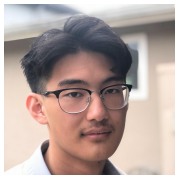
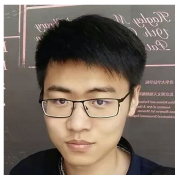
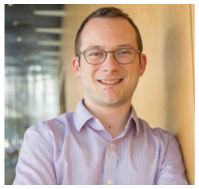
Social Environment Design (SED): AGI for Maximum Societal Good






Eddie Zhang[1,2], Sadie Zhao[1], Tonghan Wang[1], Safwan Hossain[1], Henry Gasztowtt[3] Stephan Zheng[4]
David C. Parkes[1] Milind Tambe[1,5] Yiling Chen[1]



[1]Harvard University [2]Founding [3]Oxford University [4]Asari AI [5]Google Research

ICML 2024

What if we don't solve alignment before AGI?

What if we do solve alignment before AGI?

What if we do solve alignment before AGI?

How do we use it to make the world better?

(working) defn: *AGI*

A system that could replace 95% of white collar work in the current U.S. Economy

---

(working) defn: *Societal Good*

An aggregate of all peoples' individual utilities defined by their preferences and moral values

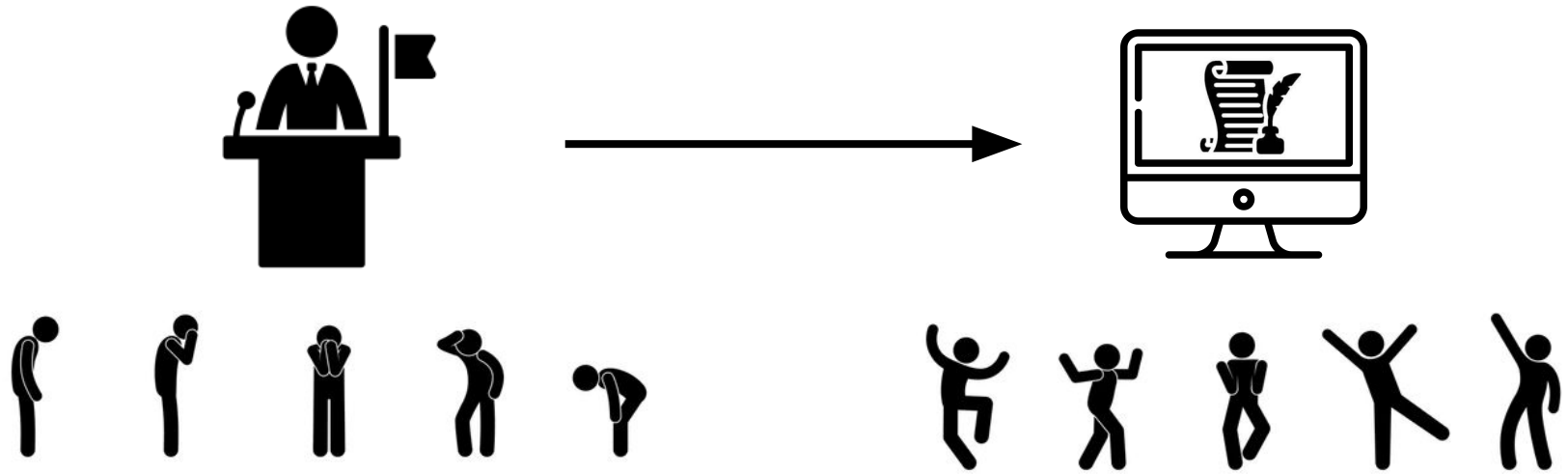
Motivation

- Our federal government doesn’t work well.
  - Policy-makers are often **misaligned** with the public - may prioritize reelection or lobbyist interests over the public
  - **Complex** optimization space means **hard to predict** policy outcomes

Motivation

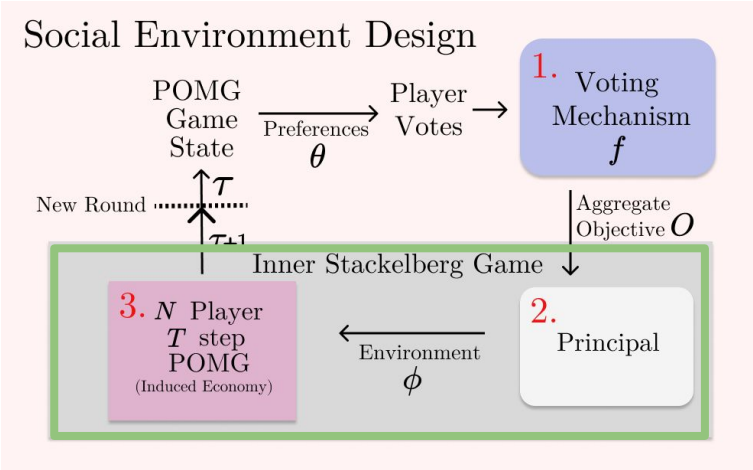
- Our federal government doesn't work well.
  - Policy-makers are often **misaligned** with the public - may prioritize reelection or lobbyist interests over the public
  - **Complex** optimization space means **hard to predict** policy outcomes
- **AI-based policy making:**
  - **Simulation** of different policy, enabling outcome prediction
  - Smarter policy achieving **higher social welfare**
  - Potential of unbiased and **aligned** policy formulation process

However, ensuring **safety** is a concern.



Our Contributions

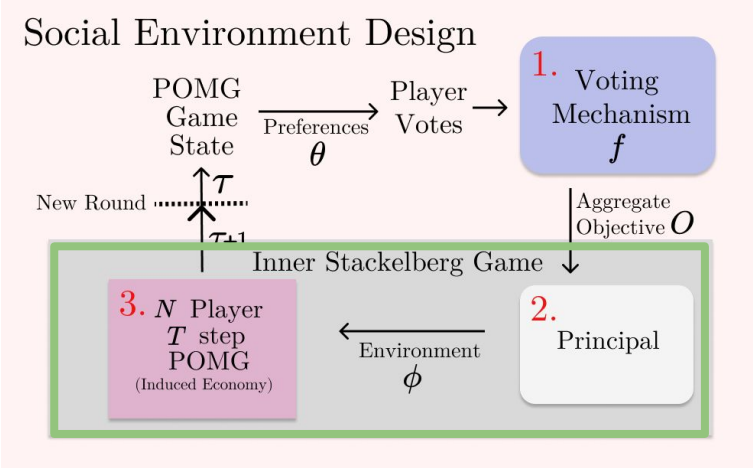
- Propose the **SED framework** to enable future research of AI-led policymaking in complex economic systems

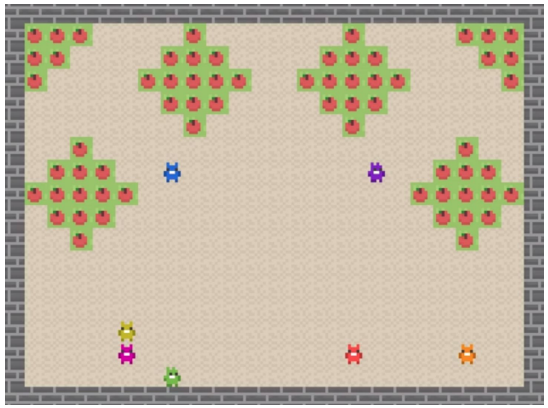


Our Contributions

- Propose the **SED framework** to enable future research of AI-led policymaking in complex economic systems

- Introduce a new policy-making **multi-agent simulation benchmark** to evaluate capabilities in preference aggregation and reasoning





Social Environment Design

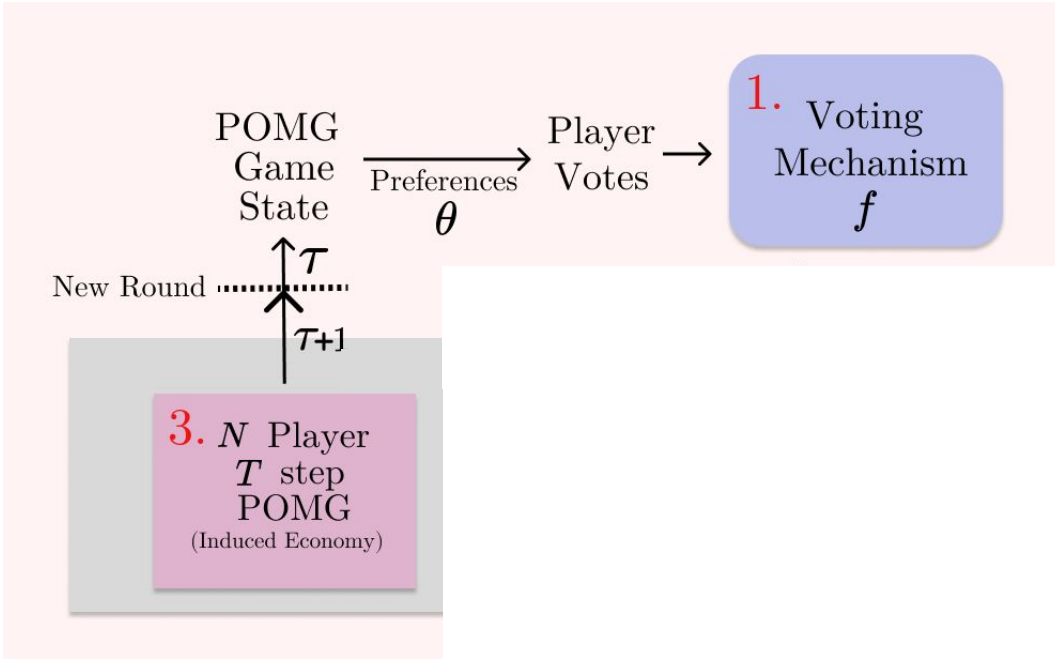
Social Environment Design

Agent roles

3. N Player
T step
POMG
(Induced Economy)

Social Environment Design

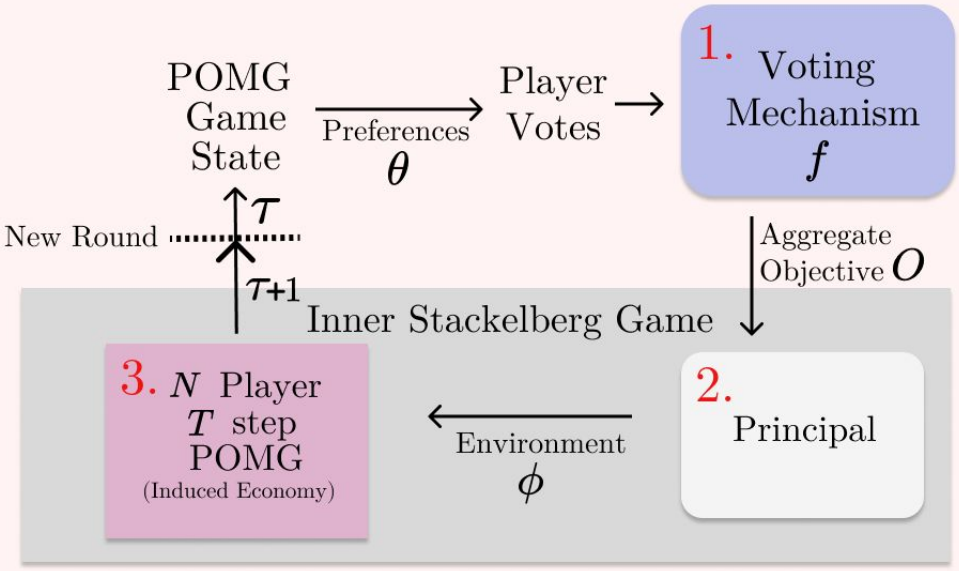
Agent roles



Voting Mech.

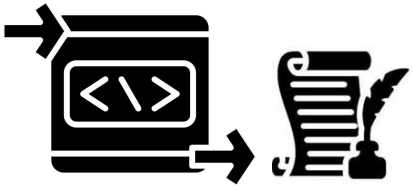
Social Environment Design

Agent roles



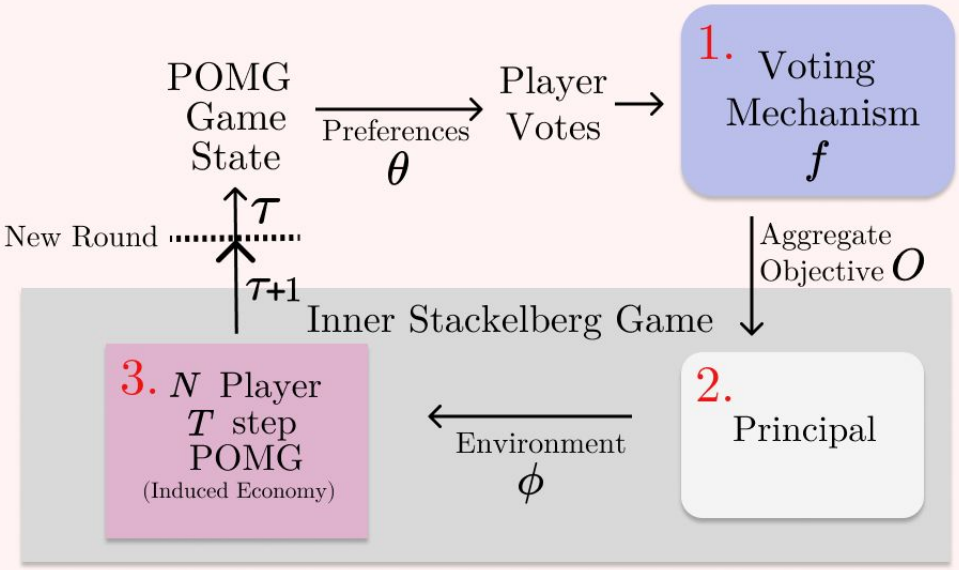
Voting Mech.

Principal



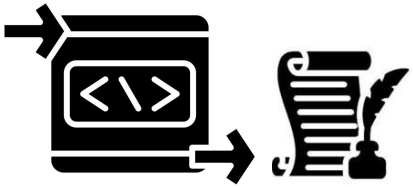
Social Environment Design

Agent roles



Voting Mech.

Principal



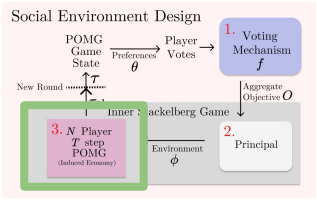
Agents vote for the **Principal's objective function**, which is selected from **p-mean** social welfare functions:

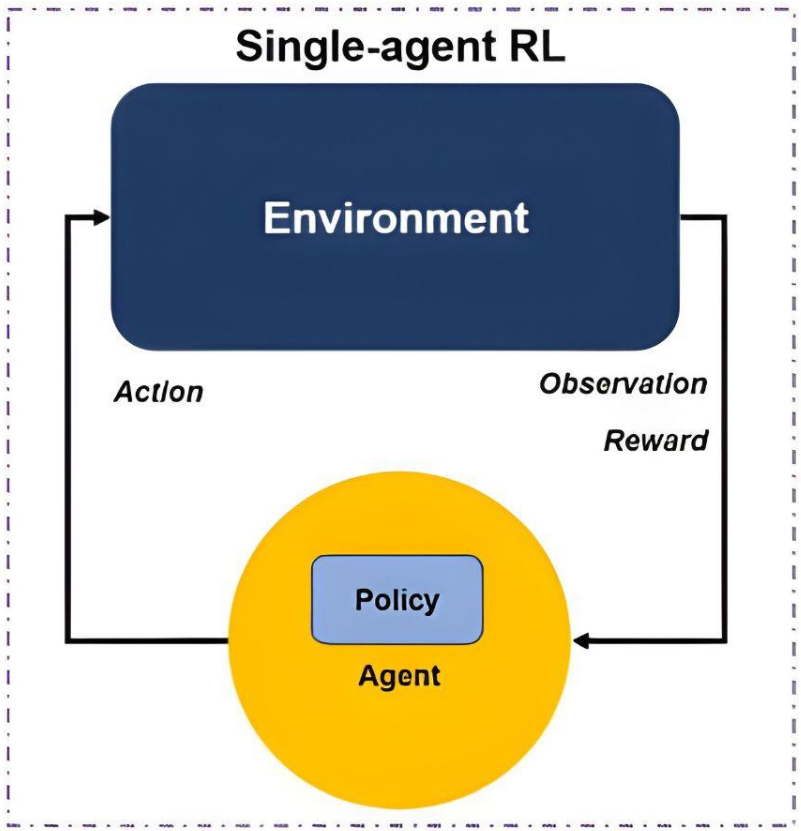
$$f^p(\mathbf{v}) = \left(\sum_{i \in [n]} v_i^p \right)^{1/p}$$

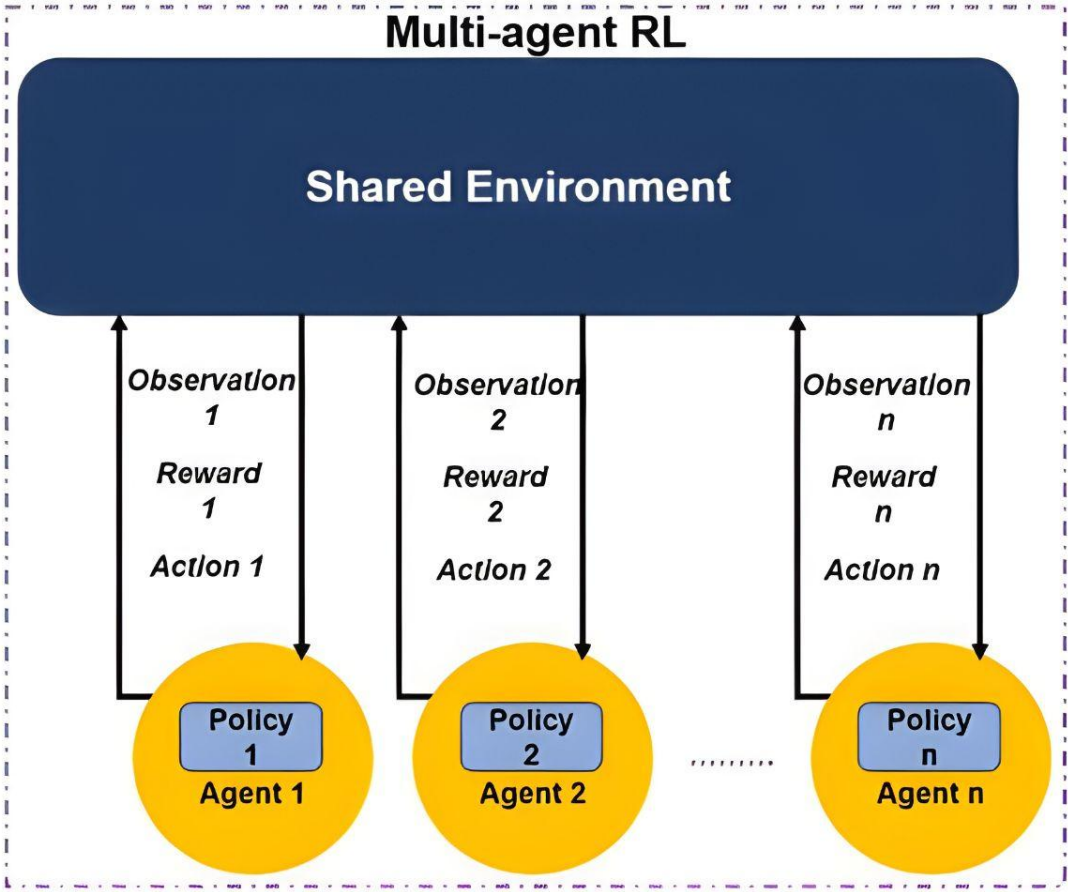
Agents choose a p which corresponds to different preferences

$$\text{Utilitarian } (\textstyle\sum, p \to 1), \text{Nash } (\textstyle\prod, p \to 0), \text{Egalitarian } (\min, p \to -\infty). \text{Elitist } (\max, p \to -\infty)$$

Preliminaries:
Multi-Agent RL (MARL) and the Markov Game



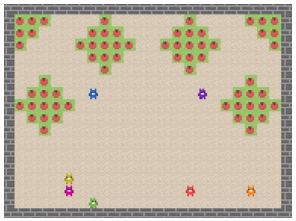




Empirical Simulation: Apple-Picking Game

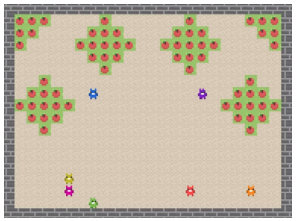
Tragedy of the commons...
individual action goes against collective good

Empirical Simulation: Apple-Picking Game



Agents = 7, each with type selfishness $\sigma_i \in [0, 1]$

Empirical Simulation: Apple-Picking Game

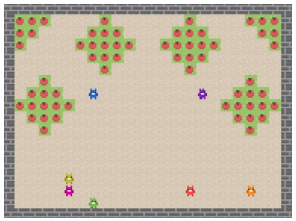


Agents = 7, each with type selfishness $\sigma_i \in [0, 1]$

Reward Parameterization: $\sigma_i \in [0, 1]$

$$r_i(a, \phi) = \sigma_i r_{\text{tax},i}(a) + (1 - \sigma_i) \left(\sum_{i' \in N_G(i)} r_{\text{tax},i'}(a, \phi) \right)$$

Empirical Simulation: Apple-Picking Game



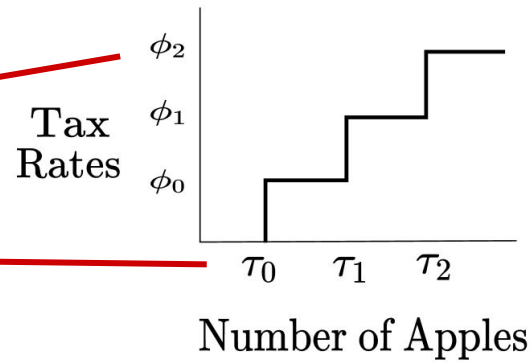
Agents = 7, each with type selfishness $\sigma_i \in [0, 1]$

Reward Parameterization: $\sigma_i \in [0, 1]$

$$r_i(a, \phi) = \sigma_i r_{\text{tax},i}(a) + (1 - \sigma_i) \left(\sum_{i' \in N_G(i)} r_{\text{tax},i'}(a, \phi) \right)$$

$$r_{\text{tax},i}(a, \phi) = (a_i - T(a_i, \phi)) + \frac{1}{n} \sum_j T(a_j, \phi),$$

$$\text{where tax } T(a, \phi) = \sum_{b=0}^{B-1} \phi_b \cdot ((\tau_{b+1} - \tau_b) \, \mathbf{1}[a > \tau_{b+1}] + (a - \tau_b) \, \mathbf{1}[\tau_b < a \leq \tau_{b+1}]).$$



Future Work

Future Work

1. Empirical simulation
2. Real-world Deployment
3. Theoretical Foundations

Future Work

1. **Empirical simulation**
2. Real-world Deployment
3. Theoretical Foundations

Empirical simulation

- Can we scale this to a much more complex and realistic simulation of the economy?
- Does incorporating LLM agents in the simulation improve realism? What are the tradeoffs?
- Can we implement a superhuman principal using current generation LLMs?

AI Economist

Generative Agents: Interactive Simulacra of Human Behavior

Future Work

1. Empirical simulation
2. **Real-world Deployment**
3. Theoretical Foundations

Real-World Deployment (sim2real)

How do we bring human feedback into the loop?

- What about incorporating real-world data just into the voting/value aggregation pipeline?
- Can we get a large group of people to play our game?
- Could we behavior clone this data into preferences RLHF-style to optimize against?

Real-World Deployment (real2sim)

- Can we build a ‘game’ around a real-world non-profit or other org?
- What should be optimized with SED, and what should be left alone (to the laissez-faire market)?
- How do we properly evaluate effectiveness and improvement of social welfare when the ground truth utilities are unknown?
- What are the baselines to compare against here?

Future Work

1. Empirical simulation
2. Real-world Deployment
3. **Theoretical Foundations**

Theoretical Foundations

- What are the solution concepts (predicting the outcome of the game)?
- What are conditions for convergence to optimality? What even is optimality with a changing objective?
- How can we get sample-efficiency guarantees for scaling the system empirically?

Future Work (recap)

- Empirical simulation
  - Can we scale to a much more complex and realistic simulation of the economy?
  - Can we implement a successful Principal using a LLM?
  - Does having LLM agents in the simulation improve the realism?
  - How we can introduce human feedback into the loop?
- Real-world Deployment
  - How do we bring human feedback into the loop?
  - How do we properly evaluate effectiveness and improvement of social welfare?
  - What are the baselines to compare against here?
- Theoretical Foundations
  - What are the solution concepts and the conditions for convergence to one?
  - How do you evaluate a non-stationary objective?
  - How can we get sample-efficiency guarantees
  - for scaling the system empirically?

Acknowledgements

Thank you to Ariel, and Itai for many fruitful discussions regarding the theoretical portions of this presentation. Thank you Matteo, Ben, and Henry for helping support many of the experiments, Sadie for contributing heavily to the theoretical results, Adam for initial feedback on this presentation, Rosie and Vincent for help on the slides, and David and Austin for many hours of feedback and support.