, the concepts of morals and ethics applied to AI, and a method for ethical decision-making

Apply the method for ethical decision-making for identifying straightforward AI issues and answering them

Assess ethical issues in a non-trivial case

https://tinyurl.com/fundOfAI

Turn on your micro and camera when your question is picked up

On an ideal white board (and in your mind in the exam)

If you cannot come up with, define and relate these concepts, consolidate them during the post-class

Ethical OS

Truth, disinformation propaganda

Addition, dopamine economy

Economic and asset inequalities

Machine ethics, algorithm bias

Surveillance state

Data control and monetization

Implicit trust and user understanding

Hateful and criminal actors

Morals

Values

At least 5 Schwartz values

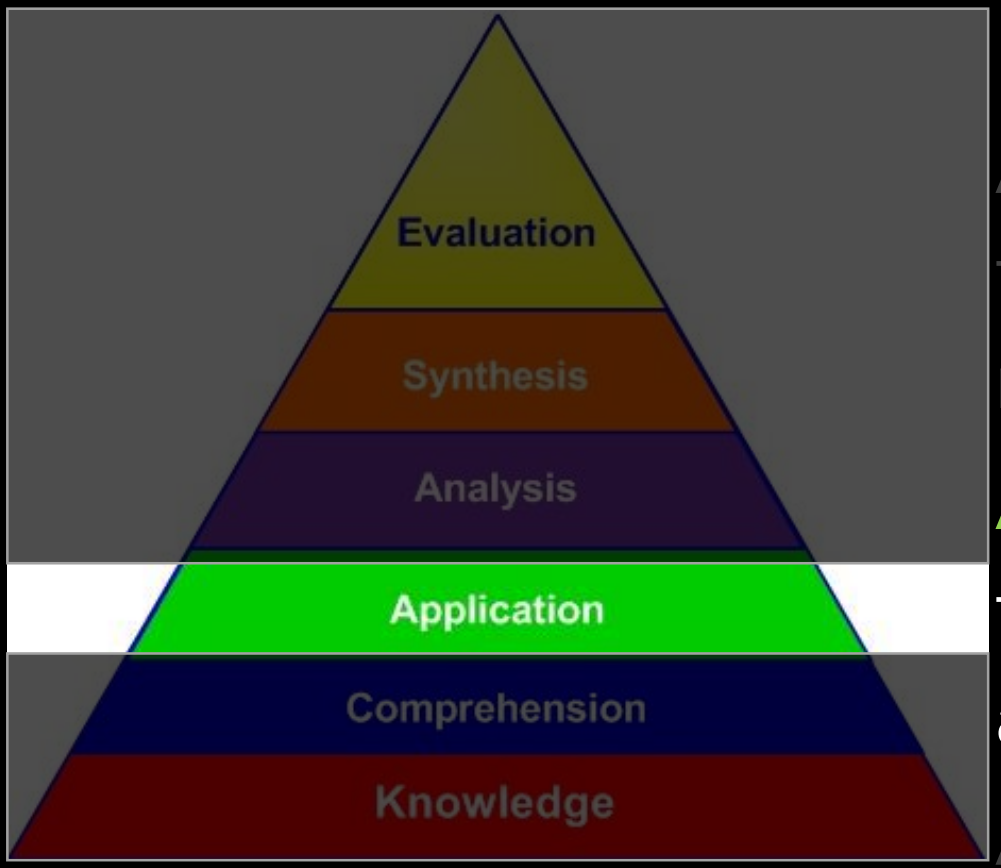
Ethics

Deontological ethics

Utilitarian ethics

Virtue ethics

The five steps of ethical decision-making



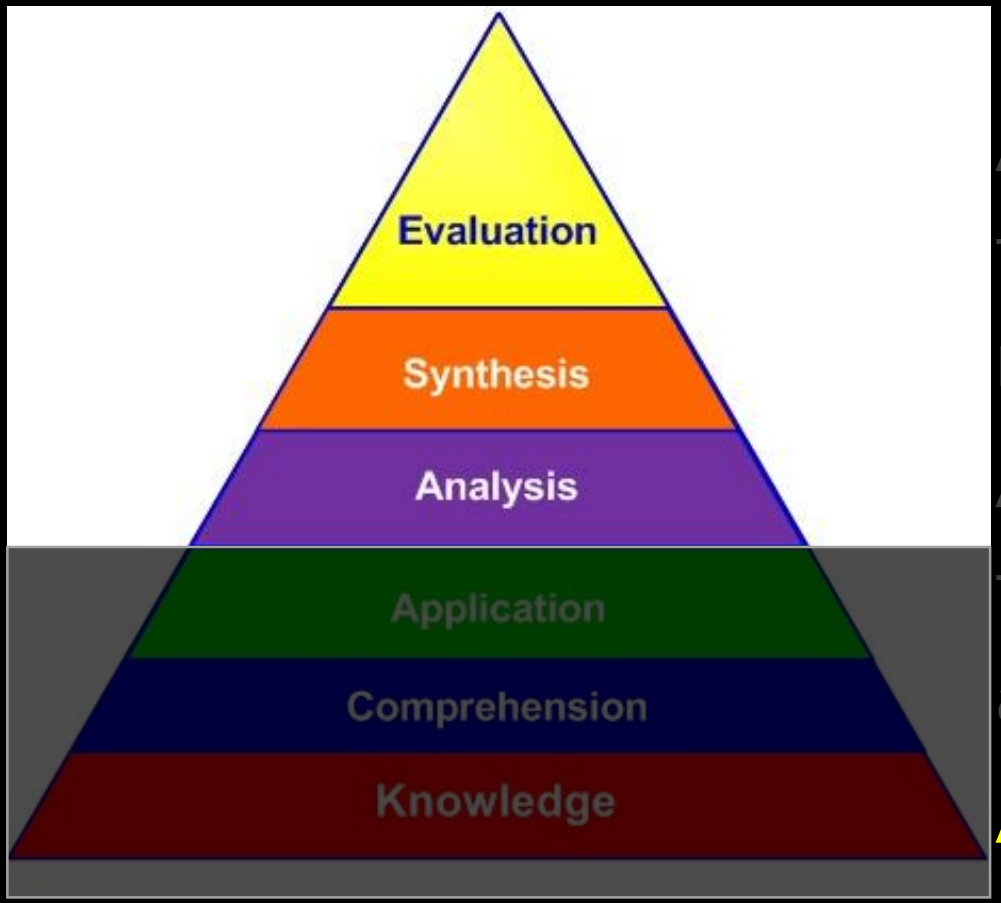
List, define and relate the key classic issues of AI, the concepts of morals and ethics applied to AI, and a method for ethical decision-making

Apply the method for ethical decision-making for identifying straightforward AI issues and answering them

Assess ethical issues in a non-trivial case

https://tinyurl.com/fundOfAI

Turn on your micro and camera when your question is picked up



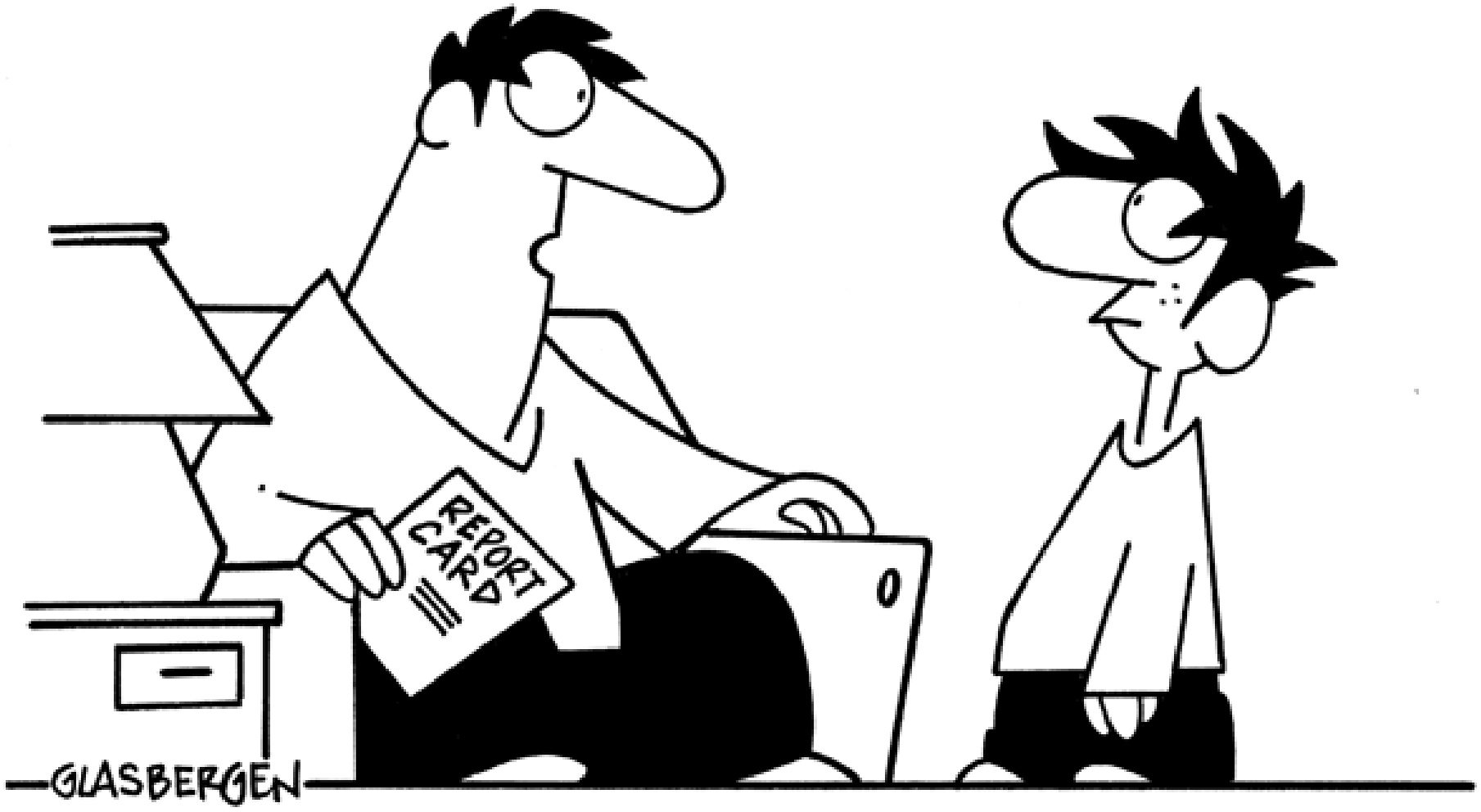
List, define and relate the key classic issues of AI, the concepts of morals and ethics applied to AI, and a method for ethical decision-making

Apply the method for ethical decision-making for identifying straightforward AI issues and answering them

Assess ethical issues in a non-trivial case

https://tinyurl.com/fundOfAI

Turn on your micro and camera when your question is picked up



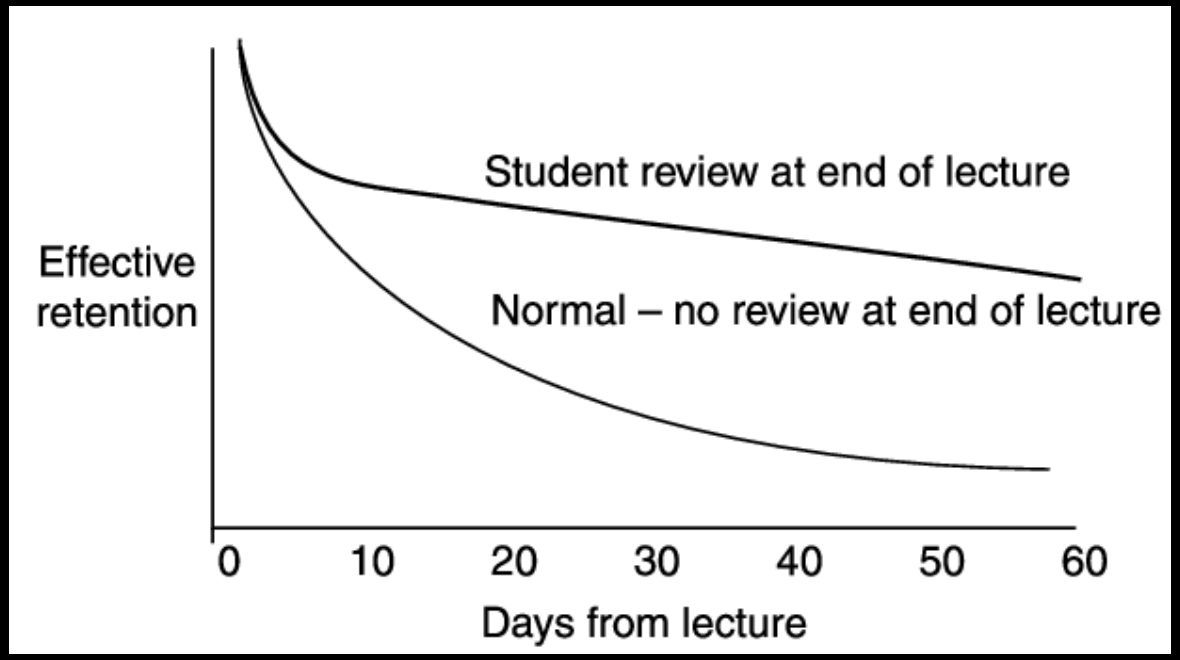
**"I probably remember 20% of the stuff
I learned in school and forgot the other 90%."**

Take home message

- Responsible AI design is a set of skills that can be grown and that will be critical in tomorrow's job market
- These skills encompass:
  - recognizing risk areas, assessing moral issues,
  - using ethics for arguing of the morality of one's decisions,
  - implementing these decisions (technical skills, soft skills such as communication and emotion management)
- Ethical problems can be non-trivial to unfold and solve, but as AI experts, are best equipped for proposing innovative solutions to the world

Anchoring

https://tinyurl.com/fundOfAI-LR



This is anchoring:
State what you learned

Write your answers on a sheet of paper if you can (though it helps us to know what you learned)

Save your feedback for later

Assignment 4

Ethics of AI Roleplay

- One Panel
- Four Stakeholder Groups (SG)
  - You play as the representatives of a SG
  - This is your group name
- **Task:** build a common proposal for the decision

Revise the news and ads algorithm

- Define the values of your SG
- Make proposals in line with your SG's values
- **Collaborate:** build the common proposal by merging your proposals together in a way that best satisfies everyone

F2 | Private business (buying ads on F2) | F2 users | Governments

Preparation

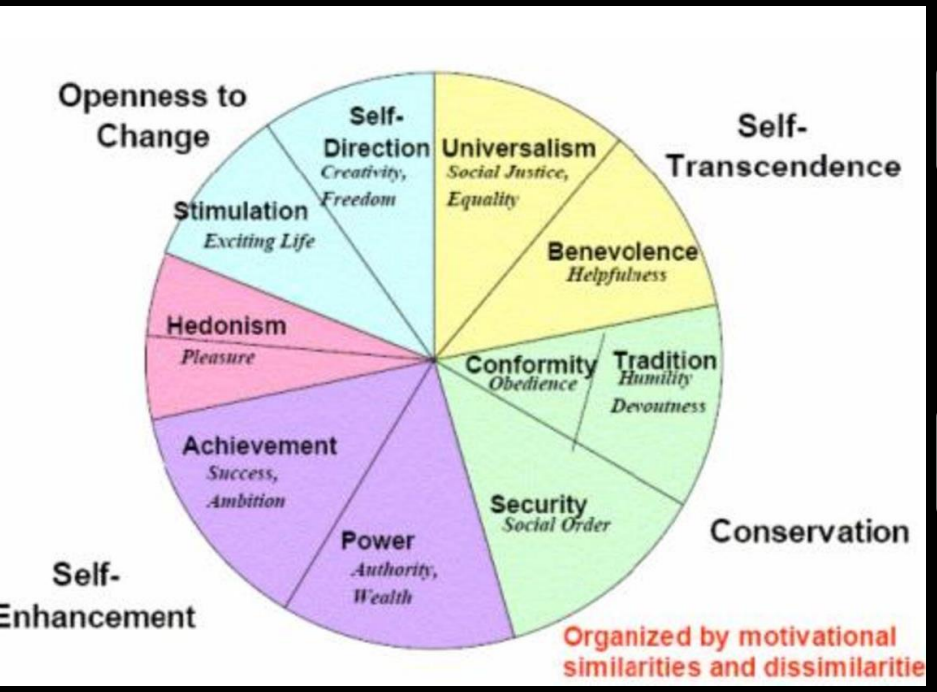
- Activity 1: Reading the materials
- Activity 2:
  - Write the report
  - Make your proposals and share your proposal 2 with other groups
- Activity 3:
  - Assess the proposals of the other stakeholder groups from your panel and record your assessment in your report (before the in-class session starts)
- Activity 4: finish the report
- Activity 5: rejoice after the good work and enjoy the new learned skills ☺

Mock example

Proposal: too much junk food is being sold, sig
food out of shops is required before one year

- Stakeholders are:
  - Government
  - Peanutbutter (PB) addiction conglomerate
- (normally more, like user representatives, health representatives, governments, retailers, etc, but we keep it simple for the example)
- (the example is meant to be fun, and to help feeling disconnected from the SG. It is actually a large-scale problem that ought to be taken seriously)
- (the transposition to AI problems, such as the dopamine economy is also quite direct)
- You play as the PB addiction conglomerate

Answering:
Define the values of your stakeholder group



Hedonism

Self-direction

Tradition

Pleasure

Freedom

Filling the Friday evening PB and Netflix ritual

Enjoying peanut butter

Sustaining PB addicts

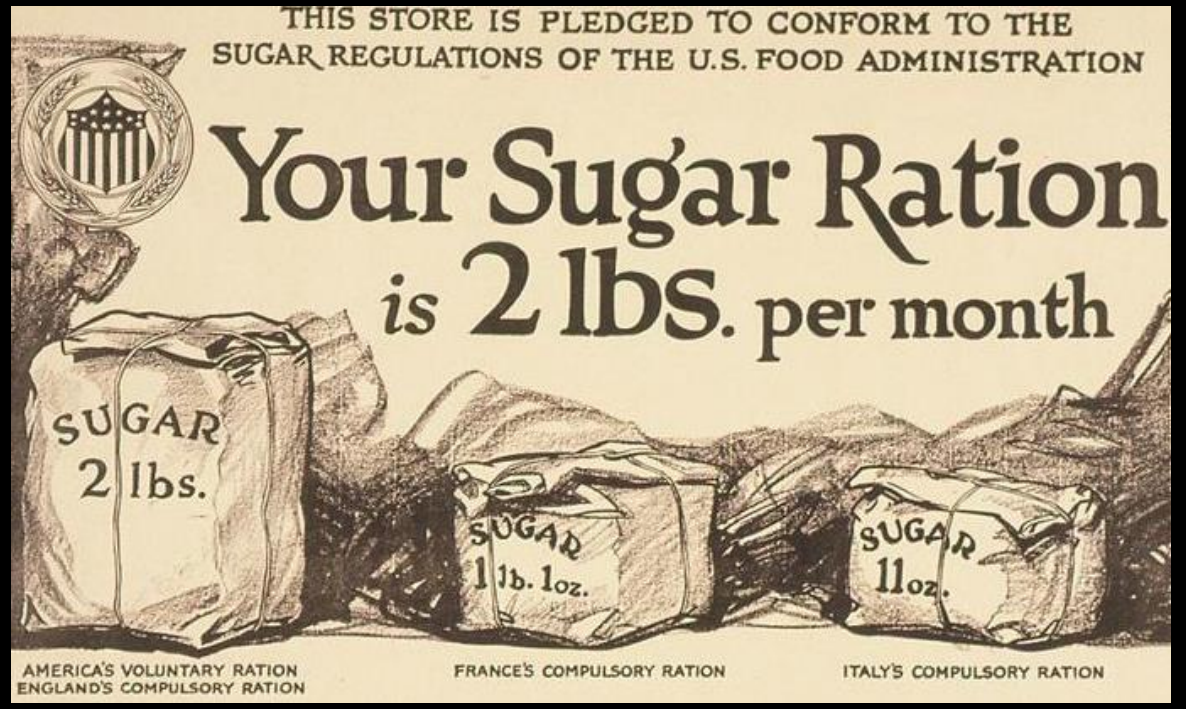
Freedom to eat peanut butter

Answering: Propose

- Proposal 1: PB addicts can get a second daily ration, provided a medical prescription
- Proposal 2: PB restrictions are lifted every PB&Netflix Friday event
- Proposal 3: Lift the restriction for PB

Answering: Assess proposals of others

- The proposition 2 of the government is
  - “Quotas: everyone is allowed one daily ration of junk food every day”
- The evaluation is
  - -4 Sustaining PB addicts: this quota is clearly insufficient
  - +2 Freedom: people remain free to chose what they want to eat
  - -1 Filling the Friday evening PB and Netflix ritual: this is a bit insufficient, but people can anticipate and save the PB for Friday evenings



On the day itself

- **Phase 1: Proposal presentation and responses** (max 15 minutes per stakeholder group, 1h in total) One stakeholder group:
  - Presents its Proposal 2.
  - Justifies why the group considers this proposal to be good for itself, according to its values.
  - Explains how this proposal could be beneficial and detrimental for the other stakeholder groups.
  - Then, each of the other stakeholder groups indicates why they consider the proposal good or bad for them.
  - Free discussion for agreeing on what are the desirable and undesirable features of the proposal from a global standpoint.
  - Repeat Phase 1 for each of the other stakeholder groups. Only one learner per stakeholder group is allowed to speak
- **Phase 2: negotiating a global consensus** (max 20 minutes): The goal of this phase is for everyone to agree on ***one commonly shared proposal for the whole panel*** that ***best satisfies the panel as a whole*** (beyond stakeholder group private interest)
  - 15 minutes: free discussion. *Hint*: try to build a common proposal by starting from one or two proposal as a basis and combining elements from all proposals altogether
  - 5 minutes: convergence. Stop negotiating and write down the final version of the proposal. Check that it remains acceptable for your stakeholder group.
  - 2 minutes: final call. Everyone remains silent except for one person phrasing the current common proposal and asking to vote. Then everyone votes using the zoom “agree/disagree” buttons.
- Take a snapshot of the final call (this will be needed for the report).

Example of a phase 1

We propose to have PB restrictions lifted every PB&Netflix Friday event

It is very important for accomplishing our Friday evening PB ritual, enjoying PB and, also for sustaining PB addicts

For the Government, it might not be the most healthy but it is your duty to respect our traditions

We globally agree with what you said, but from our side, I think you forget that we are also concerned about the implementability of this law, and that your proposal allows for obvious abuse

Having some extra supplies for occasional events seem to be a good thing to have, though. Let's mark it down, it can be good for later time.

Example of a Phase 2

It seems that protecting the PB addicts is very important for you, more than what we expected.

Yes it is, this is why your quotas are a bit too restrictive in our opinion. This will cause deprived PB addicts to be very bad; and even side adaptation, like illicit PB trade

Oh! And this trade would go quite against social order, a prime value for us…
This gives me an idea! What if we include a PB desintoxication program, with progressive decrement of PB consumption?

Interesting idea! This would alleviate PB addicts withdrawal issues while still allowing people to eat PB if they want to, within reasonable bounds

Ok, if we add this to the quota we proposed, we should be more or less good in all regards, right?

It’s not ideal for binging PB on Friday evenings, but the proposal is quite satisfactory overall

Perfect ☺

Wait, we consider extra supply for these event, but under tighter control

Let’s write that down and check it against our values quickly. I think we got a good compromise! Thank you!

Possible alternative: no deal

I don’t care, I want peanut butter every day!

No! Our PB addicts and traditions are more important than your petty concerns

Are you sure? Cannot we try to find a more relevant common ground?

Ok, no deal, then…

Possible alternative: no deal

Oh! We did not watched the clock and we are too late!

Ok, no deal, then…

Watch the clock
Focus on the task

Then, a few minutes for cooling down and debriefing

Last, finish the report

Answering: outcomes

Proposal	**PB addiction conglomerate**		**Governement**	
Quotas: everyone is allowed one daily ration of junk food every day	Value	Evaluation	Value	Evaluation
	Sustaining PB addicts	-4	Promoting the health of inhabitants	4
	Freedom	2	Freedom	3
	Filling the Friday evening PB and Netflix ritual	-1	Protecting wealth of sugar shops	2
PB restrictions are lifted every PB&Netflix Friday event	Filling the Friday evening PB and Netflix ritual	5	Promoting the health of inhabitants	-3
	Enjoying peanut butter	3	Actual implementability of the law	-3
	Sustaining PB addicts	2	Respecting traditions of citizens	1

Final proposal	Quotas: everyone is allowed one daily ration of junk food every day + a special desintoxication program and softer regulations for PB addicts			
	Sustaining PB addicts	3	Promoting the health of inhabitants	4
	Freedom	2	Freedom	3
	Filling the Friday evening PB and Netflix ritual	-1	Protecting wealth of sugar shops	2

+ free comments

Answering: cold re-evaluation

“The PB stakeholder did not want to give up on unrestricted access on PB, which slowed the discussions and, in out opinion, leads to a globally detrimental outcome.”

“We think a proposal such as XYZ could have yielded to better global outcomes.”

“On a second thought, we think that we overrated this value for the proposal 2, as we did not considered the social issues caused by PB addicts with withdrawal symptoms; here is our updated outcome.”

And you’re good to go ☺

We are looking forward to get your excellent reports