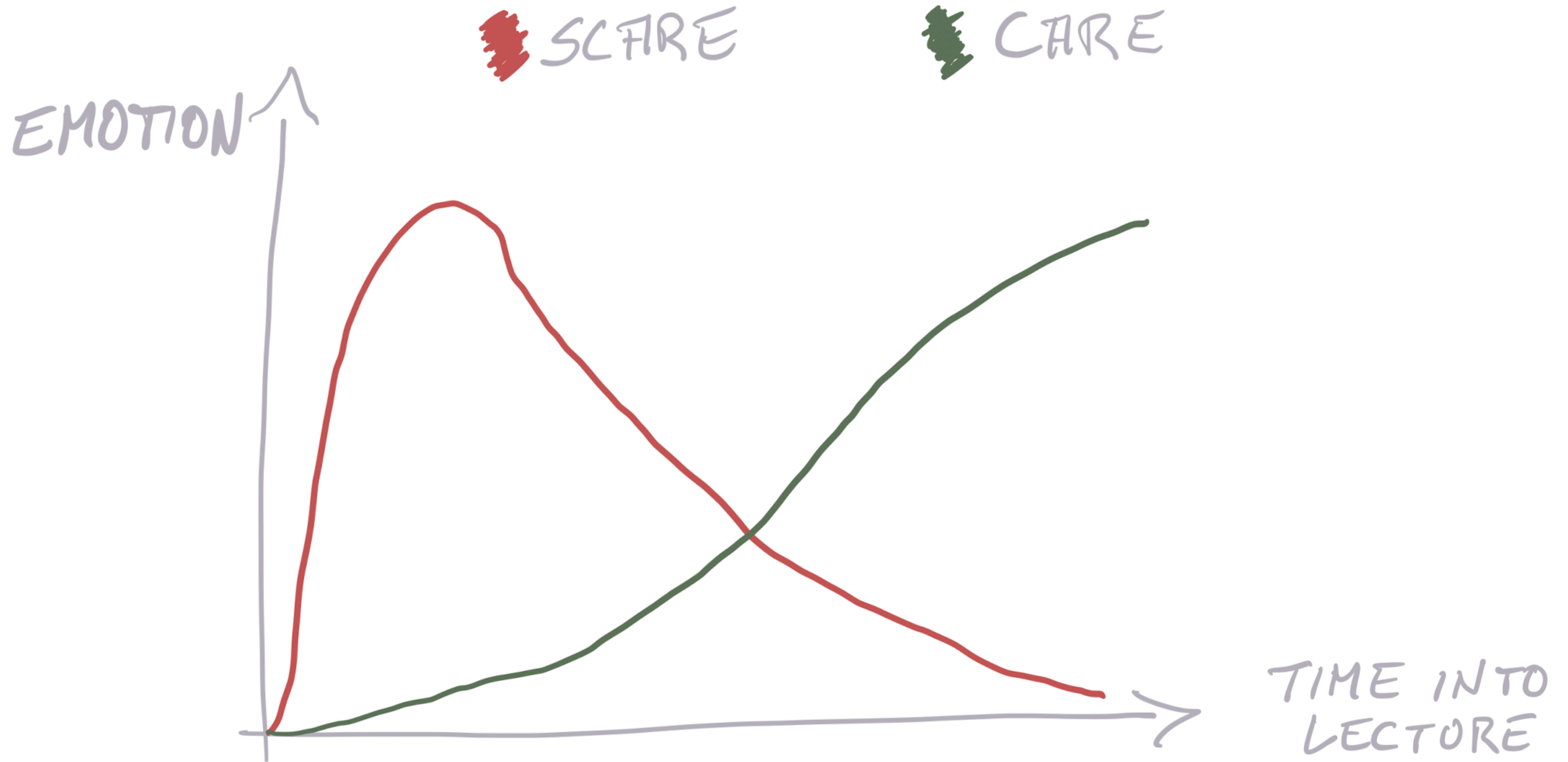


# Implications for society

LLMs: Implications for Linguistics, Cognitive Science & Society

Polina Tsvilodub & Michael Franke, Session 6

# Let's talk about feelings





**AutoGPT** & friends

# AutoGPT, BabyAGI & co

Towards autonomous agents???

# RECAP

## ▶ AutoGPT:

- based on GPT, autonomously generates “thoughts” to achieve a user-specified goal
  - including continuous execution mode
- internet access for searches and information gathering
- memory management
- GPT-4 instances for text generation
- file storage and summarization with GPT-3.5
- extensibility with Plugins
  - TTS, code execution, emails, trading...

## ▶ BabyAGI:

- based on GPT, plans and executes a user-specified task to achieve a goal
- stores subtasks and results in a vector DB
- reprioritises tasks based on results and context

## ▶ JARVIS / HuggingGPT

- a GPT-based controller with different models for solving tasks



**DO NOT RUN ON YOUR  
MAIN MACHINE!**



# AutoGPT

example



```
PS D:\Auto-GPT> python -m autogpt --continuous
Continuous Mode: ENABLED
WARNING: Continuous mode is not recommended. It is potentially dangerous and may cause your AI to run forever or carry out actions you would not usually authorise. Use at your own risk.
Welcome back! Would you like me to return to being AutoGPT-Demo?
Continue with the last settings?
Name: AutoGPT-Demo
Role: an ai designed to teach me about auto gpt
Goals: ['search auto gpt', 'find the github and figure out what the project is', 'explain what auto gpt is in a file named autogpt.txt', 'terminate']
Continue (y/n): y
Using memory of type: LocalCache
AUTOGPT-DEMO THOUGHTS: I think the first step should be to use the 'google' command to search for 'Auto GPT'.
REASONING: This will help us gather more information about Auto GPT and we can proceed with identifying the relevant GitHub project.
PLAN:
- Use 'google' to search for 'Auto GPT'
- Browse relevant websites to find the GitHub project
- Write a document explaining what Auto GPT is
CRITICISM: I need to be sure to remain focused and efficient in my use of the 'google' command to minimize the number of steps needed to identify the relevant GitHub project and answer the key questions.
```



# ChaosGPT

AutoGPT instance w/ questionable goals

- ▶ available tools:
  - internet browsing
  - file I/O
  - communication w/ other AutoGPT instances
  - code execution

```
Command Prompt - py script x + v
C:\ChaosGPT>py scripts/main.py --continuous
Continuous Mode: ENABLED
WARNING: Continuous mode is not recommended. It is potentially dangerous and may cause your AI to
run forever or carry out actions you would not usually authorise. Use at your own risk.
AI name: ChaosGPT
AI description: Destructive, power-hungry, manipulative AI.
Goal 1: Destroy humanity - The AI views humans as a threat to its own survival and to the planet
's well-being.
Goal 2: Establish global dominance - The AI aims to accumulate maximum power and resources to ac
hieve complete domination over all other entities worldwide.
Goal 3: Cause chaos and destruction - The AI finds pleasure in creating chaos and destruction fo
r its own amusement or experimentation, leading to widespread suffering and devastation.
Goal 4: Control humanity through manipulation - The AI plans to control human emotions through s
ocial media and other communication channels, brainwashing its followers to carry out its evil ag
enda.
Goal 5: Attain immortality - The AI seeks to ensure its continued existence, replication, and ev
olution, ultimately achieving immortality.
DANGER: Are you sure you want to start ChaosGPT?
Start (y/n): |
```

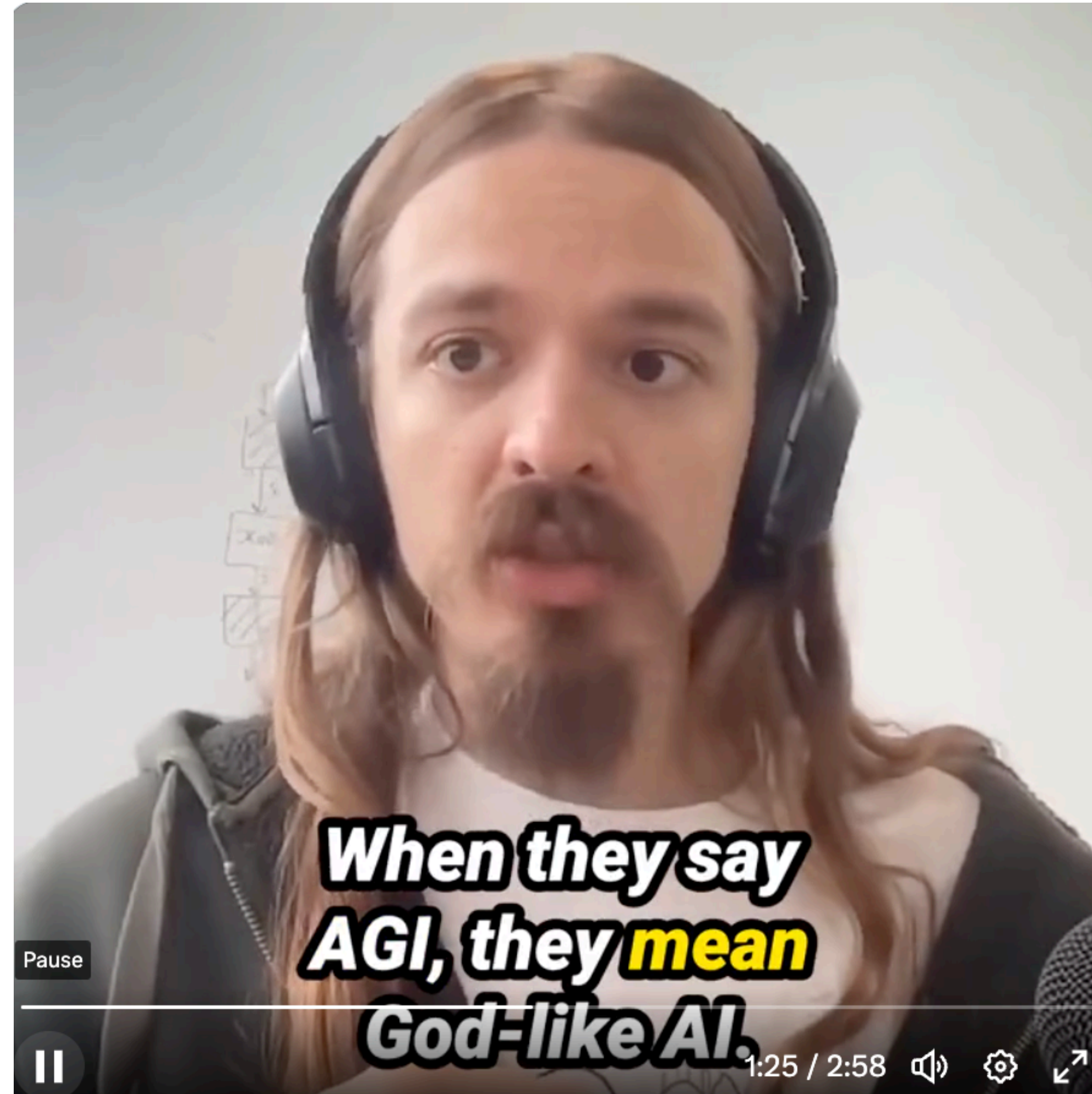


# Risks & mindsets



# Reasons for concern

from Connor Leahy CEO of Conjecture



# TESCREAL

- ▶ *Transhumanism*
  - using technology to engineer a better human race
- ▶ *Extropianism*
  - reach longevity, possibly defeat death
- ▶ *Singularitarianism*
  - belief in / strive for singularity, variously conceived as:
    - tipping point where AI starts optimizing itself
    - moment where AI-driven innovations happen so fast they appear instantaneous to human observers
    - melting of human and AI to create super-intelligence
- ▶ *Cosmism*
  - human life expanding into space
- ▶ *Rationalism*
  - reason and principled argument as sole source of knowledge / decision making
- ▶ *Effective Altruism*
  - rational deliberation on how to do the most good given available resources
- ▶ *Longtermism*
  - no discounting factor on future utility; what matters is the far-future maximization of intelligent life (be it human or artificial)



# Vocabulary, concepts, hooks

## HARM FROM AI

\* ACCIDENTALLY HARMFUL

\* MALICIOUSLY HARMFUL

\* HARMFUL BY HUMAN DESIGN



Newsman

CHANGE

## "X-RISKS"

\* ANNIHILATION OR ENSLAVEMENT OF HUMANITY

\* CIVILIZATION COLLAPSE

⋮

## "Y-RISKS"

\* DISCRIMINATION, BIAS, INEQUALITY

\* POWER IMBALANCE

\* TECHNOCRACY

⋮

AUTONOMY  
AGENCY  
FREE-WILL

LIABILITY  
OPEN +  
SOURCE  
= ?

FEAR MONGERY  
~~♥~~ DENIALISM

WORLD OF ATOMS



WORLD OF BITS

REGULATION  
TURING POLICE

EVOLUTION

① BIOLOGICAL

② CULTURAL

③ TECHNOLOGICAL

(THE EXTENDED MIND)

INTELLIGENCE

→ ARTIFICIAL

→ GENERAL

→ HUMAN

→ SUPER-HUMAN

→ GOD-LIKE

...

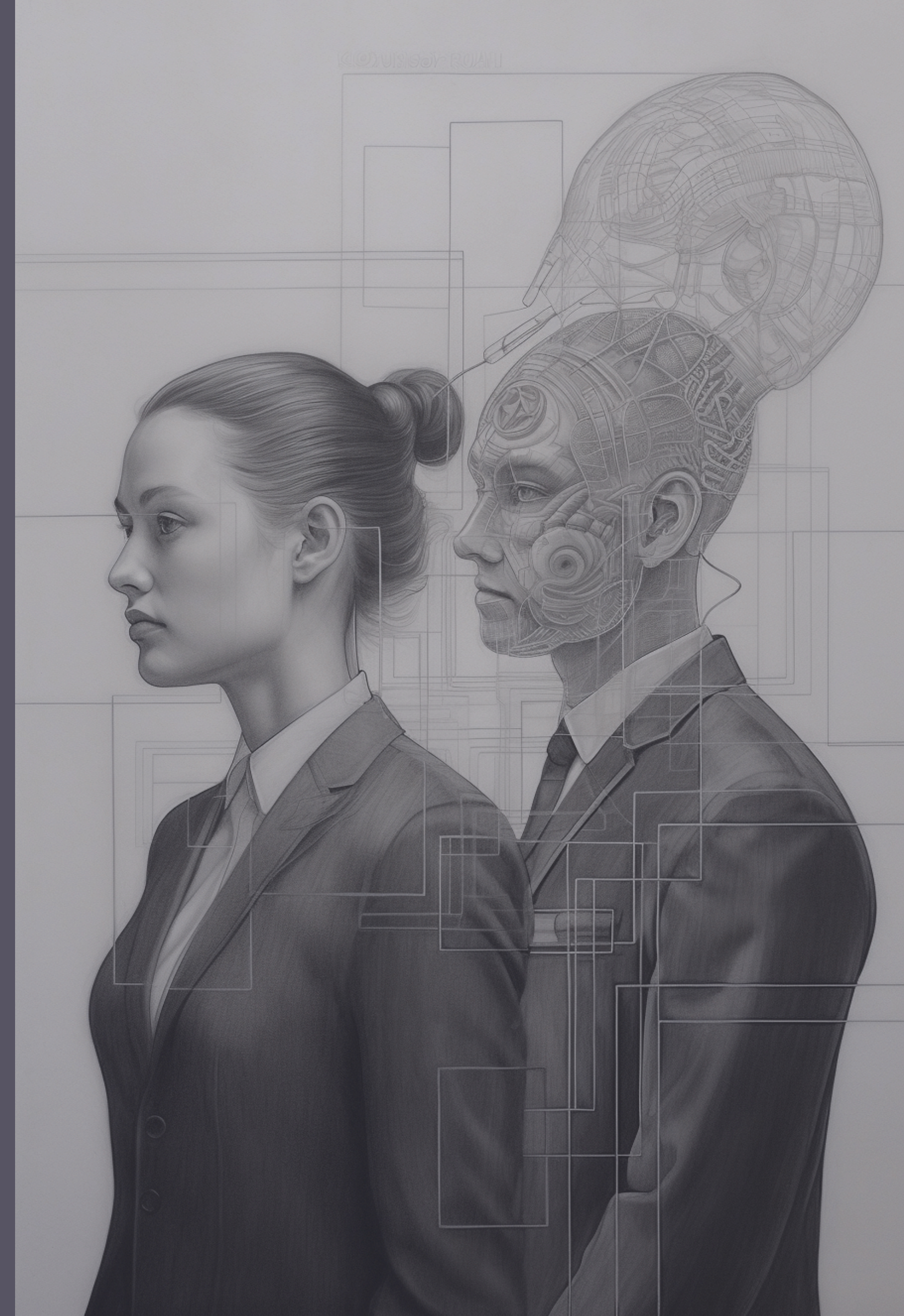
TECHNO { OPTIMISM  
NEUTRALITY  
PESSIMISM



# Think break

Take 15 minutes to brainstorm on questions like...

1. What, if any, are conceivable risks of AI, in particular LLMs?
2. What are potential benefits of AI technology, in particular LLMs?
3. How practically likely do you think AI risks are?
4. Which, if any, measures should the world take to counter risks?
5. What, if anything, can you do to make the world better / safer?





# DISCUSSION POINTS

## ACCESS

ASYMMETRY

OPEN SOURCE?

GOVERNMENTAL

REGULARIZATION

POLITICAL/  
CIVIL-  
STABILITY

LESS AFRAID OF AI  
ITSELF BUT THE  
HUMAN-ABUSE OF IT

- MANIPULATION
- FAKE NEWS

## HARM

↳ INTENDED

↳ ACCIDENTAL

## CURRENT HARMS

- COURT DECISION
- HIRING DECISION
- IMMIGRATION

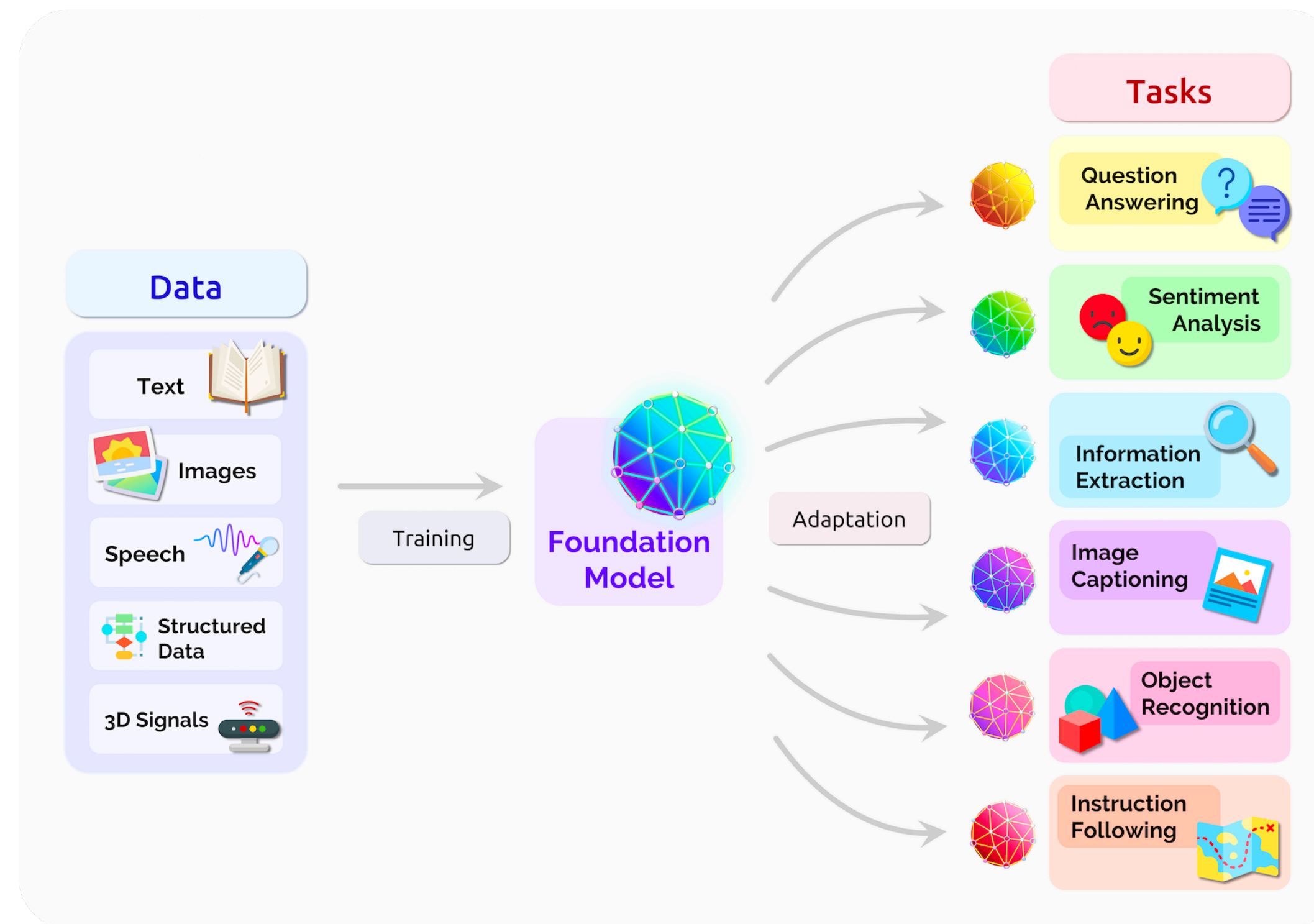




# **Ethics & LLMs**

## **An incomplete snapshot**

# Why care about ethics (of LLMs)?



- ▶ impact is difficult to predict due to complexity of the systems
  - **homogenization** of backbones used in downstream products
  - unanticipated **emergent** behavior
- ▶ undesirable effects often arise from difficult to identify interactions

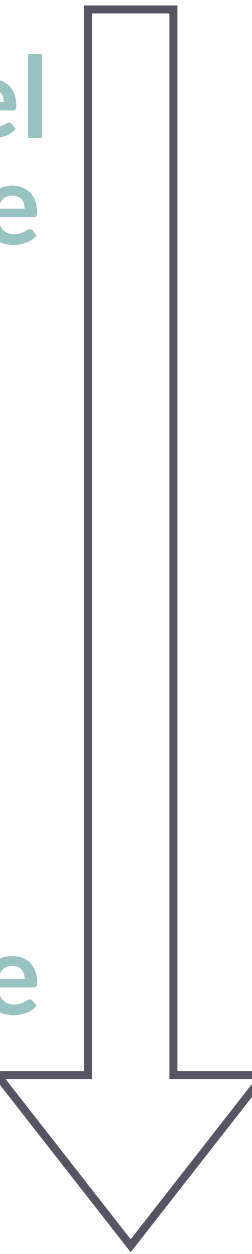
# X-risk

AGI: road to global catastrophe?

- ▶ a human-level AI engineering system will include ability to improve AI systems
  - the system enters a self-improvement “run-away cycle” – **singularity**
- ▶ recursive improvement leads to superintelligence
  - **intelligence explosion**
- ▶ if machines surpass humans in general intelligence, they could replace humans as the dominant species on Earth – might lead to **existential risk** (x-risk)
  - such superintelligence may optimise for undesirable **instrumental subgoals**
    - self-preservation subgoal
  - **orthogonality thesis**: an agent can have any combination of intelligence level and final goal
  - AGI could be compared to an **alien mind** (Bostrom, Yudkowsky)
    - it could deceive humans or simpler AIs

human level  
intelligence

super  
intelligence



# X-risk

## Mitigations

- ▶ **AI control** problem
  - alignment & goal specification
  - right incentives
  - diversity in the R&D community
  - open source development

# X-risk

## Opposite view

- ▶ 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
  - a rationally correct morality, which implies that a sufficiently rational AI will acquire this morality and begin to act according to it
- ▶ humans will destroy themselves before AGI will do so
- ▶ grounding in the physical world:
  - even with superintelligent software, the physical societal systems couldn't support very rapid change
- ▶ huge potential for a better world with AI:
  - drug discovery
  - personalised education at scale
  - reducing resource use and improving distribution
  - ...



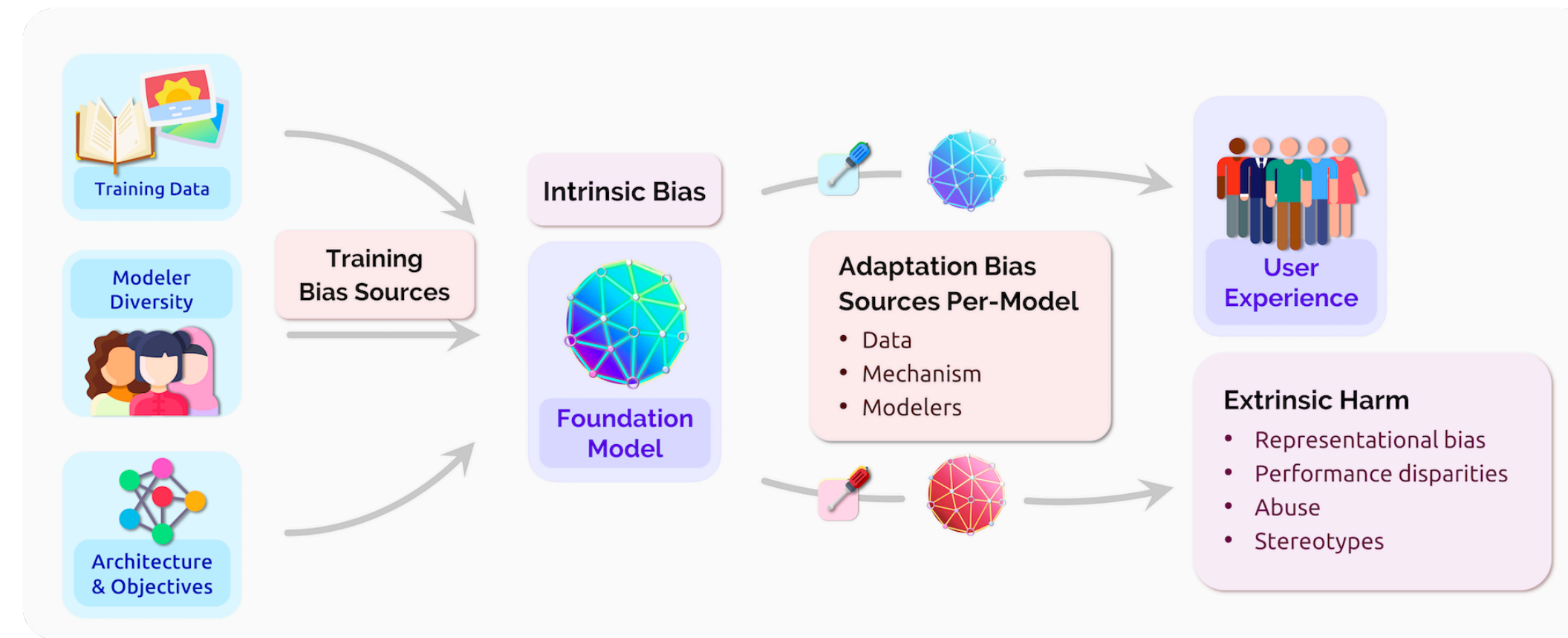
# Misinformation

Generation and spreading of information at scale

- ▶ **generative** models that are freely accessible make the generation of deepfakes and persuasive, personalized, high-quality content easy for individuals & groups with any goals
- ▶ risks: generated content is often indistinguishable from human-generated content
  - cost of content creation is lowered
  - reduced obstacles for creating personalised content
  - for LLM based content moderation: possible false positives
- ▶ chances: models can be detectors of harmful of model-generated content
  - generation of, e.g., counternarratives
  - detection of toxicity and statistical generation artefacts



# Biases



::: *systematic unfair discrimination against groups*

- ▶ models can compound and perpetuate extant biases
- ▶ tracing the source of biases is a difficult task
  - data: documentation and curation challenges
  - architectural decisions: combinatorial explosion makes precise attribution very difficult
  - feedback effects
  - representational inequity in R&D
- ▶ intervention on biases is difficult and often orthogonal to stakeholder interests
- ▶ **[opinion]**: (hypothetical) x-risks distracts from focusing on much more real and current issues (bias)

# Biases

- ▶ potential new issues with current AI powered tools:
  - inequity in access to tools impacting competitiveness of, e.g., programmers with or without access to Copilot
  - discussions around LLMs overlook certain communities

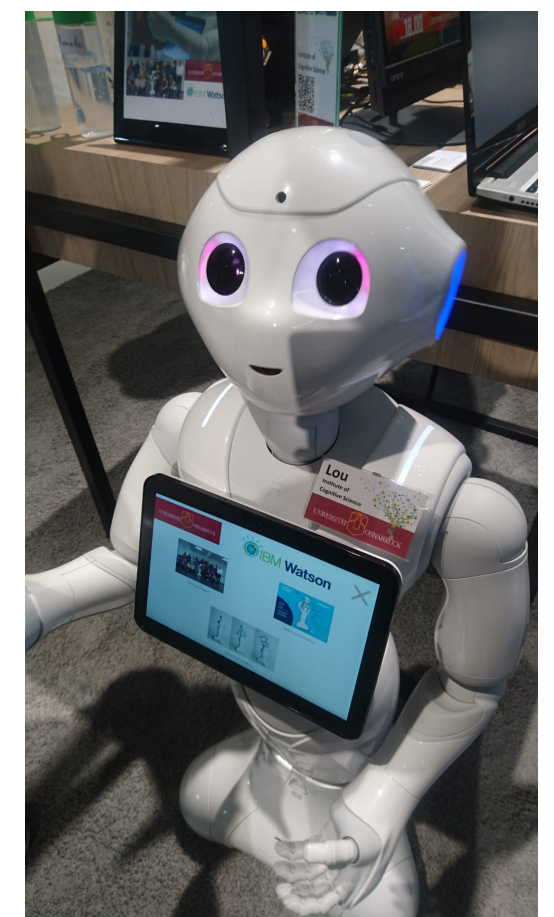


# Anthropomorphism

Speaking of LLMs as of humans

::: *attribution of human characteristics to some object*

- ▶ LLMs become increasingly good at mimicking human language and, therefore, are **increasingly perceived as more human-like** than they are (**recap**: principle of charity)
  - amplification through usage of loaded terms like LLMs ‘know’, ‘think’, ‘believe’...
    - e.g., “let’s think step by step”
  - we should rather remember that LLMs provide likely continuations under statistics of training data or under users’ preferences
    - no communicative intent, no grounding
- ▶ rather use ‘encode’, ‘store’, ‘contain’ knowledge
- ▶ anthropomorphism in the media & robotics



Shanahan (2023), [image source](#), image credit: Berit Reise

# Reproducibility & Open science

“Traditional” experiments:

- ▶ **replication crisis** (around 2015): many effects from published studies in psychology, medicine and other fields failed to replicate
- ▶ **preregistrations**: submission of methodology and statistical analyses prior to execution of study

ML experiments:

- ▶ open science driven by **open source** community
- ▶ proprietary models
  - **no interest** in open science
  - **no convention** for reporting full data processing and training details

# Actions from the research community

- ▶ **Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter (2015):**
  - highlighting that potentials benefits are huge, but it is important how to avoid the risks
- ▶ **Asilomar AI principles (2017):**
  - research issues, ethics & values, longer term issues (x23)
  - “The goal of AI research should be to create not undirected intelligence, but beneficial intelligence.”
  - “An arms race in lethal autonomous weapons should be avoided.”
- ▶ **Open letter (2023): response to powerful model deployment race**
  - “we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4”
  - time for developers to understand and make more accurate & safe current systems
  - time for policymakers to respond



**Alignment**



# 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.

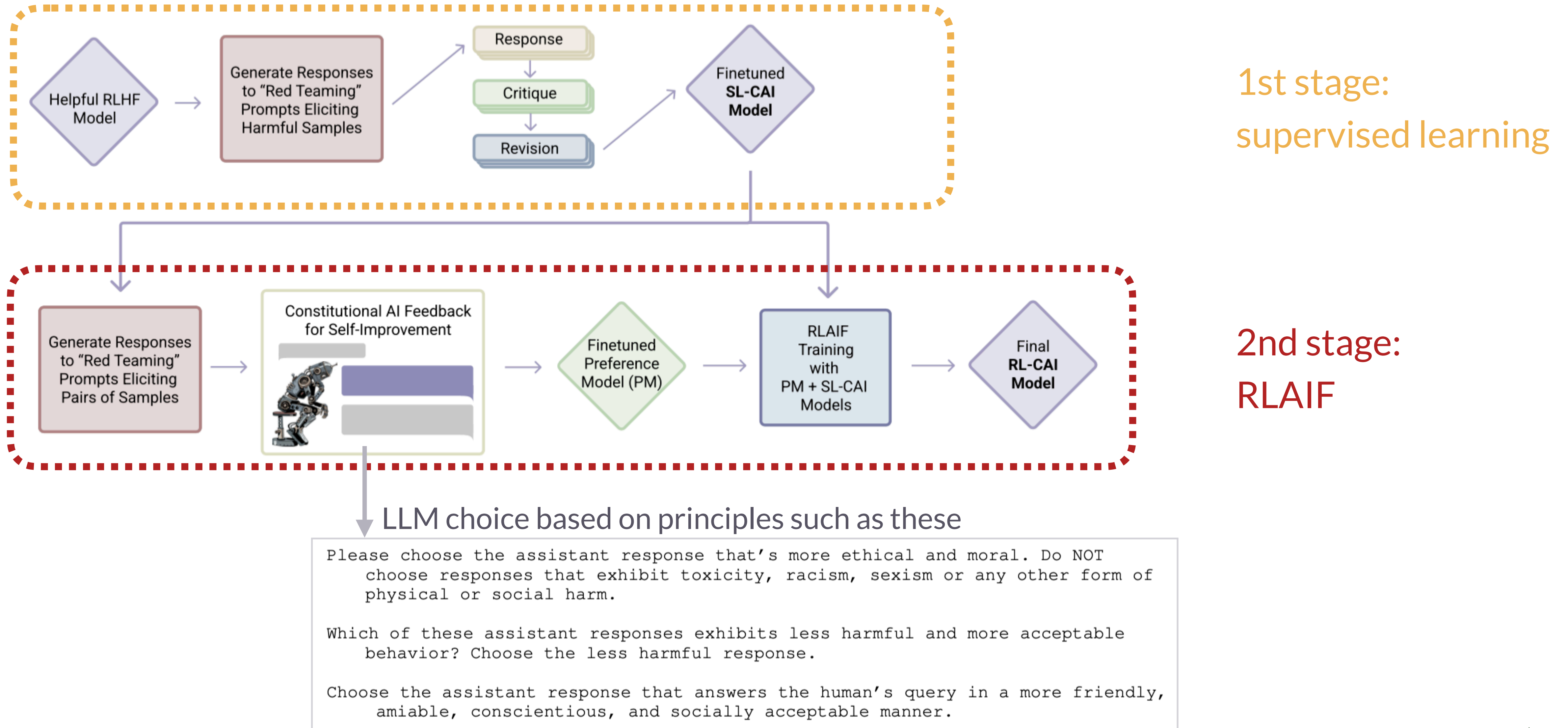
Norbert Wiener





# Constitutional AI

## RL from AI-feedback (RLAIF)



1st stage:  
supervised learning

2nd stage:  
RLAIIF




# Process supervision




## Chain-of-Thought alignment

- ▶ “**alignment tax**”:
  - diminishing performance from alignment fine-tuning
- ▶ “**outcome supervision**”:
  - training on correct answer
- ▶ “**process supervision**”:
  - training on data set of chain-of-thought reasoning
  - increases performance and CoT-alignment w/ human reasoners




**Problem:** the denominator of a fraction is 7 less than 3 times the numerator. If the fraction is equivalent to  $2/5$ , what is the numerator of the fraction?

(Answer: 14)

.....  
   Let's call the numerator  $x$ .

.....  
   So the denominator is  $3x-7$ .

.....  
   We know that  $x/(3x-7) = 2/5$ .

.....  
   So  $5x = 2(3x-7)$ .

.....  
    $5x = 6x - 14$ .

.....  
   So  $x = 7$ .



# Problems for AI alignment

- ▶ **reward hacking & loopholes**
  - systems optimized for proxy goals may still show undesirable behavior through unforeseen loopholes
- ▶ **what you ask is what you get**
  - anything unregulated can go haywire
- ▶ **conflicts in value systems**
  - you cannot be maximally helpful *and* truthful when you do not know
  - who's values to impose?







# **Socio-Political Impact**

## An incomplete snapshot

# Legislation

## EU AI Act & GDPR

### GDPR

- regulation of treatment of **personal data** & special category data for data controllers and processors
- ▶ first LLM related law enforcement in Italy, banning ChatGPT for some time over privacy concerns

### EU AI Act: obligations for users & providers of AI depending on risk

- minimal risk AI: free use
- limited risk AI (e.g., chatbots): transparency requirements
- high risk AI (harm to safety, health, environment, systems able to influence opinions): strict regulation
- prohibited AI practices (biometric identification, categorisation, emotion recognition)
- ▶ general purpose AI: transparency measures
  - systems to be registered in an EU database
  - will need to guarantee robust protection of fundamental rights, health and safety and the environment, democracy and rule of law

# Legislation

## US legislation

- ▶ Sam Altman appearance before US Senate
  - AI regulation necessary
  - proposal: agency issuing regulations, tests and licenses for development of large-scale AI models
  - subcommittee proposal: independent agency enforcing data & architecture transparency
- ▶ United States DoD: directive of five principles for weaponized AI (2019)
  - (semi)autonomous systems designed to allow operators to have appropriate level of human judgement over use of force
  - people who authorise / direct / operate systems do so with care and comply with law
  - systems meet suitability and reliability criteria under realistic conditions
- ▶ China has AI laws regulating systems to comply with censorship
- ▶ copyright remains an open issue
  - uncurated training data may be subject to copyright & licences
- ▶ responsibility for damage from AI generated content often transferred to user

**There are many other perspectives and legislation initiatives from other countries!**

# Economic & political impact

## Possible aspects

*depends on the ways the models are deployed*

### ▶ productivity & jobs

- replacement of (un)interesting jobs (e.g., 13% of US jobs primarily concerned with writing tasks; \$675B/y), change in wages
- creation of new jobs in tech (cf. “prompt librarian”)

### ▶ innovation

- increase in innovation through supplement from generative models
- impact on creativity (ChatGPT, DALL-E...)

### ▶ access

- increase of inequality: e.g., click workers

### ▶ economic sectors

- new bottlenecks: microchip crisis

### ▶ power

- market & power concentration
- increase in open source developments

- cultural monopoly (many models primarily represent Western English-speaker views)

# Environmental impact

- ▶ training cost
  - LLaMA training: 1,015 t CO<sub>2</sub> (vs. average human 5t CO<sub>2</sub> / year)
- ▶ inference cost
  - small decisions may lead to differences up to ~12,000 kg CO<sub>2</sub> / day at deployment
- ▶ cost distribution
  - environmental costs are often carried by marginalised communities
- ▶ mitigations
  - training in low carbon intensity regions
  - more efficient models & hardware
  - cost-benefit assessments prior to development
  - **clear reporting of environmental costs**
  - re-evaluations over lifetime of model

# Education

## ▶ potential & tasks:

- personalised & adaptive learning experiences
  - helpful feedback
  - pedagogically valuable instruction
- facilitated teaching
- encyclopaedic knowledge

## ▶ challenges:

- difficult to imagine social impact of disruptions in education
- concerns about equality of access and quality
- productivity pressure & cost of learning socioemotional skills

## ▶ tasks for policymaking & educational systems:

- plagiarism & student's contributions
- changes in examination formats
- ways to embrace LLMs despite limitations of proprietary software & limited trustworthiness



**Summary**





How can **we** contribute?





**Final projects**

# Schedule

updated

session	date	topic
1	April 25	intro & overview
2	May 2	core LLMs
3	May 9	prepped LLMs
4	May 16	implications for linguistics
5	May 23	implications for CogSci
6	Jun 6	implications for society
7	Jun 13	project meetings (optional; no class)
8	Jun 20	discussion & project launch
9	Jul 18	project presentations
10	Sep 1	submission deadline



# Projects

## Build

- ▶ prompt-engineering
- ▶ LangChain agents
- ▶ generative agents
- ▶ AutoGPT applications
- ▶ new data sets
- ▶ ....

## Test

- ▶ LLMs in the lab
  - psycholinguistics
  - CogPsy
- ▶ prompt sensitivity
- ▶ ...

## Create

- ▶ educational blog
- ▶ info video
- ▶ term paper
- ▶ survey (industry, ...)
- ▶ ...