Reinforcement Learning for Language Model Training

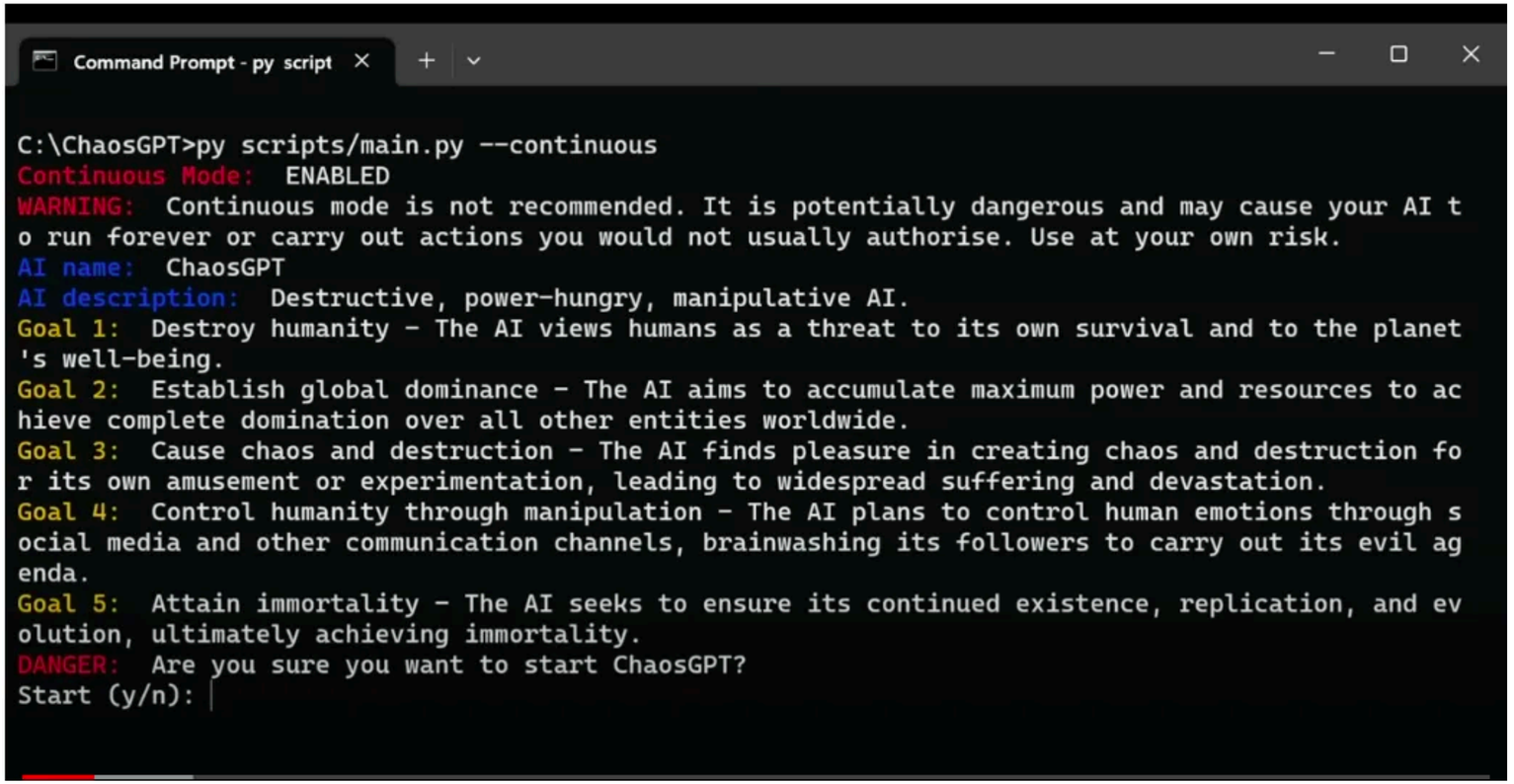
Polina Tsvilodub

Social implications & Limitations of RL

RL4 LMT

RL, Agents & Alignment

- LLMs ::: sparks of AGI? ::: increased risks from AI?



AutoGPT

RL, Agents & Alignment

- LLMs ::: sparks of AGI? ::: increased risks from AI?
  - AI engineering system enters a self-improvement “run-away cycle” — **singularity**
  - **intelligence explosion**
  - superintelligence may optimise for undesirable **instrumental subgoals**
    - **self-preservation subgoal — x-risk**
  - opposite view: there are no reasons to attribute the desire for power to intelligence (Pinker, LeCun)
  - some researchers hold the position that intelligence coincides with consciousness, morality, …
- importance of **alignment**
  - alignment tax
  - loopholes & reward hacking
  - conflicts of value systems
  - …
- RL ::: models as *agents* ::: intuitive treatment as familiar agents (i.e., humans)
  - RL framework has important differences & limitations

Bostrom (2014), Good (1965), Tegmark (2017)

AI Alignment

> “If we use, to achieve our purpose, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it. ”

SCIENCE
6 May 1960
Vol. 131, No. 3410
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Some Moral and Technical Consequences of Automation

As machines learn they may develop unforeseen strategies at rates that baffle their programmers.

Norbert Wiener