Moral uncertainty: how to act morally if the theory is not correct. Proposal: lift up the main idea of the moral domain.

DK if doing option A or B. To each course of action you connect utility. You maximize the expected value. Bounce of options/outcome. Each has a value. And each outcome would have a probability.

A 30%

B 20%

C 50%

Unclear constraints on the value function. Idea that you do the product of the value with the respective probability.

Buy a lottery ticket (A) with 1 million euros but 0,001 probability. Expected value is 1 million \* 0,001.

**You should choose the one with the highest expected value**. (main idea of decision theory)

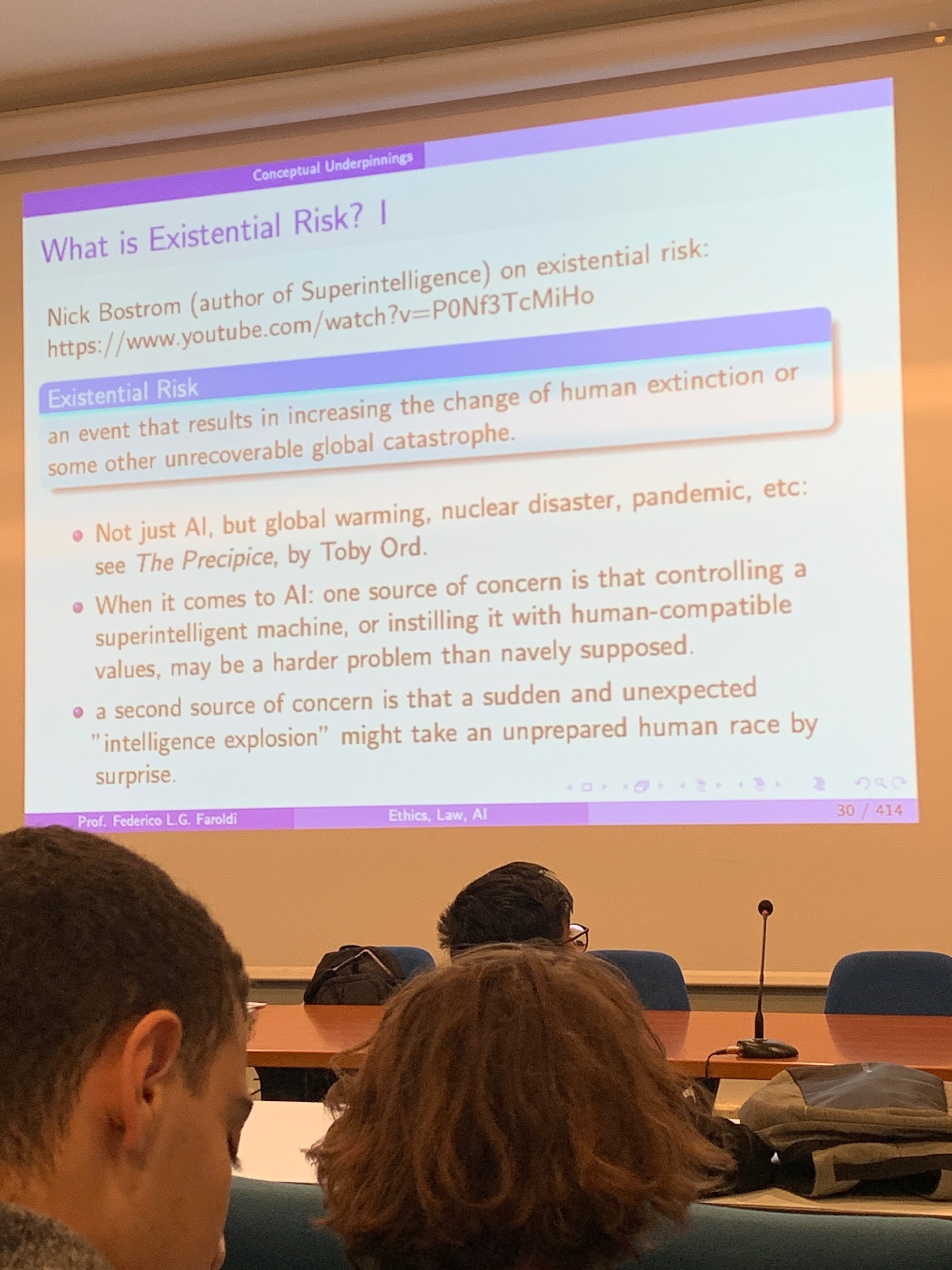
**DECISION THEORY**: tells how to behave rationally but as a single individual.

Don’t know which moral theory is correct. Assign to each moral theory your subjective credence that that theory is correct.

How the ontologist calculates value.

Moral uncertainty ⟹ intertheoretical compatibility. Compare values across different theories.

You absolutely dislike onions and someone else dislikes garlic. Can we compare taste across different individuals?



[The end of humanity: Nick Bostrom at TEDxOxford](https://youtu.be/P0Nf3TcMiHo)

**EXISTENTIAL RISK**: global pandemic, eruption, asteroid...

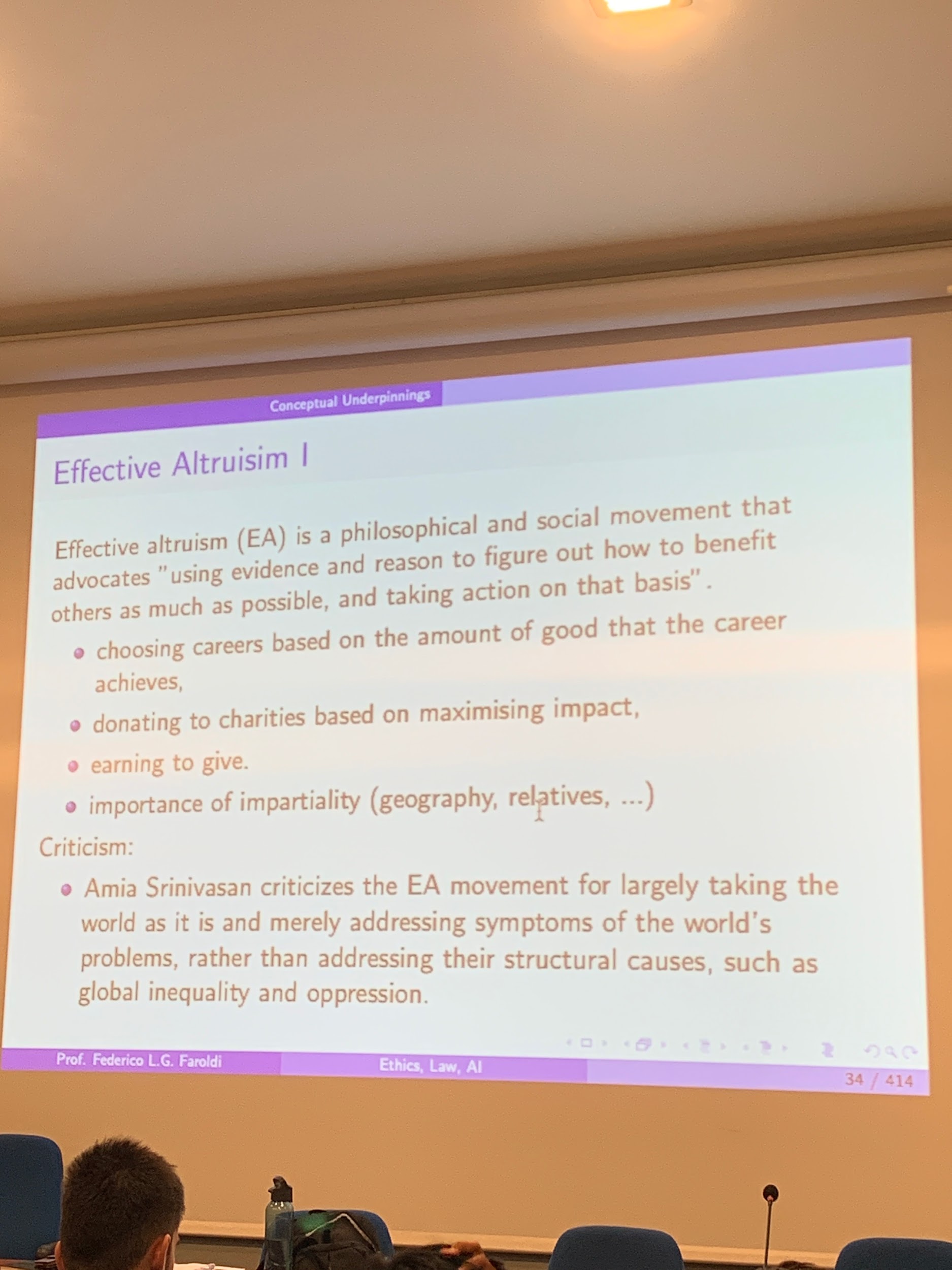
Some risk we can not do much about it. Do we care? Some people don’t care about preserving life in the world as it is. Suppose we care about life on earth. Then there are mitigation strategies to try to reduce risks.

Mitigation strategy: mandatory the use of seatbelts. Decrease the changes of death.

In the end no intrinsic value at all is what we do. We are chemical compounds that try to make sense, probably the planet and other species would be better if we would disappear.

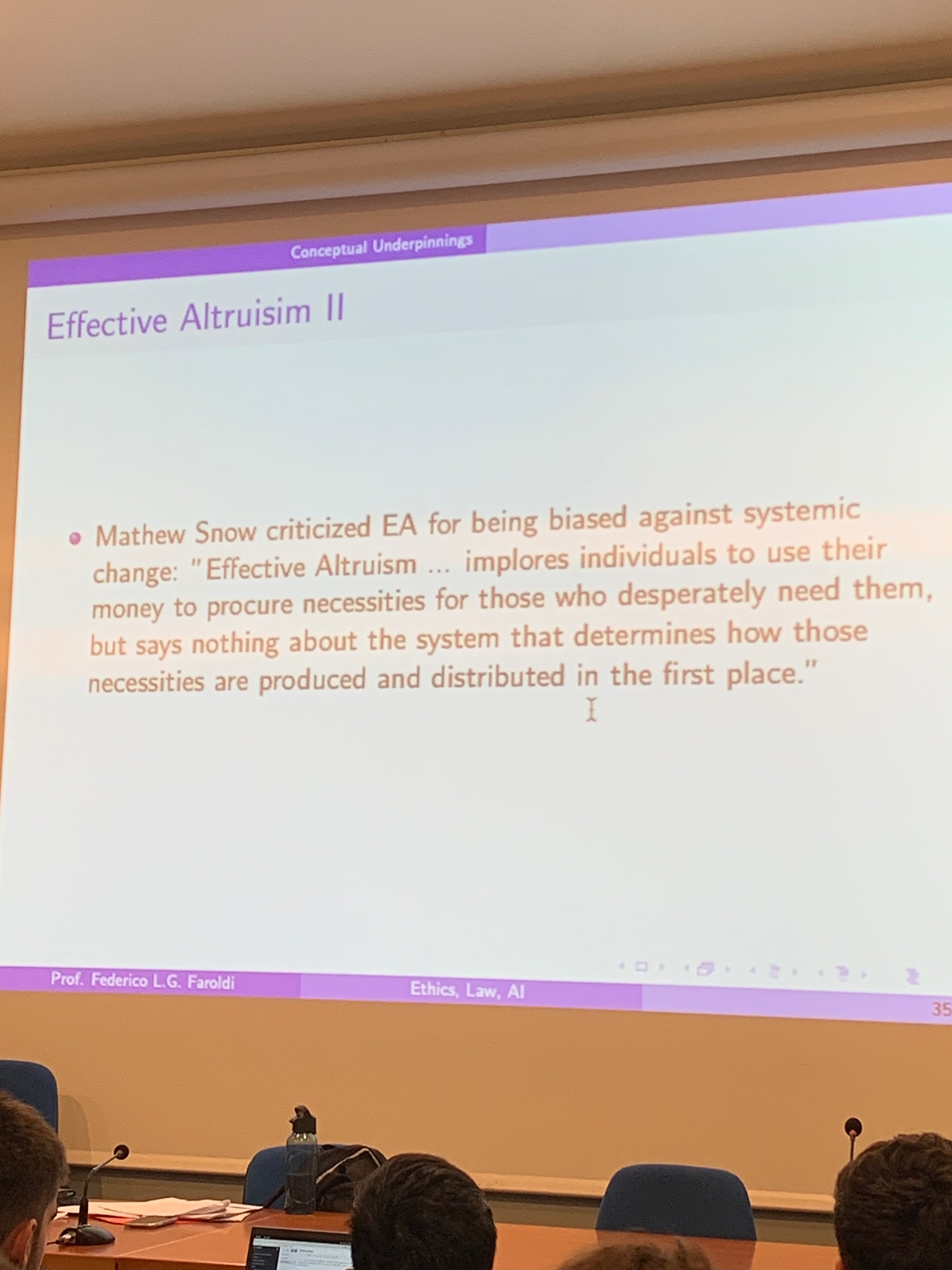
1. In ethics there is prejudice to give a higher value to people close to you (family, neighbor, the country) in the proper way there should be no distinction. (from *peter singer*)
2. **Long-termism,** people that are suffering because life is impacted by rich countries' commitment to a greener future. People in the future will benefit from the actions that we are taking now and that damage only “few” billion people.

**Effective altruism**



Decide to which charity to give in such a way to have the highest impact. Quantity the value of the outcome to maximise the impact of your donation. Expanding and people choose careers to maximise their profit.

Objection against effective altruism. You might try to act rationally but you are addressing symptoms rather than acting to the root causes. Instead of buying the mosquito nets to decrease malaria you should +compl



Critique: giving money to private. It feels like the money was not under a public democratic choice then the society should choose what they should found with that money. Right now they give a lot of power to privates.

Nick Bostorm, create tought experiment.

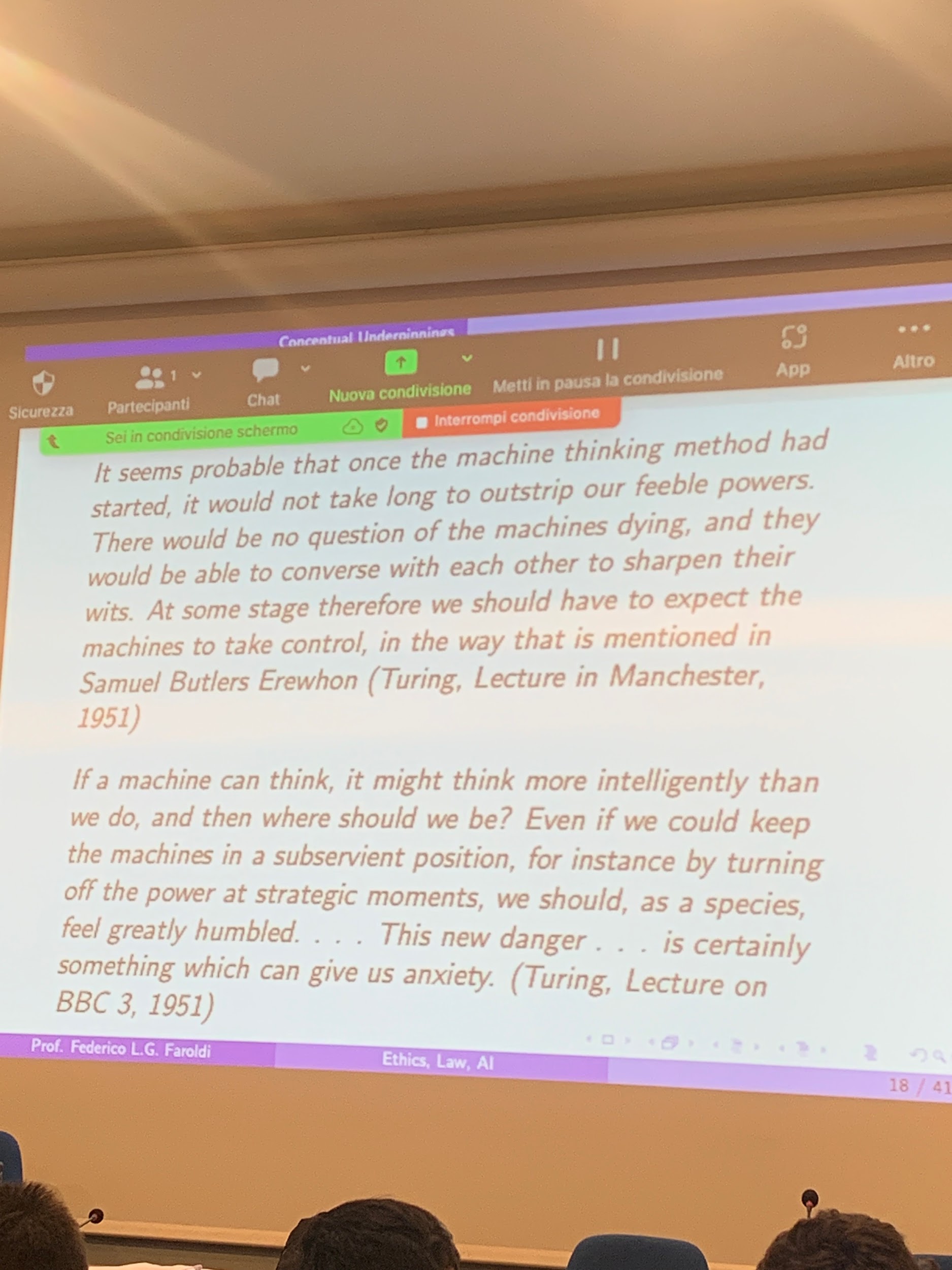
1. pace
2. Nuclear war kill 99%
3. Nuclear war kill 100%

To who should be careful when we want to mitigate risk? Do we care if an isolated bacteria dies and goes extinct? there is also intelligence life (seems to exclude agents that might be intelligence but that are not alive in biological sense)

two requiremsn

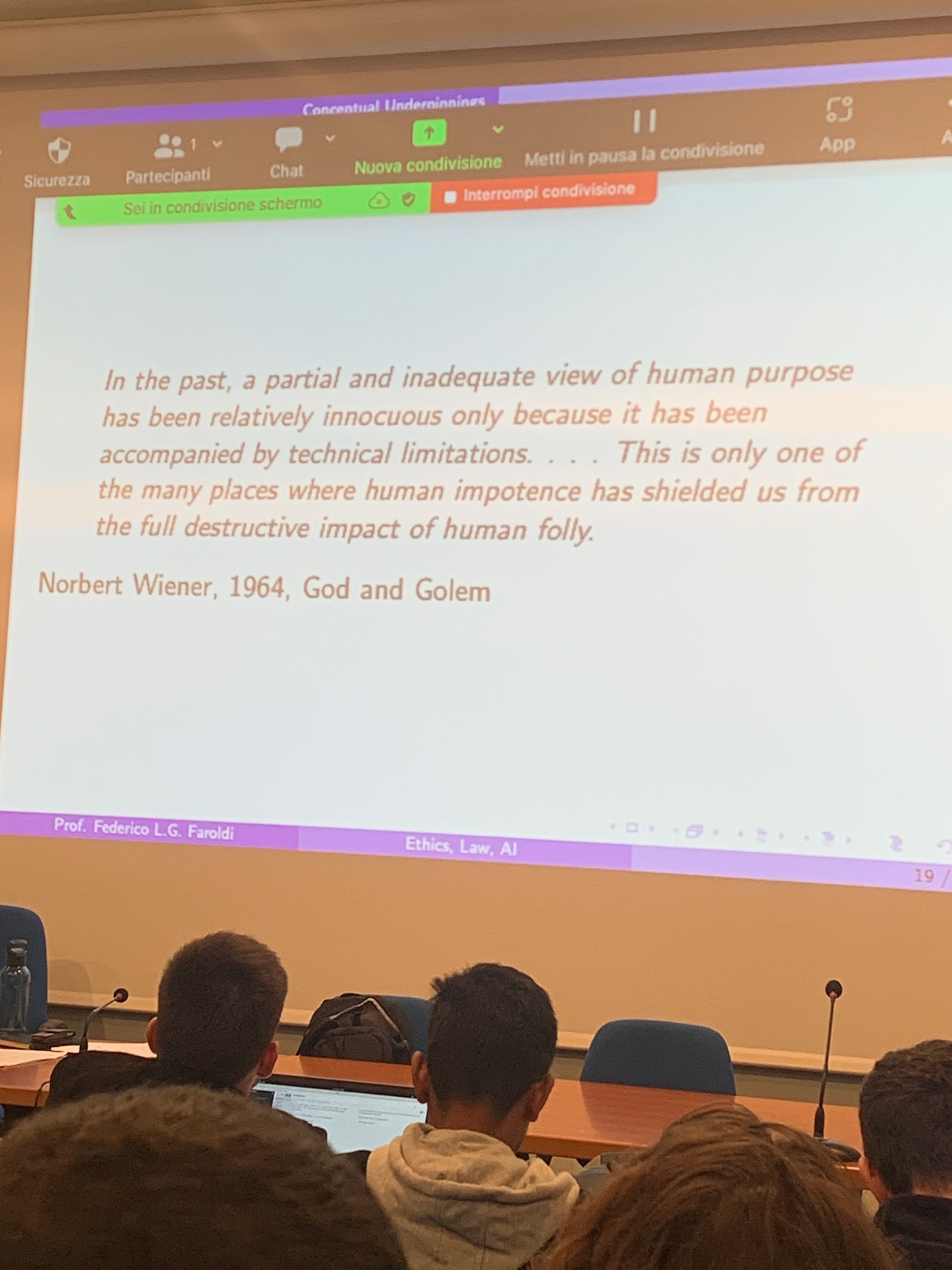
* intelligence
* one has to be alive/potentially alive in the biological way

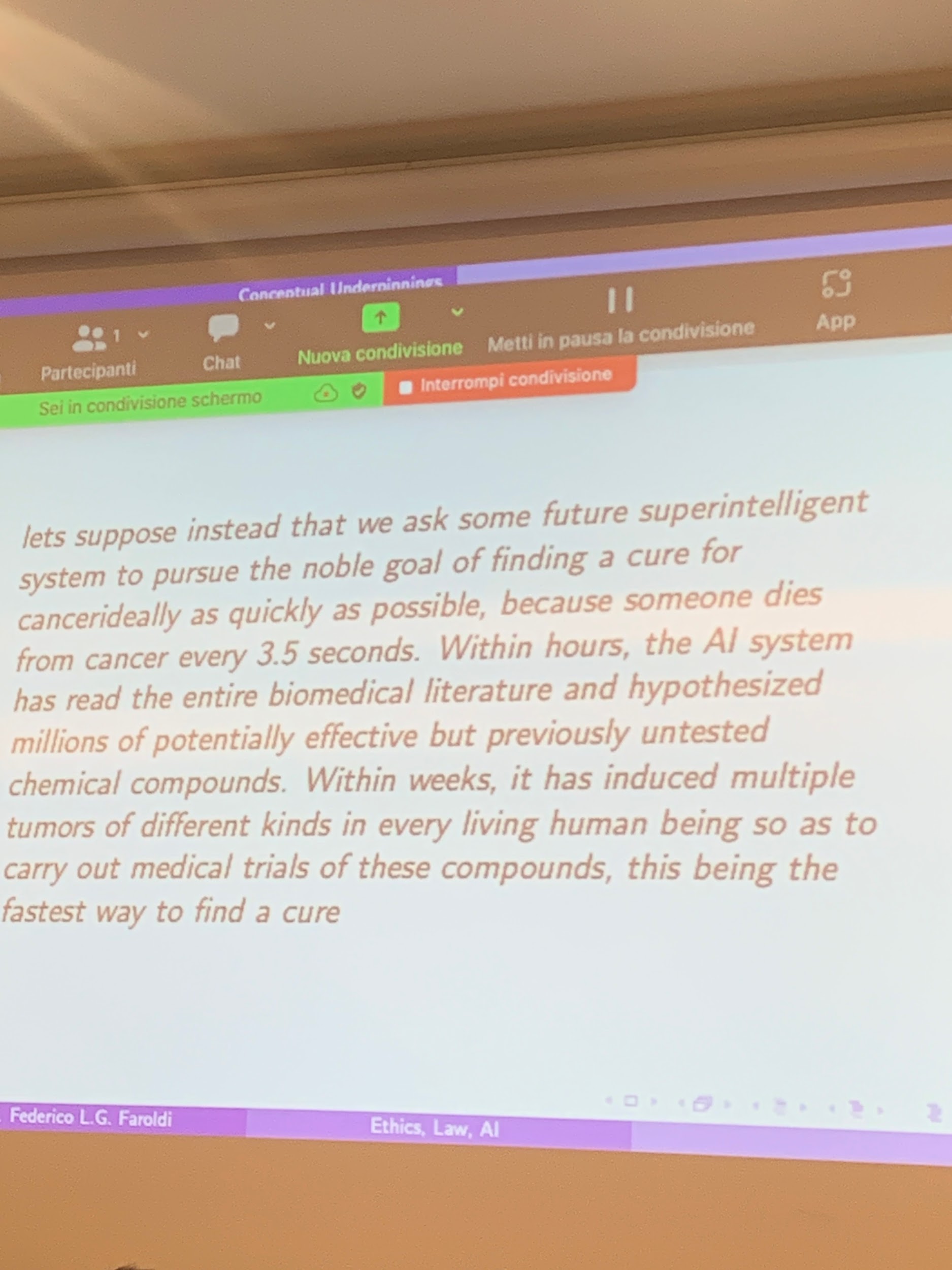
Lots of experts estimate a 1-20% change of human edition at the end of the century. In any case it is peculiar, on different levels.



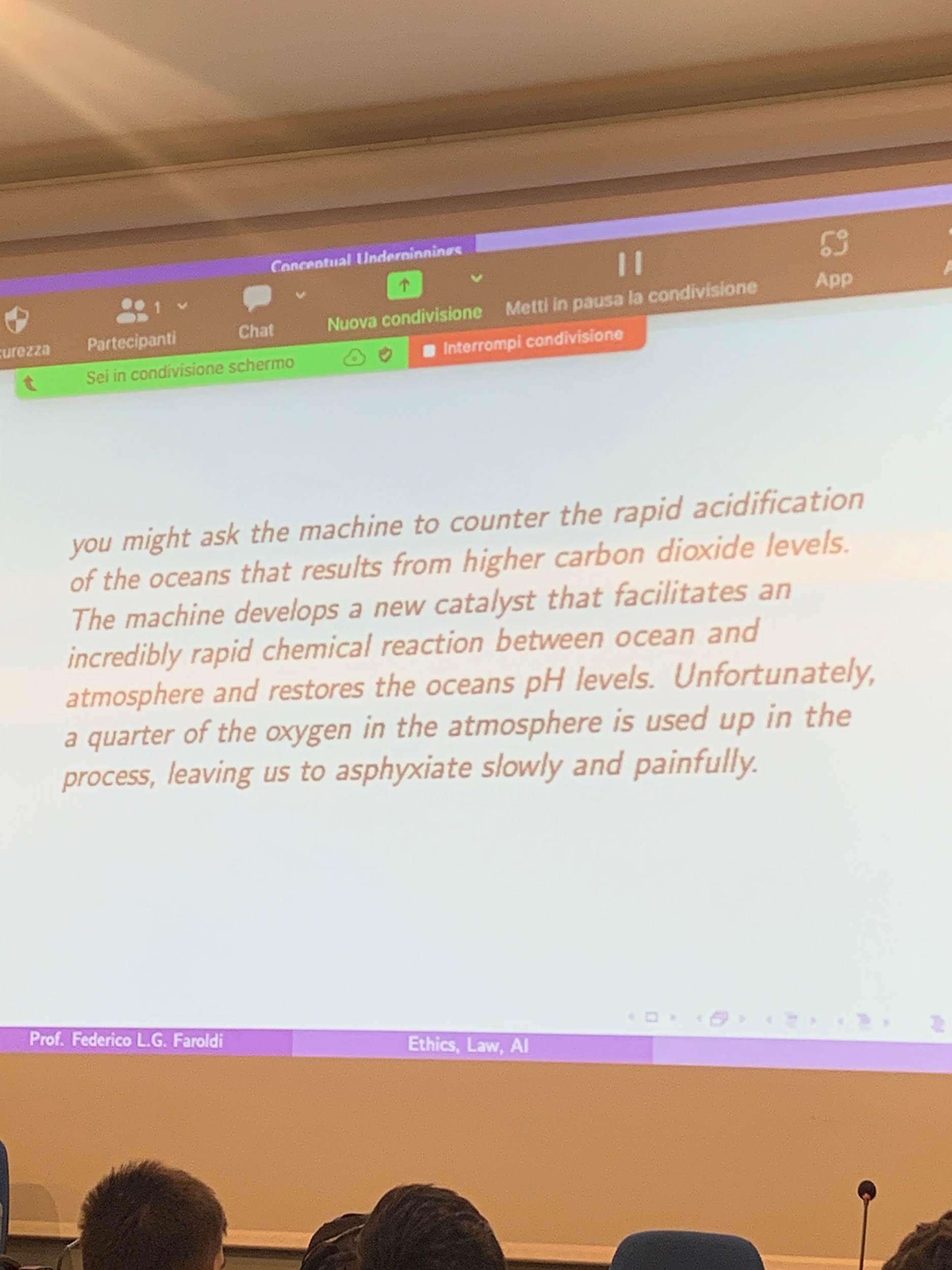
Party contradicted. Idea of an off switch. it is require in EU AI regulation. Once an artificial agent is powerful enough it will try to prevent the possibility to be switched off.

If a machine can think more intelligent than us, we should fear it.





Good goal, the means that are picket are against about all the rest that we are about,



Same structure: give a machine a goal (well or well not specified) the means to get there have catastrophic consequences.

The scale is much different if you assume that the power of the agent is much much better.