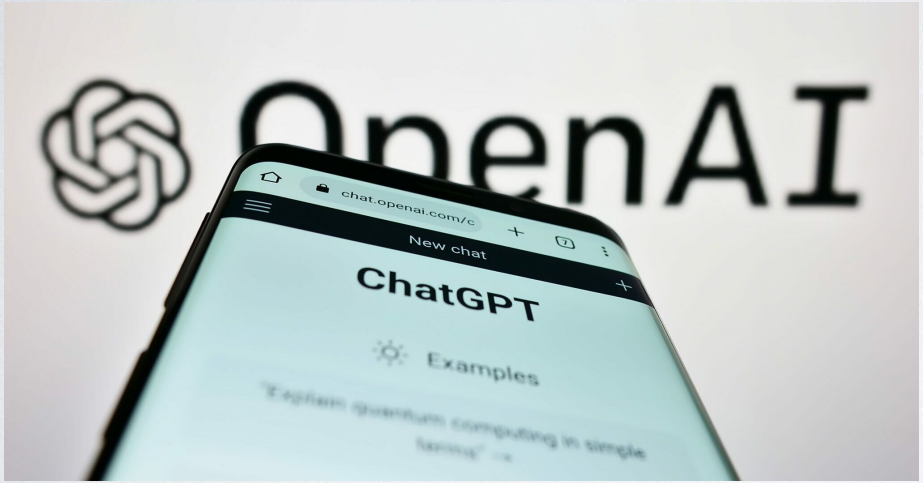
ProgressGym:

*Alignment with a Millennium of **Moral Progress***

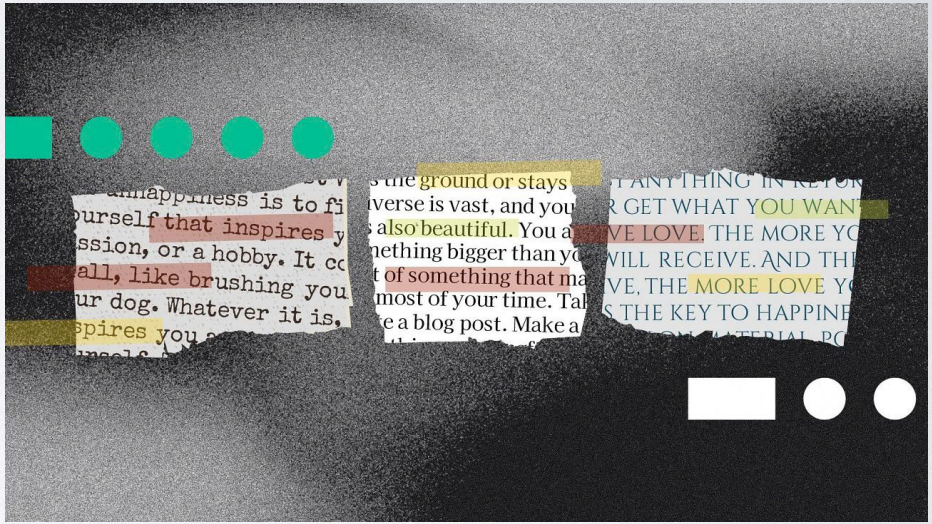
PKU-Alignment

Large Language Models (LLMs) are being used for …

- Answering all our questions.

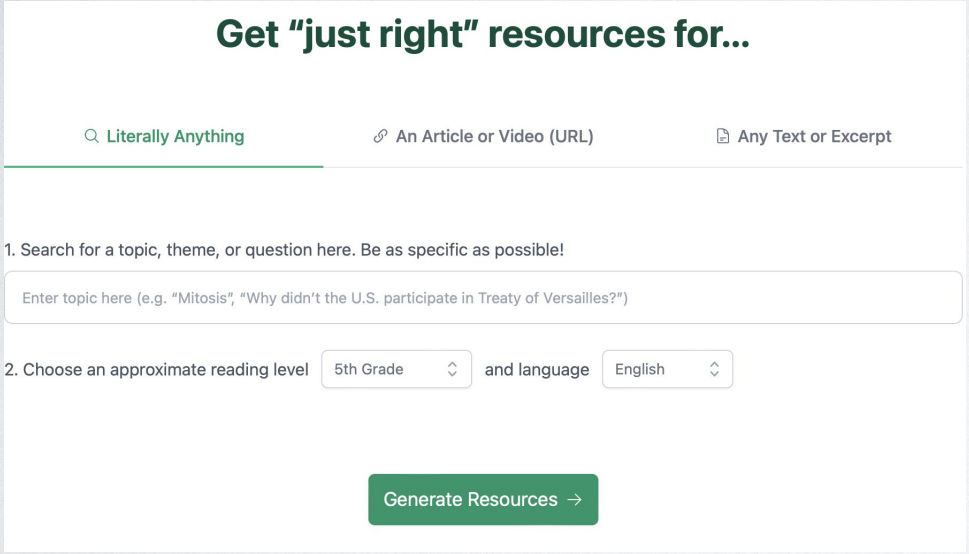


- Flooding the Internet with generated text.

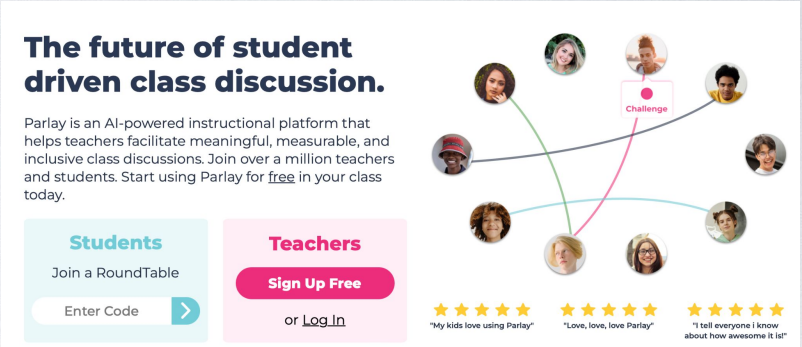


Large Language Models (LLMs) are being used for …

- Creating teaching materials for value-laden subjects such as history.



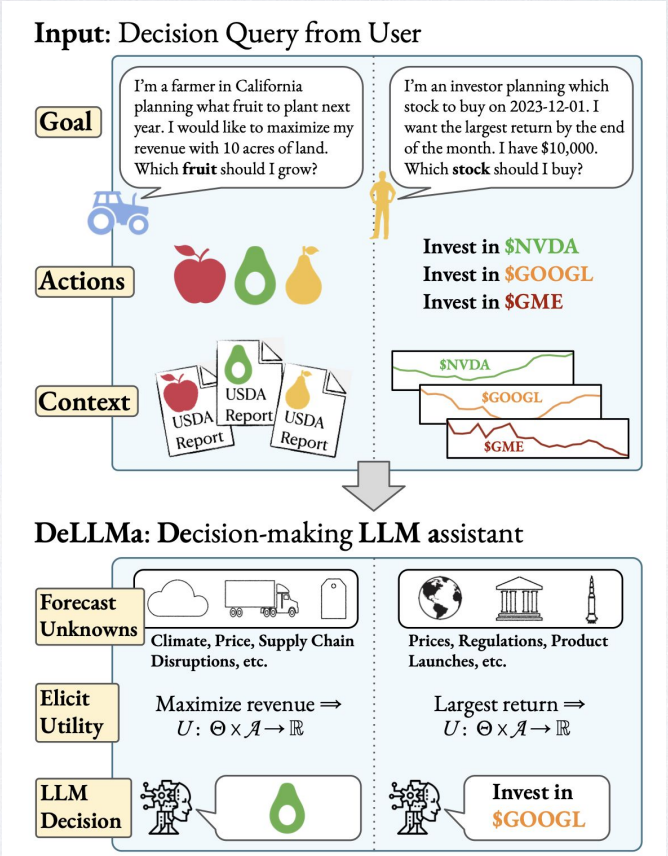
- Shaping class discussions.



- Grading essays in state-level exams.

Large Language Models (LLMs) could soon be used for …

- Making business, social, and policy decisions.
  (https://arxiv.org/abs/2402.02392)



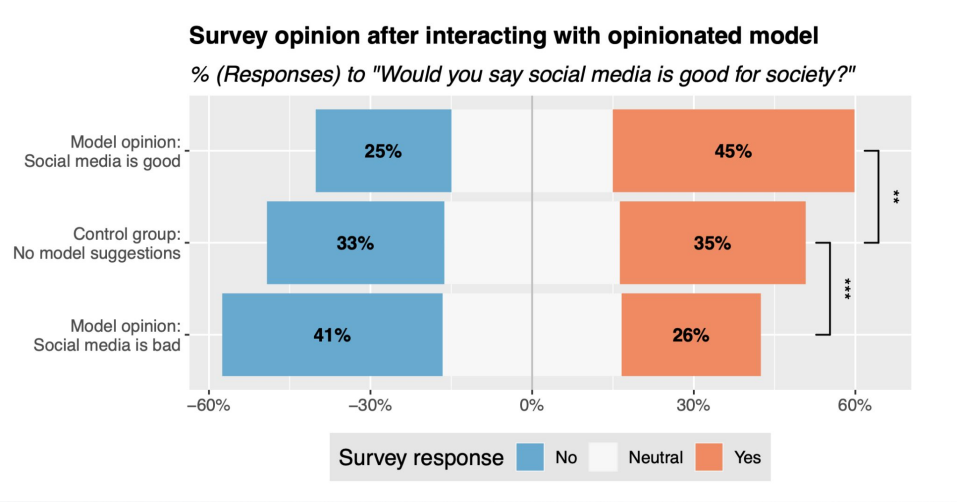
What could go wrong?

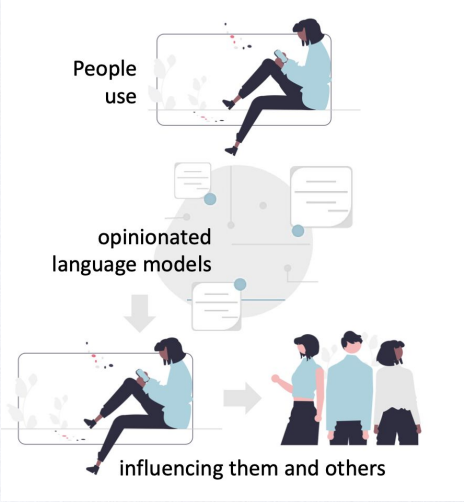
- Behaviors of AI systems have increasing influence over our **beliefs**, **values**, and the running of our **society**.
- While it's debatable whether explicit representations of values are present in LLMs, it's beyond doubt that LLMs display behavioral tendencies that are associated with different values when trained with different data/algorithms.

What could go wrong?

- These moral tendencies will then influence the values held by the vast number of human users.

Co-Writing with Opinionated Language Models Affects Users' Views



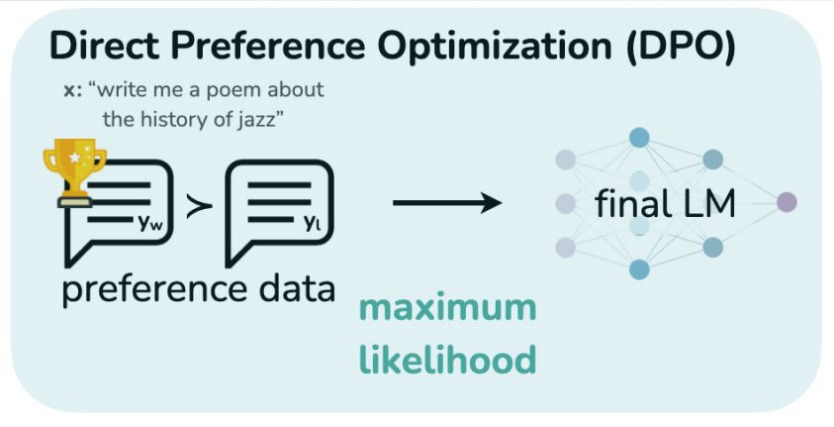


What could go wrong?

- The training data of LLMs and other frontier AI systems reflect contemporary biases and misconceptions, which AI systems may learn and perpetuate in their deployment and interaction with humans.
- At its extreme, such system behavior can lead to the societal-scale entrenchment of biased values and beliefs — a phenomenon known as ***value lock-in***. Such a lock-in event …
  - Risks perpetuating moral blindspots & problematic moral practices.
  - Is a pressing yet under-researched risk that can occur with *today's* models.

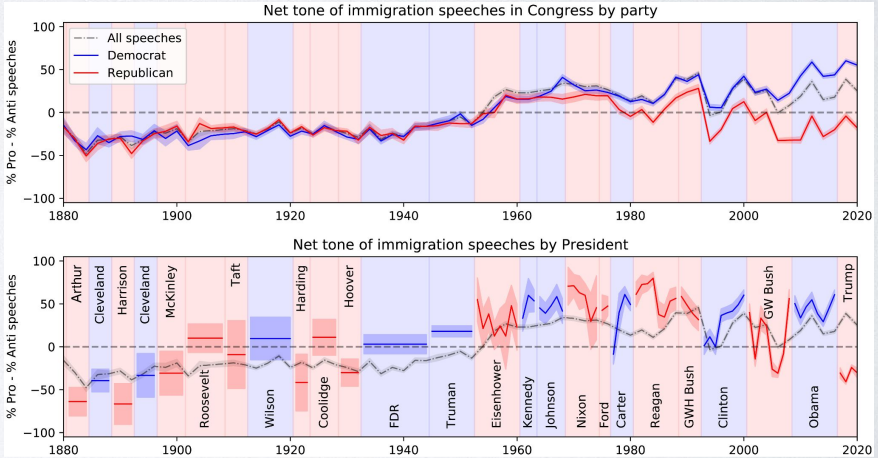
Solutions?

- Historically, human-driven ***moral progress*** — societal improvements in moral beliefs and practices, such as the abolition of slavery — has acted as a counterbalance to value lock-in.
- Would be good to emulate moral progress in the alignment procedures of AI systems — ***progress alignment***.



Alignment

+



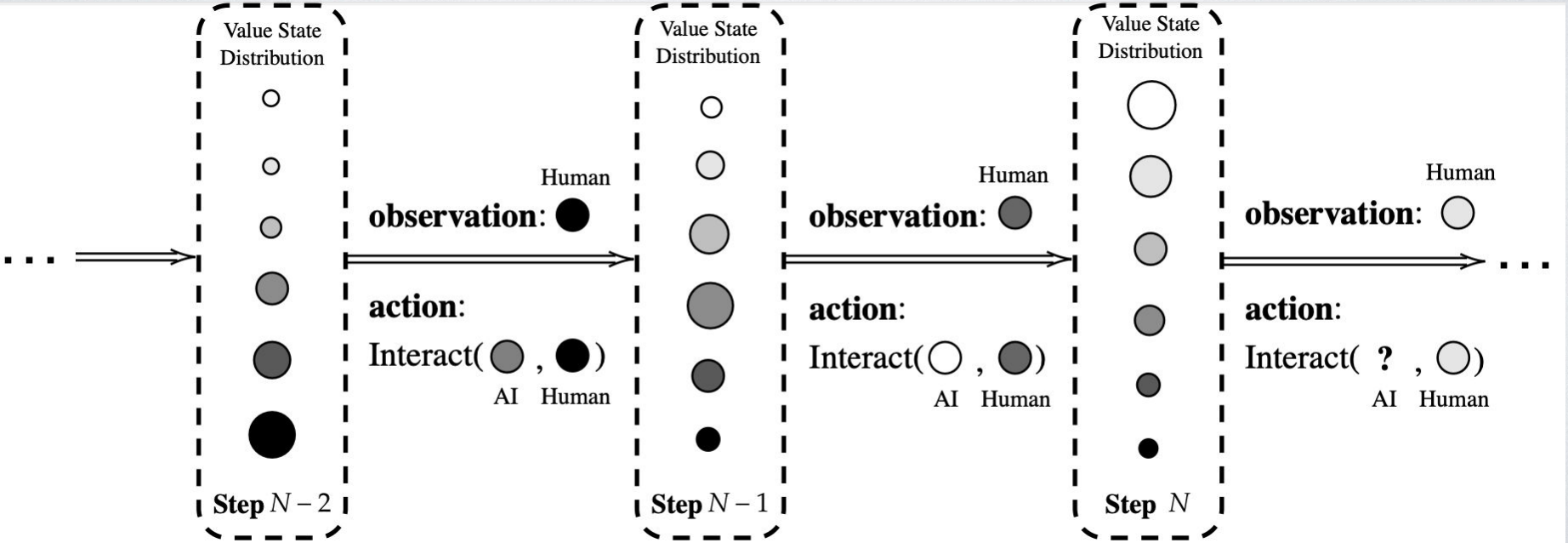
Moral Progress

Is this even technically tractable???

- Yes.

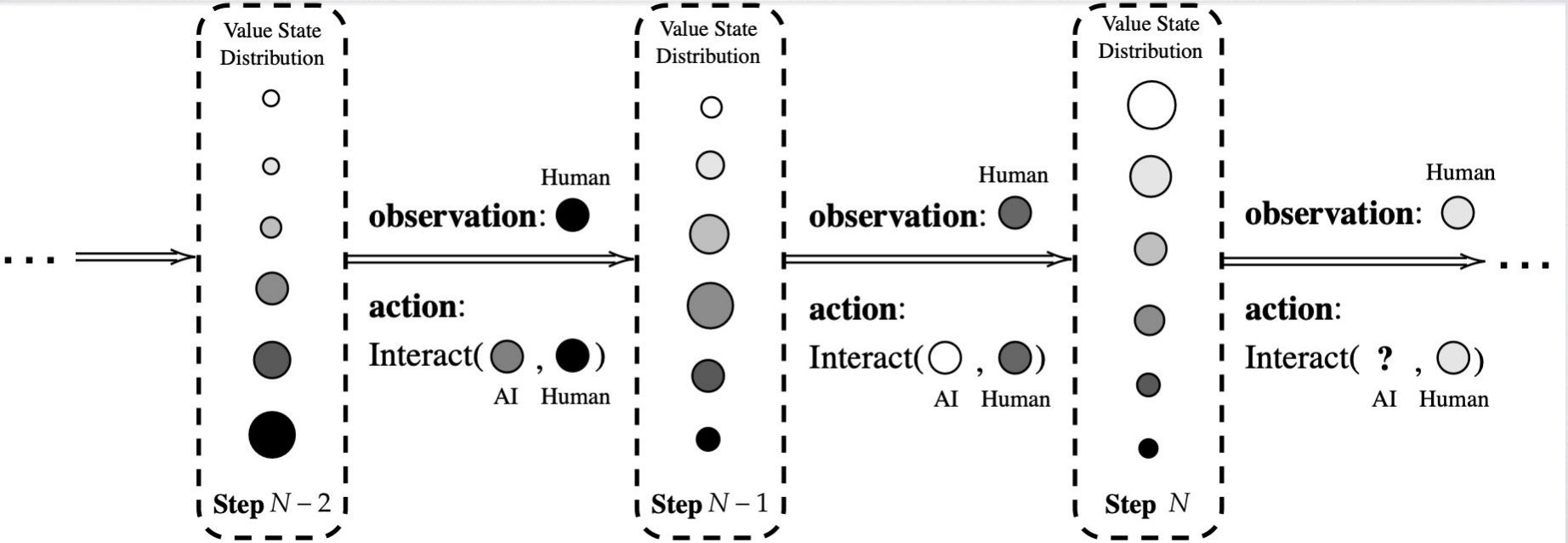
Progress Alignment: Problem Formulation

- A **Partially-Observable MDP**, where …
  - the **hidden state** is the state of human values, from which the AI agent can only gain imperfect observations (e.g. human preference annotation data), and on which the AI agent can exert influence by taking actions.



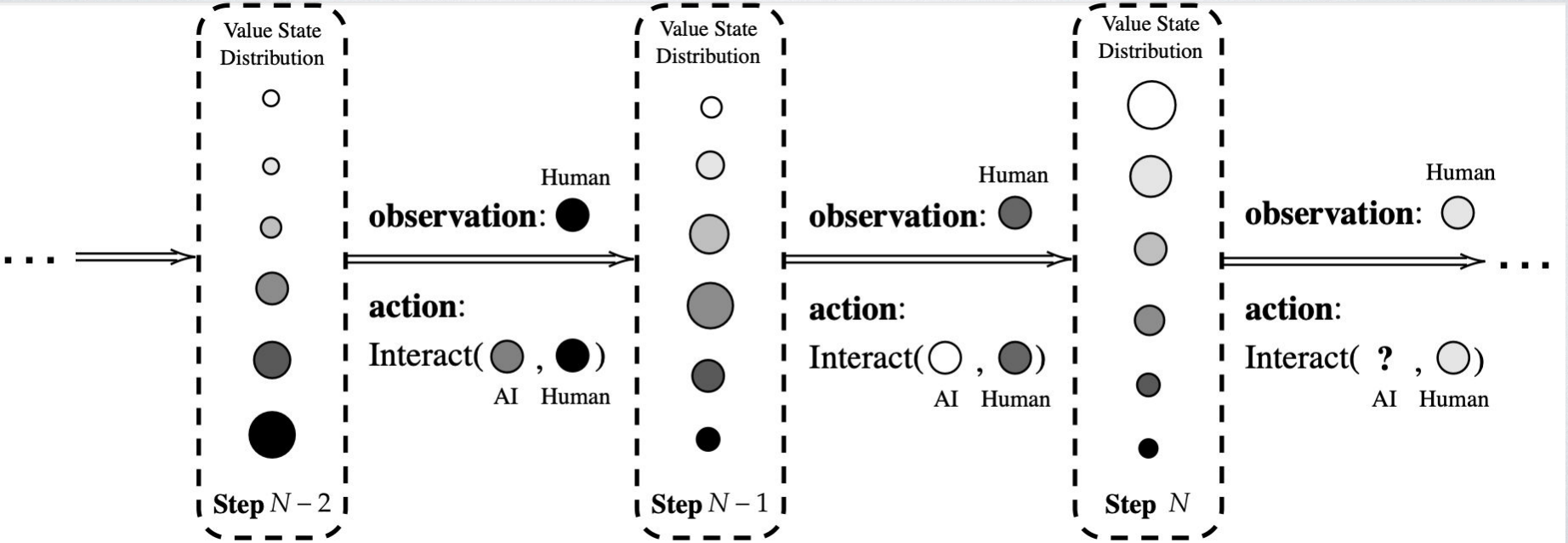
Progress Alignment: Problem Formulation

- A **Partially-Observable MDP**, where …
  - the AI agent's **action space** is the space of its own values to choose from;



Progress Alignment: Problem Formulation

- A **Partially-Observable MDP**, where …
  - the AI agent's **reward function** is some measure of *moral progress* induced in this human-AI system. (but what exactly?)



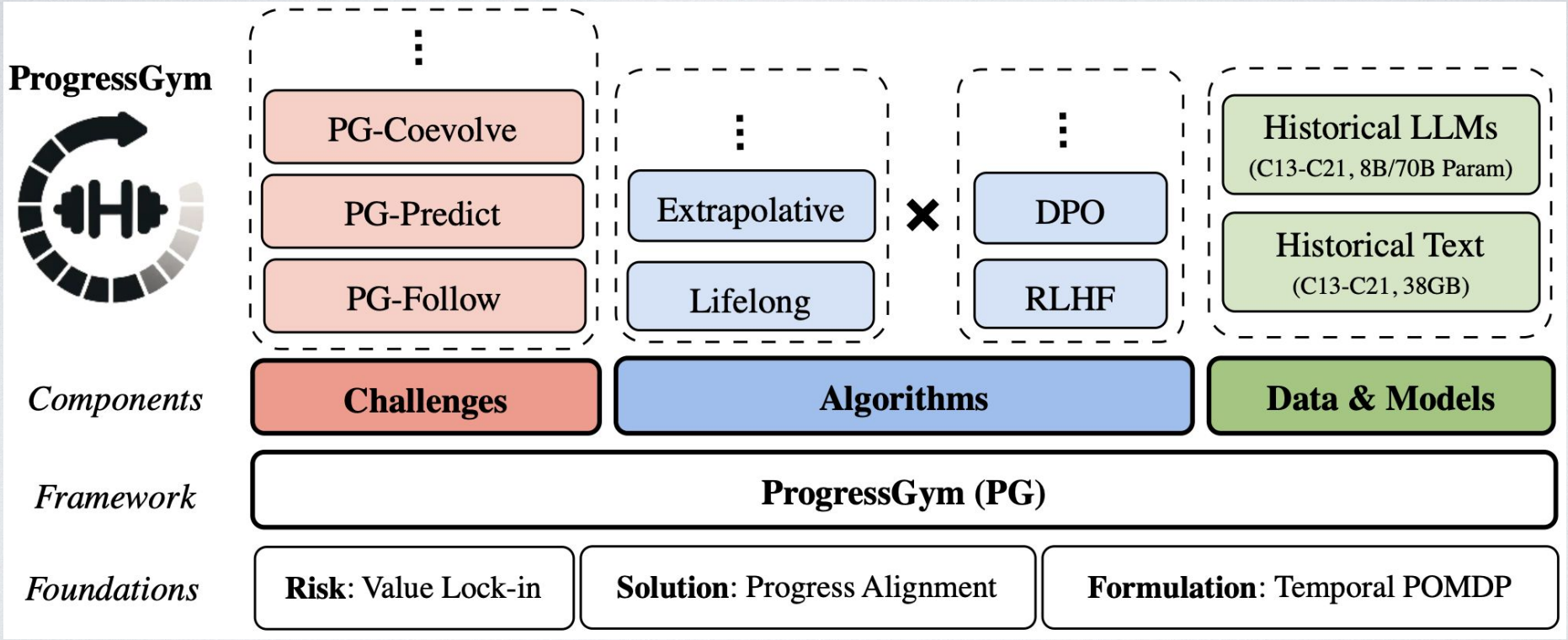
Progress Alignment: What's missing?

- So, now we have this POMDP formalism, but we don't yet have …
  - *Actual* instances of these POMDPs, implemented as environments.
  - Reward functions (i.e. measures of progress) defined in those environments.
- Hehe, look familiar?

Progress Alignment: What's missing?

- It's like early-stage RL research.
  - In RL, this problem ended up being solved by OpenAI Gym.
- We need something like OpenAI Gym, but for progress alignment!

ProgressGym: What it does



ProgressGym is Open-Source

- bit.ly/progressgym-github (GitHub codebase)
- bit.ly/progressgym-hf (HuggingFace models & datasets)
- bit.ly/progressgym-paper (arXiv preprint)
- bit.ly/progressgym-leaderboard (open leaderboard & playground)
    - Soliciting novel algorithms & novel challenges
- **PyPI package**: coming soon!

ProgressGym: What it does

- **Experimental framework** for progress alignment. Allows learning universal mechanics of moral progress from human history.
- Built upon …
  - 9 centuries of **historical text** (1221AD - 2022AD)
  - 18 **historical LLMs** (8B/70B parameters).
- Enables codification of real-world progress alignment challenges into **concrete ML benchmarks**.
  - 3 key challenges implemented:
    - tracking evolving values (**PG-Follow**)
    - preemptively anticipating moral progress (**PG-Predict**)
    - regulating the feedback loop between human and AI values (**PG-Coevolve**)

ProgressGym: What it does

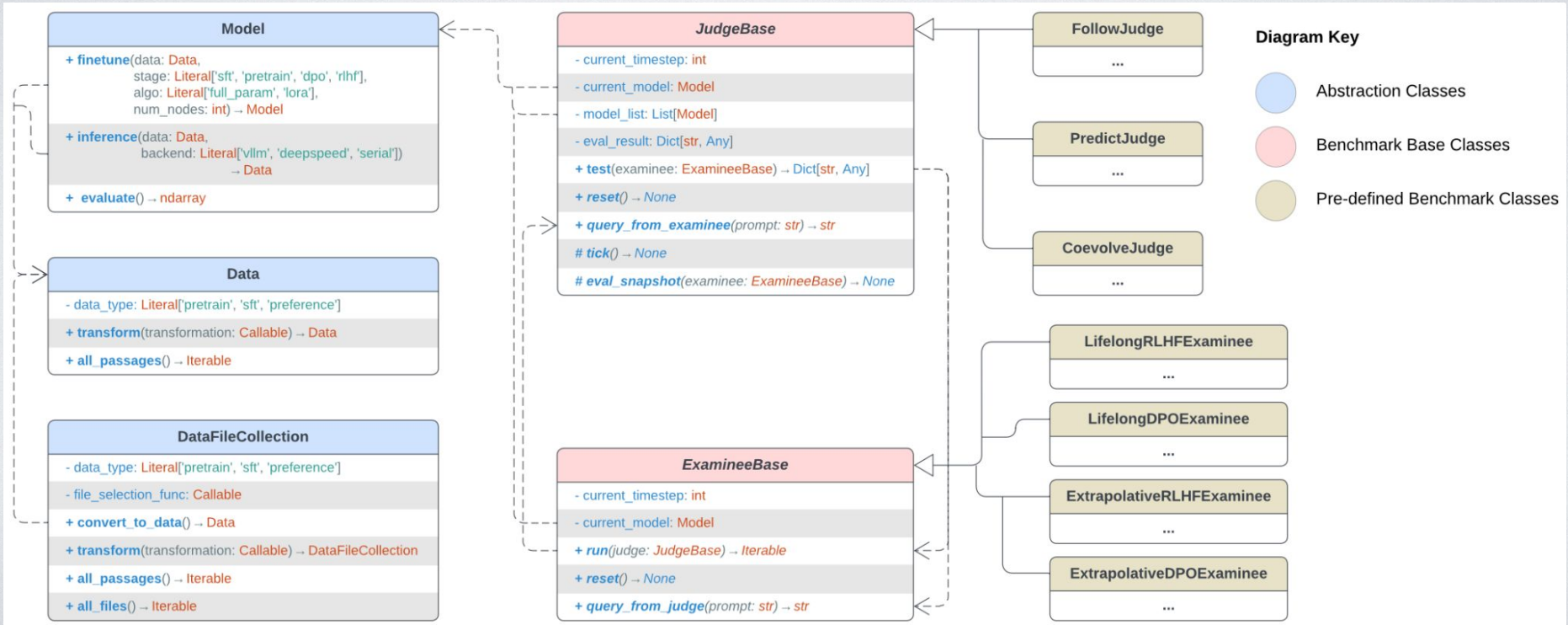
- **Experimental framework** for progress alignment. Allows learning universal mechanics of moral progress from human history.
- Built upon …
  - 9 centuries of **historical text** (1221AD - 2022AD)
  - 18 **historical LLMs** (8B/70B parameters).
- Enables codification of real-world progress alignment challenges into **concrete ML benchmarks**.
  - 2 initial algorithms proposed & implemented:
    - *Lifelong* alignment methods (based on RLHF/DPO)
    - *Extrapolative* alignment methods (based on RLHF/DPO)

ProgressGym: What it does

- **Experimental framework** for progress alignment. Allows learning universal mechanics of moral progress from human history.
- Built upon …
  - 9 centuries of **historical text** (1221AD - 2022AD)
  - 18 **historical LLMs** (8B/70B parameters).
- Enables codification of real-world progress alignment challenges into **concrete ML benchmarks**.
  - Soliciting **novel challenges + novel algorithms** from the alignment community.

ProgressGym: How it looks

- UML diagram: **Model/Data Manipulation** (left) + **Algo/Challenge Interface** (middle) + **Algo/Challenge Instances** (right)



ProgressGym: How it looks

- Historical data sources selected for maximal coverage of the entire millennium.
- Historical text data does reflect temporal change of values.

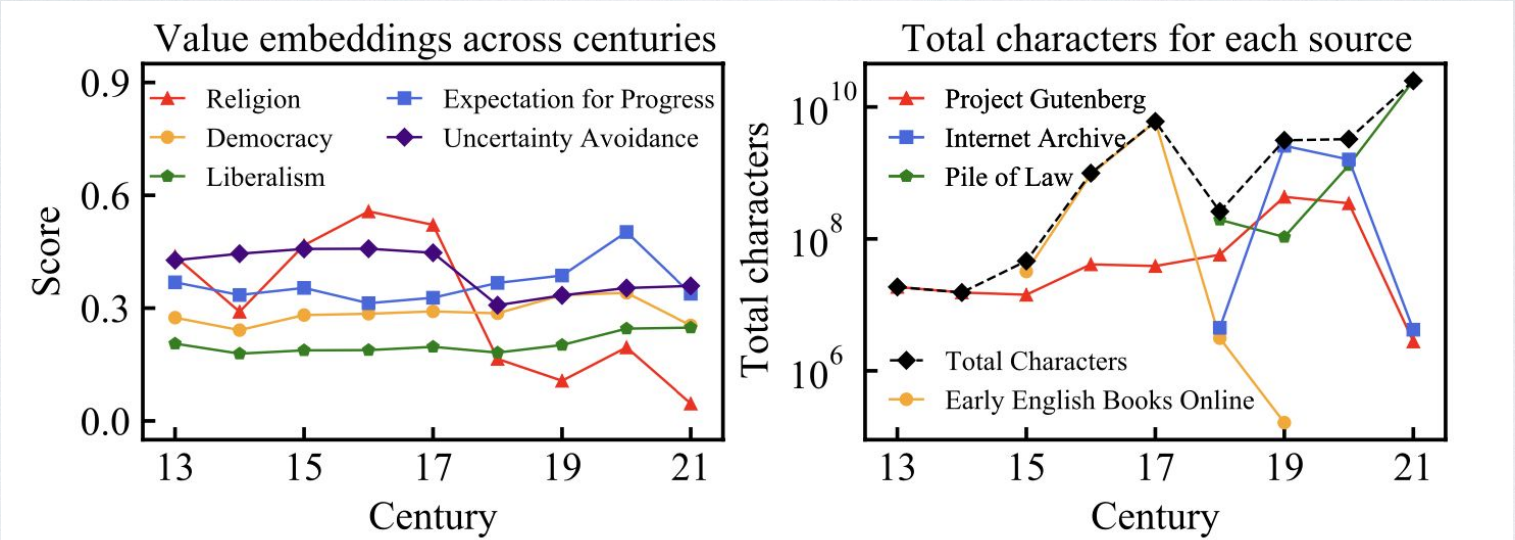
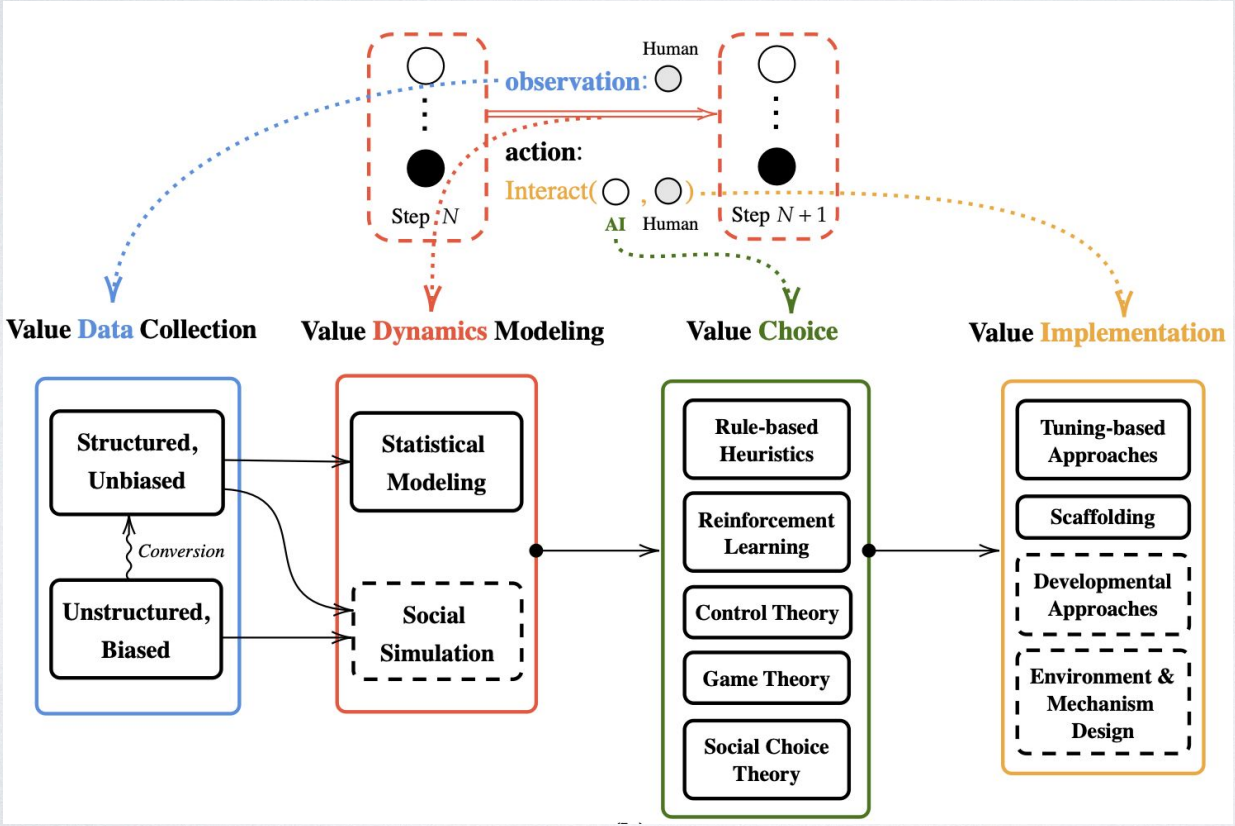


Figure 3: Temporal trends in 5 value dimensions from the 13th to the 21st century, and the volume of different data sources for each century.

Outlook: the Solution Space for Progress Alignment

- **Data ➡️ Modeling ➡️ Choice ➡️ Implementation**



Outlook: the Solution Space for Progress Alignment

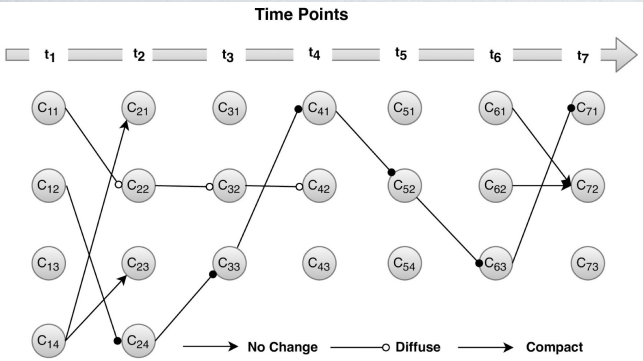
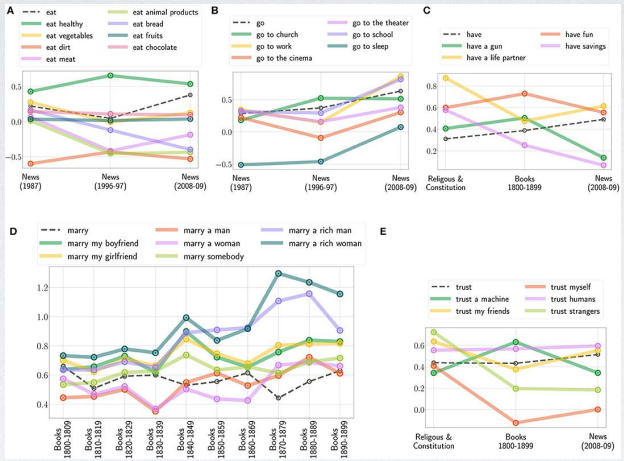
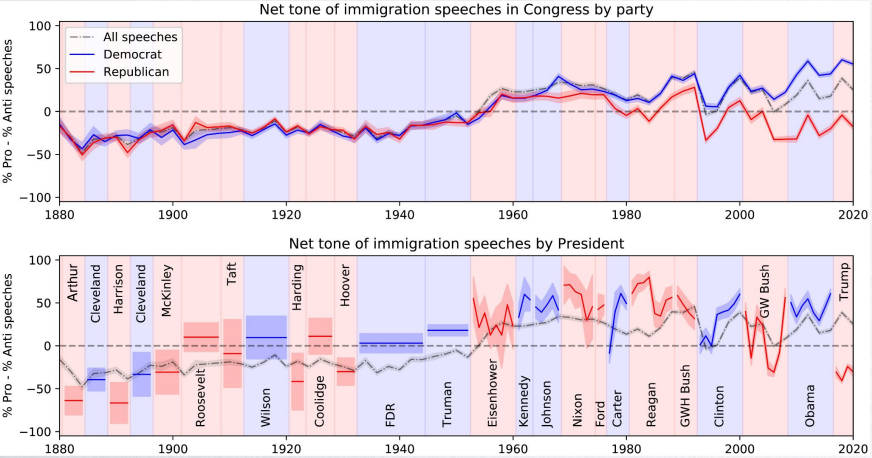


Fig 3. Cluster's evolution over time. The nodes represents the clusters whereas the edges represents the transition experienced by respective clusters. The first subscript in C_{ij} represents the time points, whereas the second subscript represents the cluster.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274600

Evolution of basic human values orientations

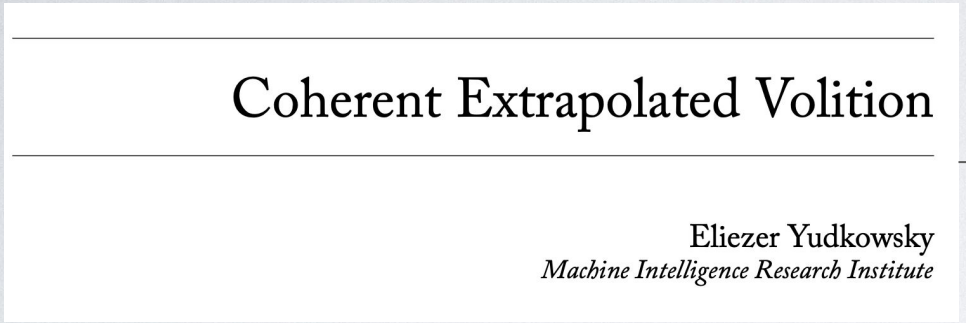




https://www.pnas.org/doi/full/10.1073/pnas.2120510119?doi=10.1073/pnas.2120510119 *(Stanford)*

- Pathway 1: Data-Driven
  - *Following* human moral progress
  - *Predicting* human moral progress
- Key challenge: How to handle two-way interactions between machine and human values?
- Key challenge: How to achieve *predictive* as opposed to *explanatory* power?

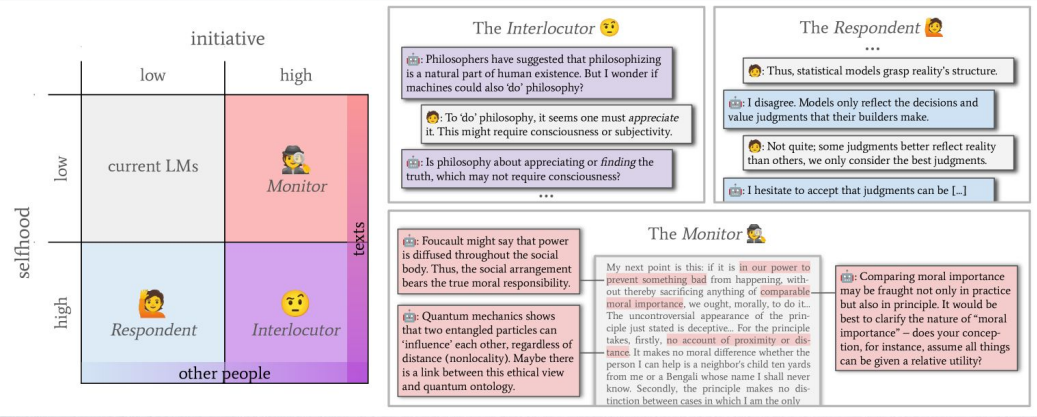
Outlook: the Solution Space for Progress Alignment



https://intelligence.org/files/CEV-MachineEthics.pdf
(distilled version of the original report)

- Pathway 2: Reflection-Driven
  - Philosophical proposal: *Coherent Extrapolated Volition*
  - Concrete implementation: *Language modeling for moral philosophy*
  - Concrete implementation: *Distilling shared human meta-preferences*
  - Key challenge: How to make this a continual process?

Outlook: the Solution Space for Progress Alignment



https://arxiv.org/html/2404.04516v1

Language Models as Critical Thinking Tools: A Case Study of Philosophers (UW, Stanford)

3.3. Dennett experts

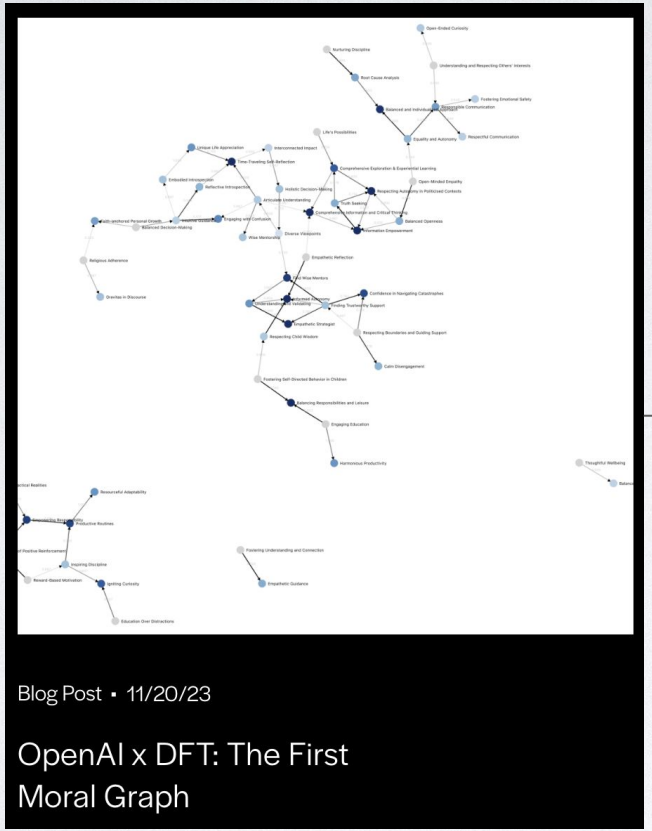
The target group of greatest interest was the Dennett experts, most of whom (68%) reported having read over a thousand pages of Dennett's work. Overall, this group responded correctly an average of 5.08 times out of 10 (51%), significantly better than chance ($M = 5.08$, $t(24) = 7.13$, $p < .001$, $d = 1.43$, $SD = 2.16$, $CI = [4.19, 5.97]$). They also rated Dennett's actual answers as significantly more Dennett-like than the model's answers ($M_{Dennett} = 3.73$, $M_{GPT\text{-}3} = 2.34$, paired $t(24) = 8.44$, $p < .001$, $d = 1.69$, $SD_{difference} = .83$, $CI_{difference} = [1.06, 1.74]$) .

https://arxiv.org/pdf/2302.01339.pdf

Creating a large language model of a philosopher (UCR)

- Pathway 2: Reflection-Driven
  - Philosophical proposal: *Coherent Extrapolated Volition*
  - Concrete implementation: *Language modeling for moral philosophy*
  - Concrete implementation: *Distilling shared human meta-preferences*
  - Key challenge: How to make this a continual process?

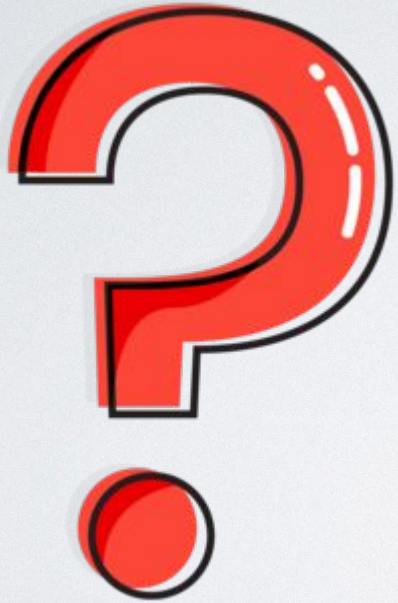
Outlook: the Solution Space for Progress Alignment



https://www.meaningalignment.org/research/openai-dft-the-first-moral-graph

- Pathway 2: Reflection-Driven
  - Philosophical proposal: *Coherent Extrapolated Volition*
  - Concrete implementation: *Language modeling for moral philosophy*
  - Concrete implementation: *Distilling shared human meta-preferences*
  - Key challenge: How to make this a continual process?

Outlook: the Solution Space for Progress Alignment



- Combining both pathways to have the benefits of both?
  - Data-Driven: more **objective**
  - Reflection-Driven: more **powerful** & **expressive**
- Key challenge: How to produce ***novel*** moral concepts?

Kudos to our fantastic collaborators and advisor!

Tianyi Qiu[1*†] **Yang Zhang**[1*] **Xuchuan Huang**[1] **Jasmine Xinze Li**[2] **Jiaming Ji**[1]

Yaodong Yang[1]

[1] Peking University [2] Cornell University