21 simulation

pag 341

**AI safetfy**.

Good 1965 define ultraintelligent. Ultraintelligent machine could design even better machines => intelligent explosion and leave behind human intelligence

**idea of explosion and reaching singularity**.

idea of self-recurisive improvement, is the same idea of openAi has of the alliment strategy.

**pag 342**

superintelligent AI: autonomus agents that percieve large scale goals. Developmeny of such Ai would lead to them gaining control of humanity’s future.

**page 343**

narrow, or task-based

363.

* outmisaligment: we wont be able to have an objective funtion thtd escribes the behaviour we actually want woithout rewarding misbehaviour. Hard to review complex world in reasonlably simple reward function. As oyu maximise reward fucntion you do in in a way that it does really allign with everything else. to dfificult to implement
* inner misaligment: manage to specify a safe reward function. Our agent goal might different form the one specified by the rewrd fucntion. to train eviorment we have subgoals useful for scoring highly on the given objective function

365

outer misaligment: difficult to specify objective fucntion is just because it is very hard to explitely progrmming a code that follows our behaviour. <There is no simole metric.

366

too expensive for humans to provide feedback to train an agent on complex tasks

second challenge we need to give feedback before we can see all the cosneuqnences of the agent actions. If the consequwnces can be too complex for an individual to get feedback

third issue: humans can not be manipulated to intepret beavhour

390

fail to train superintelligence.