**CODE GPK7H**

Overview of ethics.

Ethics define what is good and bad, what is correct and not correct to do.

* good
* right
* correct

What you would do (descriptive) in contrast with normative (what is the right thing to do vs what we should do).

3 partition in normative ethics:

* **First order**: define normative rules of what we all should do
* **Meta ethics**: investigate on the topic, what is the basic idea
  + **Logic-semantic**: *what is the meaning of good/right…*, do they follow propositional logic? what are the truth condition of ethical claims
  + **Ontological**: *do moral facts exist and if so how? Are they subjective or objective?* (or intersubjective). Money is a construct of society, is the value that we give to it
  + **Epistemic**: *how do we have moral knowledge?* find with empirical investigation?
* **Applied ethics**: basically applied first order (life ⟹ life ethics, business ⟹ business ethics, digital world ⟹ digital ethics) there is one way in which ethics of AI is a subfield of a subfield of a subfield. It is part of digital ethics, part of applied ethics and part of normative ethics.

Machine ethics. Does not feel like an application of something that we already know. Patients vs agents. Moral patient vs moral agent:

* moral patient: has moral rights, not necessarily moral duties.
* moral agent: have moral duties, has to obey ethics. Healthy adult human beings, humans under normal conditions have moral duties.
* me: both moral agent and moral patient
* donkey: moral patient only. Not hold the donkey morally responsible to kill someone as a human

History of ethics: expansion of the two spheres. 200 years ago different sensibility, not all humans were considered moral agents. Now we start to expand the sphere of moral patients to include also the environment, future generations. We think we have some duties to protect/preserve. **What about non-human but intelligent agents?** Will they have any duties? if they are going to be superintelligent?

**Utilitarianism, consequences, deontological, categorical moral reasoning, categoricalism**

**Ontology**: based on duty, what we need to do. Based on doing the right things, there is a set of absolute values that you should follow. What *should I do?*. You should follow your duties to apply a principle/rule.

**Consequentialism**: what are the consequences of one action? What counts are the final outcomes

Both try to answer the question: what *should I/we do?*

**Particularism**: it is hard to define, there are no general moral principles. Not related to duties. There are two ways in which one can get to this theory

* looking at what we do when we try to reach a conclusion about what we should do

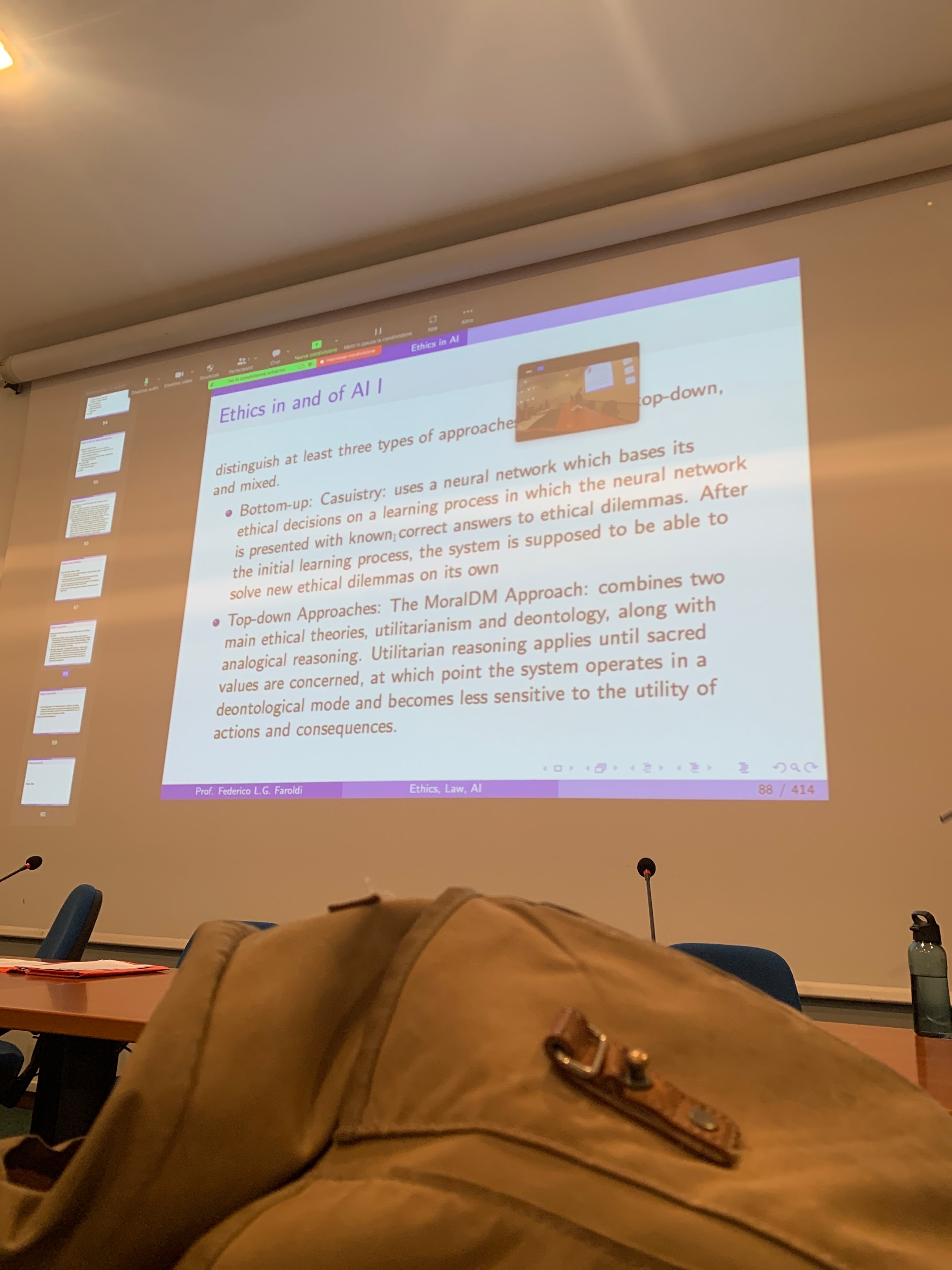
relevant factors known to you. Then face weighing for one option and against one option.

reasons have weights and it has to be balanced. From this we have all things considered obligation or prototo reason.

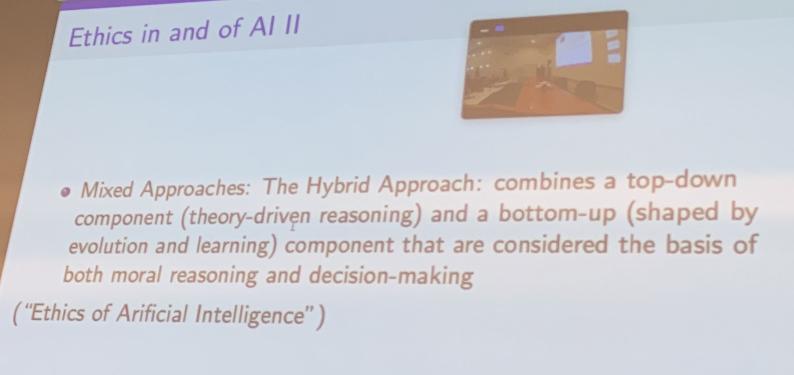
hot → ¬ run

rain → ¬ run

hot + rain → run



Pure ethical agents, sort of mixes.



Moral uncertainty. Moral knob on self driving car, switch before driving and decide what is ethically right to do, so that the car acts as you would do. How should we behave when we don’t know which ethical theory is correct?

Moral underdeterminacy or value gaps.

Value gluts

Moral uncertainty tries to provide an answer on what to do if we don’t know what moral theory is correct. Apply standard ideas of decision theory to these domain

Decision theory: how to act rationally in case of uncertainty. You have a set of options/actions. Each associated with an outcome of certain probability. Each outcome has an outcome associated with it. *What should I do? Should I buy a 10$ lottery ticket with 0.01% of win 1 mln of one of 15$ or 0.015 of winning 10 mln?* check the expected value of each lottery ticket. Idea of using expected values when we come to moral uncertainty

Moral uncertainty. What we should do if we are not sure which moral theory is correct. In the end we have to make a decision (also not taking a decision is taking a decision). Can use the notion of expected values. Problem: the value that we assign (presuppose that we actually could) even if it works out we use subjective standards, does not really solve the problem of social reason