The Role of Incentives in Interactive Learning and AI Alignment

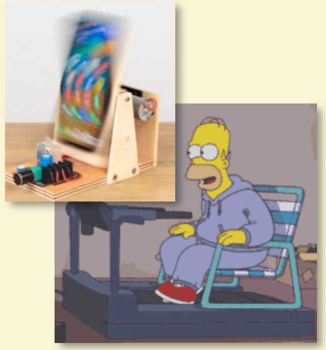
Thomas Kleine Buening

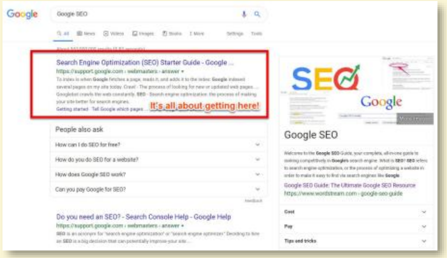
ML Seminar - Charles University
27th February 2025

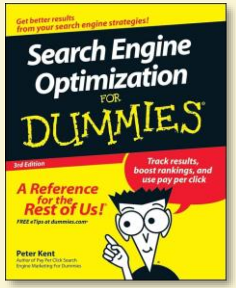
The Alan Turing Institute
UNIVERSITY OF OXFORD

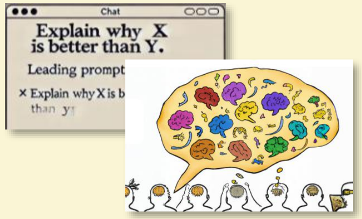
"Any system that can be gamed will be gamed.

W. Brian Arthur









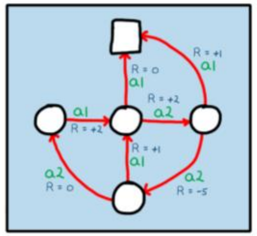
"Any system that can be gamed will be gamed.

W. Brian Arthur

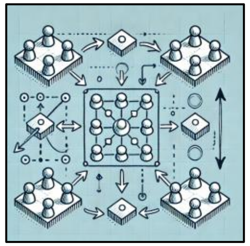
Research Agenda

Design learning algorithms that are **(1)** robust against strategic behavior and **(2)** incentivize agent behavior that is aligned with the system's goals.

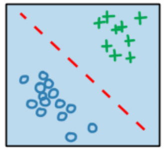
Reinforcement Learning



Mechanism Design



Supervised Learning



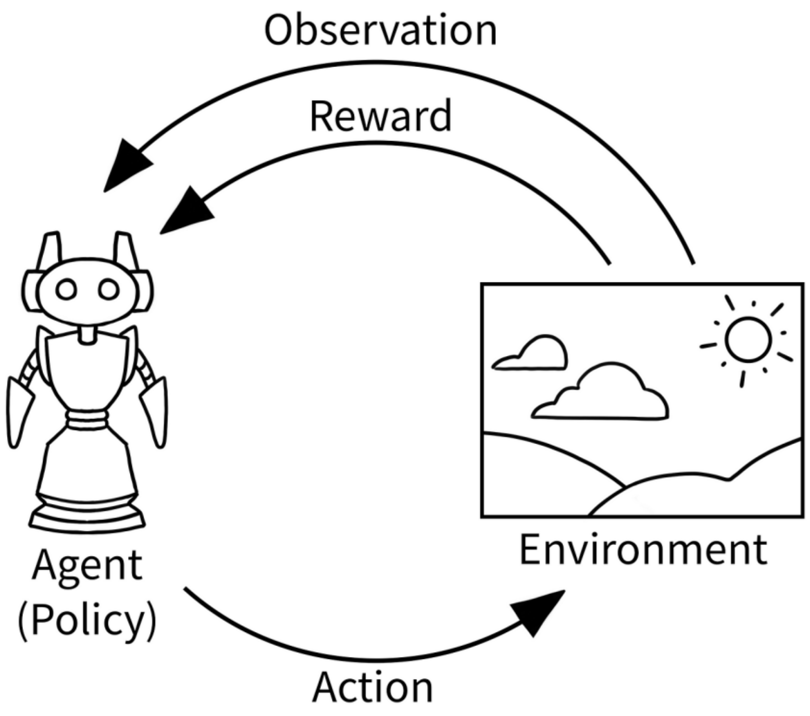
What about adversarial robustness?

- somewhat suitable to achieve **(1)**
- not at all suitable to achieve **(2)**

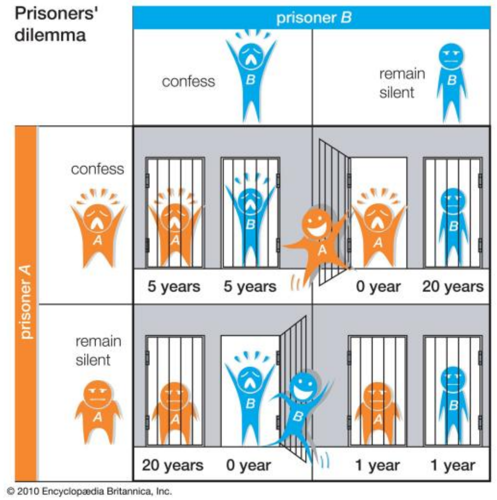
Reinforcement Learning

The Reinforcement Learning Problem

The reinforcement learning problem is the problem of *learning* how to act in an *unknown* environment, only by *interaction* and *reinforcement.*



Mechanism Design



<u>Game Theory</u> (Analyzing Games)

Given a strategic environment (i.e., a game), determine how rational agents will behave (e.g., study equilibrium).

<u>Mechanism Design</u> (Designing Games)

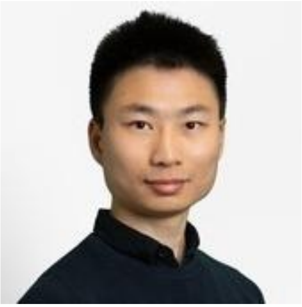
How to design a strategic environment (i.e., a game) to ensure that rational agents behave in a way that leads to a desired outcome.

(Key element: each agent holds private information).

Mechanism Design ≈ Inverse Game Theory

Strategyproof Reinforcement Learning from Human Feedback

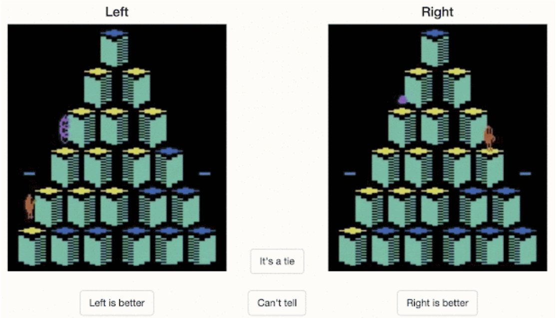
w/ Jiarui Gan, Debmalya Mandal, Marta Kwiatkowska



Reinforcement Learning from Human Feedback (RLHF)

- Popularized as a method for fine-tuning LLMs (2020 – …)
  - Used to align models with human preferences and values

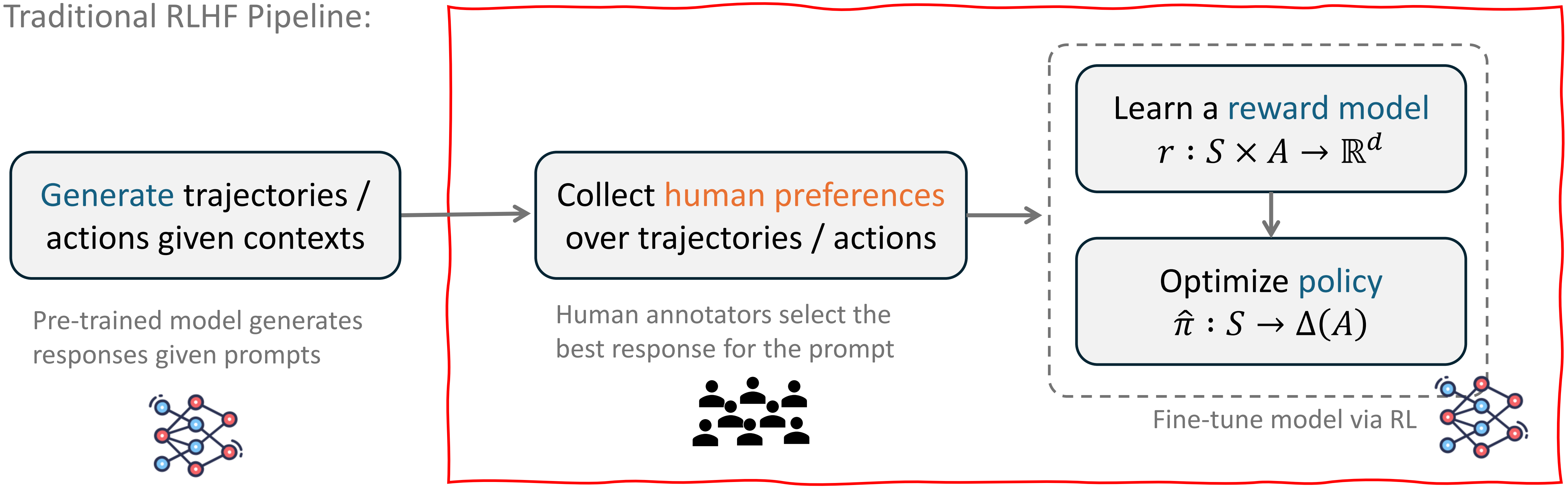
- Originates from Christiano et al. 2017 (not applied to LLMs)
  - Allows us to optimize a policy without hand-specifying a reward function instead using pairwise comparisons of trajectories



- Nowadays used as a general framework for aligning AI systems with human intentions in various applications of RL beyond LLMs such as robotics.

The Three Major Steps of RLHF

Traditional RLHF Pipeline:



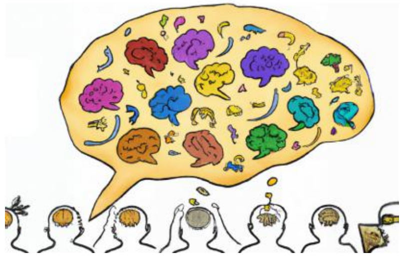
- In **Offline RLHF**, we cannot choose what trajectories to generate / compare.
- Hence, we focus on the latter two steps.

Who's preferences are we actually aligning to?

Who‘s preferences are we aligning to?

k human labelers

diverse preferences $D_1^*, \dots, D_k^*$

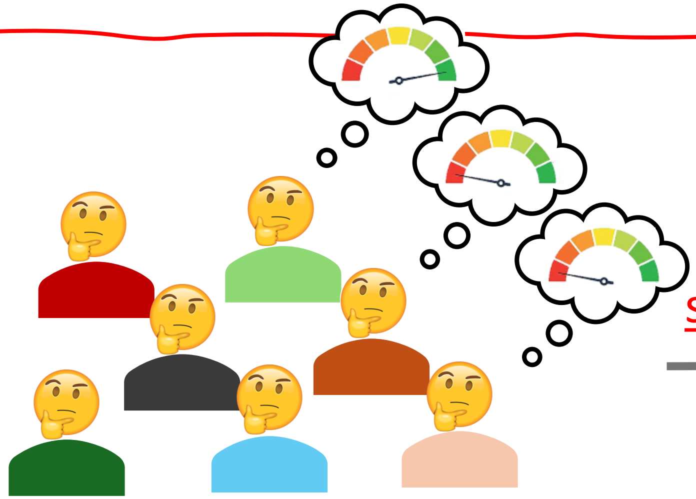


Learn reward model(s)
$r_i : S \times A \to \mathbb{R}^d$

Optimize policy
$\hat{\pi} : S \to \Delta(A)$

- **Pluralistic Alignment:**
  - There is no single set of values and preferences.
  - Compute a group-aligned policy.

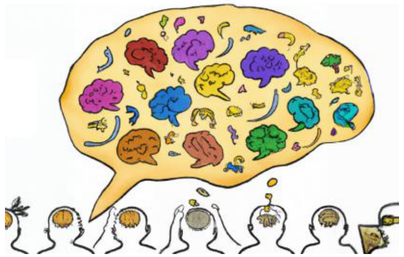
But... what kind of incentives does pluralistic alignment create?



k human labelers

"If you weigh all our opinions equally, ...“

strategic preferences $\widetilde{D}_1, \dots, \widetilde{D}_k$



Learn reward model(s)
$\tilde{r}_i : S \times A \to \mathbb{R}^d$

Optimize policy
$\hat{\pi} : S \to \Delta(A)$

- **Pluralistic Alignment:**
  - There is no single set of values and preferences.
  - Compute a group-aligned policy.
- Pluralistic alignment incentivizes malicious / strategic behavior.

How can we make RLHF robust against such strategic behavior?

In a Nutshell: Informal Problem Formulation

- Every labeler $i \in [k]$ wants the RLHF policy $\hat{\pi}$ to maximize their reward fct. $r_i^*(s,a)$:

$$J_i(\hat{\pi}) = E\left[\sum_{h=1}^{H} r_i^*(s_h, a_h) \mid a_h \sim \hat{\pi}(s_h)\right]$$

  We here assume linear reward functions $r_i^*(s,a) = \langle \theta_i^*, \phi(s,a) \rangle$.

- RLHF Objective = Maximize everyone's utility (social welfare): *(policy alignment)*

$$\mathrm{SW}(\hat{\pi}) = \sum_{h=1}^{H} J_i(\hat{\pi})$$

- Truthfulness = Report your true preferences D_i^* as represented by $r_i^*(s,a)$.
- Misreporting = Report manipulated preferences $\widetilde{D}_i$ to influence RLHF policy $\hat{\pi}$ in your favor.
- Strategyproofness = Truthfully reporting D_i^* is optimal for every labeler. *(incentive alignment)*

Trade-Offs between Incentive Alignment and Policy Alignment

Lemma (informal):

Existing RLHF methods are not strategyproof.
A single strategic labeler can make existing RLHF methods perform arbitrarily bad.

Can we reconcile incentive alignment (strategyproofness)
with policy alignment (social welfare maximization)?

In general? No.

Theorem (informal):

Every strategyproof RLHF algorithm must achieve k-times worse social welfare compared to the optimal policy, where k is the number of different labelers:

$$\mathrm{SW}(\hat{\pi}) \leq \frac{1}{k} \cdot \max_{\pi} \mathrm{SW}(\pi)$$

Incentivizing Approximate Strategyproofness

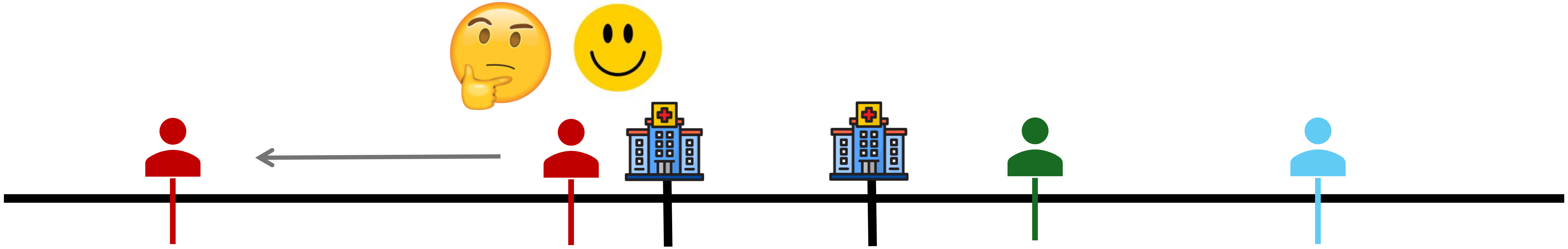
Let‘s use an idea from facility allocation: *The median is sometimes strategyproof.*

Incentivizing Approximate Strategyproofness

Let's use an idea from facility allocation: *The median is sometimes strategyproof.*

Suppose we want to decide where to build a hospital 

- Every community  wants the hospital to be as close to their homes as possible.
- Suppose we don't know the location of the communities but rely on them telling us.



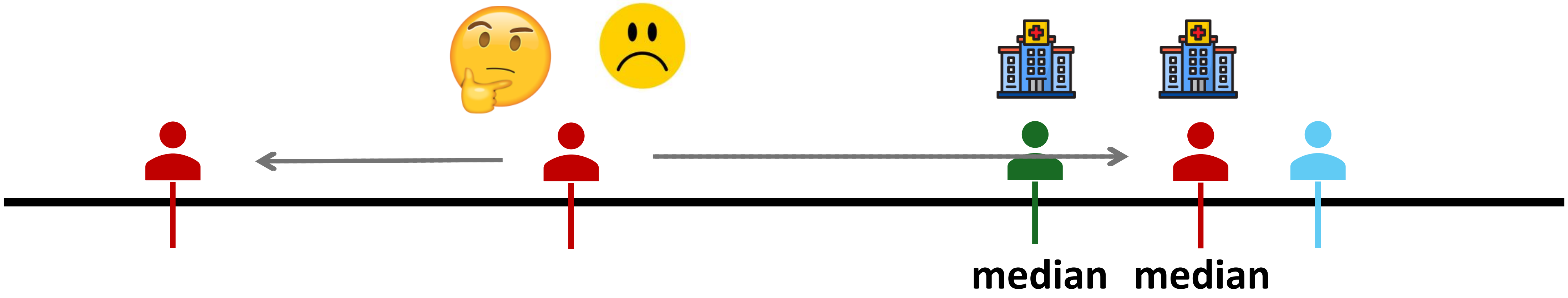
- By misreporting  can move the avg closer to their home ⇒ avg is not strategyproof

Incentivizing Approximate Strategyproofness

Let‘s use an idea from facility allocation: *The median is sometimes strategyproof.*

Suppose we want to decide where to build a hospital

- Every community wants the hospital to be as close as possible to their homes.
- Suppose we don‘t know the location of the communities but rely on them telling us.



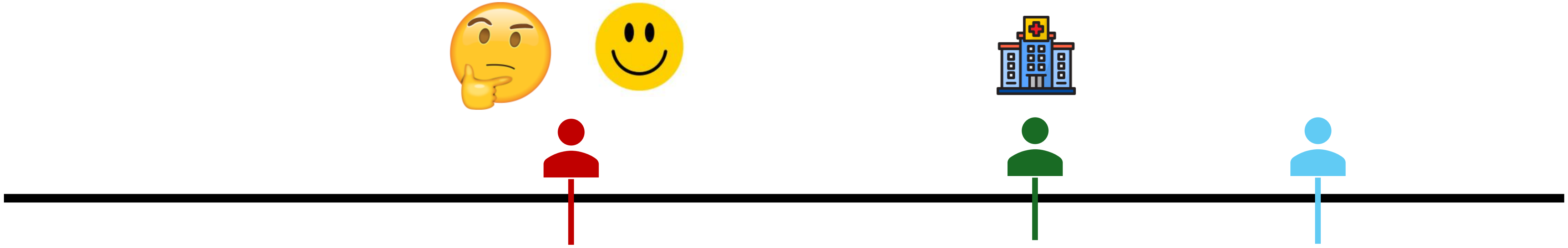
- By misreporting can move the avg closer to their home ⇒ avg is not strategyproof

Incentivizing Approximate Strategyproofness

Let‘s use an idea from facility allocation: *The median is sometimes strategyproof.*

Suppose we want to decide where to build a hospital

- Every community wants the hospital to be as close as possible to their homes.
- Suppose we don‘t know the location of the communities but rely on them telling us.



median

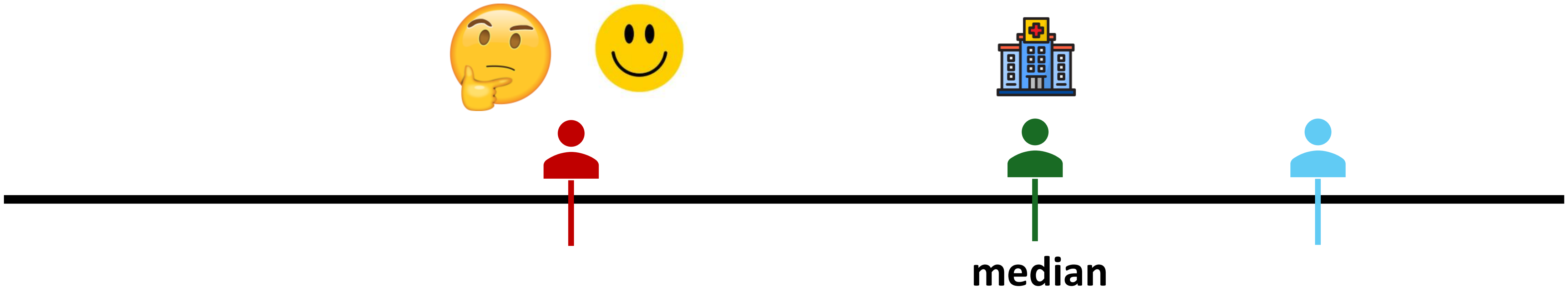
- By misreporting can move the avg closer to their home ⇒ avg is not strategyproof

Incentivizing Approximate Strategyproofness

Let‘s use an idea from facility allocation: *The median is sometimes strategyproof.*

Suppose we want to decide where to build a hospital

- Every community wants the hospital to be as close as possible to their homes.
- Suppose we don‘t know the location of the communities but rely on them telling us.



- By misreporting can move the avg closer to their home ⇒ avg is not strategyproof
- The median cannot be moved closer by misreporting ⇒ median is strategyproof

Pessimistic Median of MLEs

Algorithm 1 Pessimistic Median of MLEs (Pessimistic MoMLE)

input offline preference data sets $\mathcal{D}_1, \dots, \mathcal{D}_k$

1: **for** every labeler $i \in [k]$ **do**

2: compute the MLE $\hat{\theta}_i^{\mathrm{MLE}}$

3: construct confidence set $C_i := \{\theta \in \mathbb{R}^d : \|\hat{\theta}_i^{\mathrm{MLE}} - \theta\|_{\Sigma_{\mathcal{D}_i}} \le f(d, n, \delta)\}$

4: **end for**

5: get median confidence set $\mathscr{C} := \{\text{c-Median}(\theta_1, \dots, \theta_k) : \theta_i \in C_i \text{ for } i \in [k]\}$

6: get pessimistic estimate of the social welfare w.r.t. $\mathscr{C}$ given by

$$\underline{\mathcal{W}}(\pi) := \min_{\theta \in \mathscr{C}} \mathbb{E}_{s \sim \rho} [\langle \theta, \phi(s, \pi(s)) \rangle]$$

7: **return** $\hat{\pi} = \operatorname{argmax}_{\pi \in \Pi} \underline{\mathcal{W}}(\pi)$

Theorem (informal):

(1) Pessimistic MoMLE is $\sqrt{d/n}$-strategyproof (i.e., approximately strategyproof).

(2) Pessimistic MoMLE is suboptimal by a margin of at most $\sqrt{d/k} + k\sqrt{d/n}$.

k = #labelers, d = feature dimension, n = #samples

Main Take-Aways

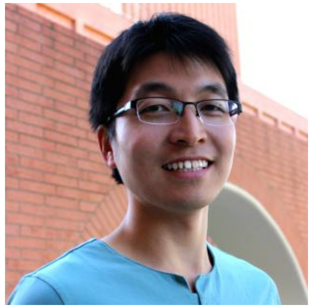
- Pluralistic alignment invites *malicious and strategic human feedback.*
- Fundamental trade-off between *incentive alignment* (discouraging strategic feedback) and *policy alignment* (maximizing social welfare).
- Social Choice Theory tells us that the median is strategyproof under certain conditions. Combining the *median* with *pessimistic* estimates, we can balance this trade-off:
  - *approximate* strategyproofness
  - RLHF *converges* to the optimal policy as #labelers and #samples increases

Strategic Linear Contextual Bandits

w/ Aadirupa Saha, Christos Dimitrakakis, Haifeng Xu

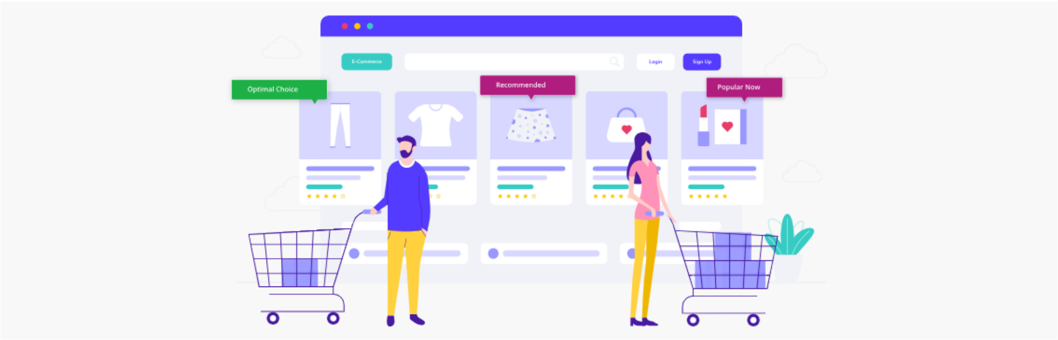
NeurIPS 2024

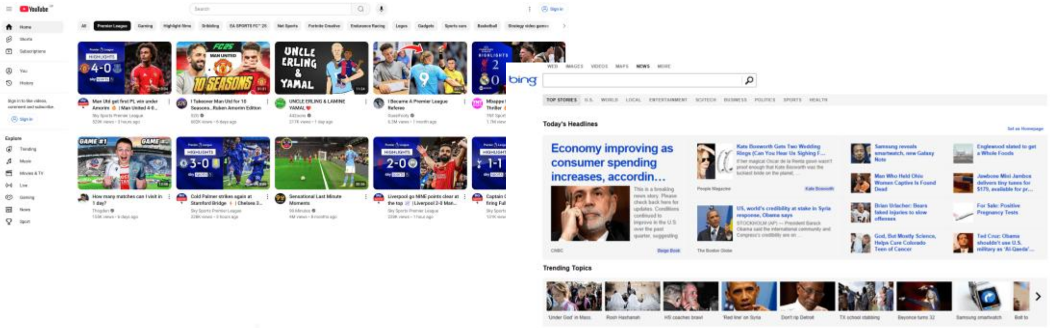


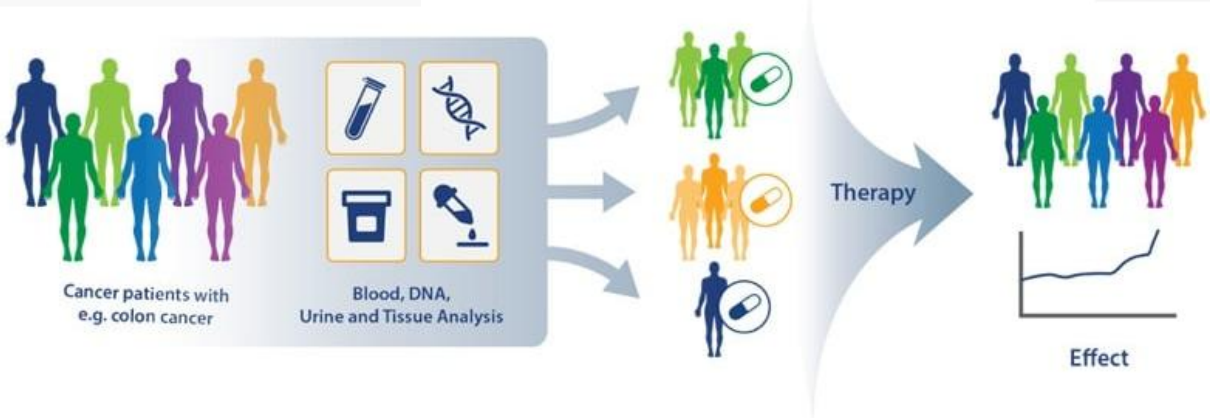


Contextual Bandits

Select the *best action* given relevant *contextual information*.



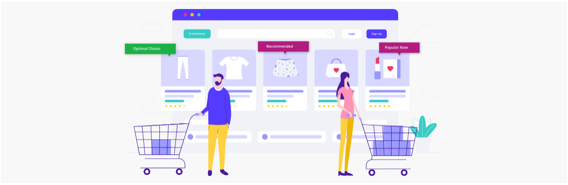


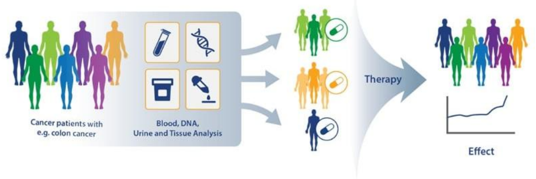


repeatedly:

1) algorithm observes relevant *contextual information*
2) algorithm takes an *action* and receives a *reward* for the taken action.

Linear Contextual Bandits





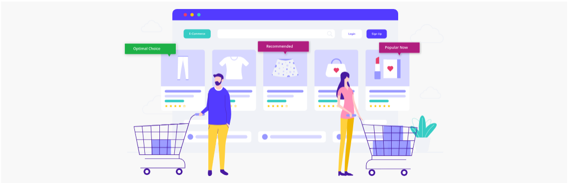
repeatedly:

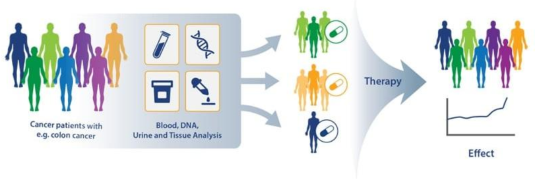
Where do these contexts actually come from?

1) algorithm observes every arm's contexts $x^*_{t,1}, \ldots, x^*_{t,K} \in \mathbb{R}^d$
2) algorithm selects arm $a_t \in [K]$ and receives reward $r_{t,a_t} \sim D\big(\langle \theta^*, x^*_{t,a_t} \rangle\big)$

agent = vendor / content creator / patient / healthcare provider

Strategic Linear Contextual Bandits





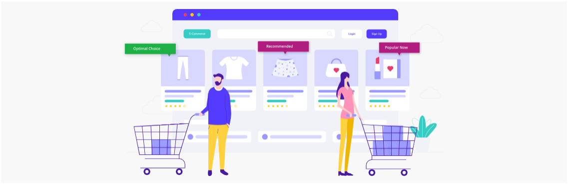
repeatedly:

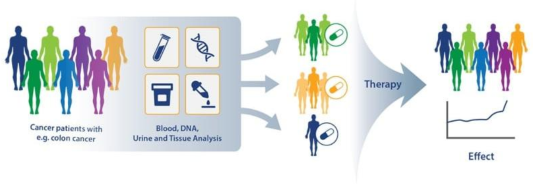
Where do these contexts actually come from?

1) algorithm observes every agent's contexts $x^*_{t,1}, \dots, x^*_{t,n} \in \mathbb{R}^d$
2) algorithm selects agent $a_t \in [K]$ and receives reward $r_{t,a_t} \sim D\left(\langle \theta^*, x^*_{t,a_t} \rangle\right)$

agent = vendor / content creator / patient / healthcare provider

Strategic Linear Contextual Bandits





repeatedly:

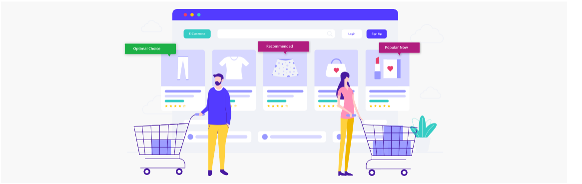
Where do these contexts actually come from?

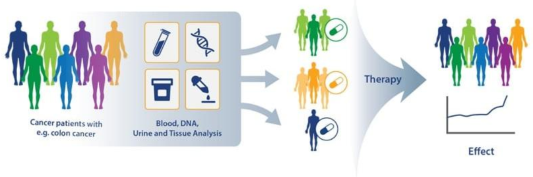
~~1) algorithm observes every agent's contexts $x^*_{t,1}, \dots, x^*_{t,n} \in \mathbb{R}^d$~~

2) algorithm selects agent $a_t \in [K]$ and receives reward $r_{t,a_t} \sim D\left(\langle \theta^*, x^*_{t,a_t} \rangle\right)$

agent = vendor / content creator / patient / healthcare provider

Strategic Linear Contextual Bandits





repeatedly:

Where do these contexts actually come from?

1) every agent $a \in [K]$ privately observes their context $x^*_{t,a}$ and reports gamed context $\tilde{x}_{t,a}$
2) algorithm selects agent $a_t \in [K]$ and receives reward $r_{t,a_t} \sim D(\langle \theta^*, x^*_{t,a_t} \rangle)$
3) agent a_t receives some utility (e.g., 1) for being selected.

Algorithm minimizes expected regret

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in [K]} \langle \theta^*, x^*_{t,a} \rangle - \langle \theta^*, x^*_{t,a_t} \rangle\right]$$

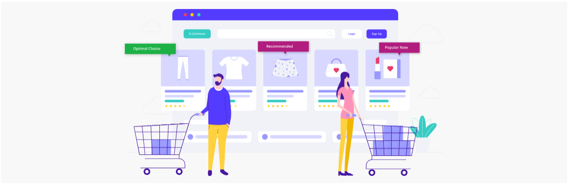
Every agent a maximizes its #selections

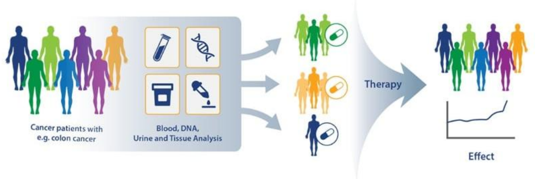
$$\mathbb{E}\left[\sum_{t=1}^{T} 1(a_t = a)\right]$$

Goal: Bound the regret assuming the agents respond to the algorithm in Nash Equilibrium.

agent = vendor / content creator / patient / healthcare provider

Strategic Linear Contextual Bandits

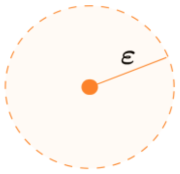




repeatedly:

Where do these contexts actually come from?

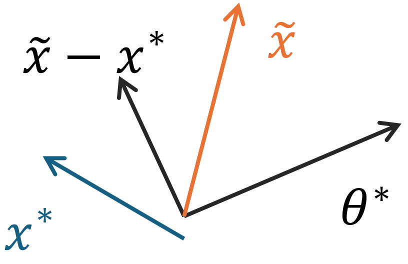
1) every agent $a \in [K]$ privately observes their context $x^*_{t,a}$ and reports gamed context $\tilde{x}_{t,a}$
2) algorithm selects agent $a_t \in [K]$ and receives reward $r_{t,a_t} \sim D(\langle \theta^*, x^*_{t,a_t} \rangle)$
3) agent a_t receives some utility (e.g., 1) for being selected.



Remarks:

- unbounded manipulation ($\tilde{x}_{t,a}$ can arbitrarily differ from $x^*_{t,a}$) $\Rightarrow$ adversarial perspective fails!
- we only observe $\tilde{x}_{t,a}$ and $r_{t,a_t} \sim D(\langle \theta^*, x^*_{t,a_t} \rangle) \Longrightarrow$ cannot infer θ?!

reward parameter θ^* and true context x^*_{t,a_t} are both unknown



$\langle \theta^*, \tilde{x} \rangle = \langle \theta^*, x^* \rangle$

Optimistic Grim Trigger Mechanism

$$\hat{\theta}_{t,a} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\sum_{\ell < t:\, a_\ell = a} \left(\langle \theta, \tilde{x}_{\ell,a} \rangle - r_{\ell,a} \right)^2 + \lambda \|\theta\|_2^2 \right)$$

$$\tilde{C}_{t,a} = \left\{ \theta \in \mathbb{R}^d :\, \|\hat{\theta}_{t,a} - \theta\|_{V_{t,a}}^2 \leq \beta_{t,a} \right\}$$

Inspired by iterated social dilemmas:

1. Independent estimates $\hat{\theta}_{t,a}$ and confidences $\tilde{C}_{t,a}$ for every a based on gamed contexts.
2. Play optimistically w.r.t. reported contexts $\tilde{x}_{t,a}$: $$a_t = \operatorname*{argmax}_{a \in A_t} \left(\langle \hat{\theta}_{t,a}, \tilde{x}_{t,a} \rangle + \sqrt{\beta_{t,a}} \|\tilde{x}_{t,a}\|_{V_{t,a}^{-1}} \right)$$
3. Eliminate a if $$\sum_{\ell \leq t:\, a_\ell = a} \left(\langle \hat{\theta}_{t,a}, \tilde{x}_{\ell,a} \rangle - \sqrt{\beta_{\ell,a}} \|\tilde{x}_{\ell,a}\|_{V_{\ell,a}^{-1}} \right) > \sum_{\ell \leq t:\, a_\ell = a} r_{t,a} + 2\sqrt{n_t(a) \log(1/\delta)}$$

LCB of reported reward > UCB of observed reward

- Estimates $\hat{\theta}_{t,a}$ can be incorrect and $\theta^* \notin \tilde{C}_{t,a}$. Why doesn‘t this matter?
- Intuition: We care about making good decisions, not learning θ^*. Before elimination:

$$\sum_{t \leq \tau_a :\, a_t = a} \left(\langle \hat{\theta}_{t,a}, \tilde{x}_{t,a} \rangle - \langle \theta^*, x_{t,a}^* \rangle \right) \leq \sum_{t \leq \tau_a :\, a_t = a} \sqrt{\beta_{t,a}} \|\tilde{x}_{t,a}\|_{V_{t,a}^{-1}} + 4\sqrt{n_{\tau_a}(a) \log(T)}.$$

Main Results

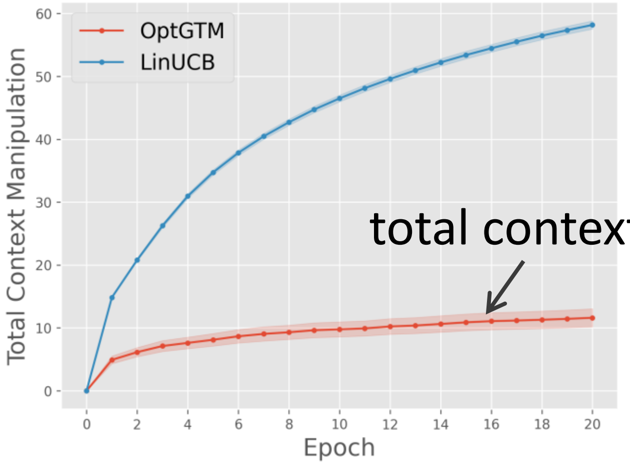
Theorem (informal): OptGTM satisfies:

approximately incentive-compatible

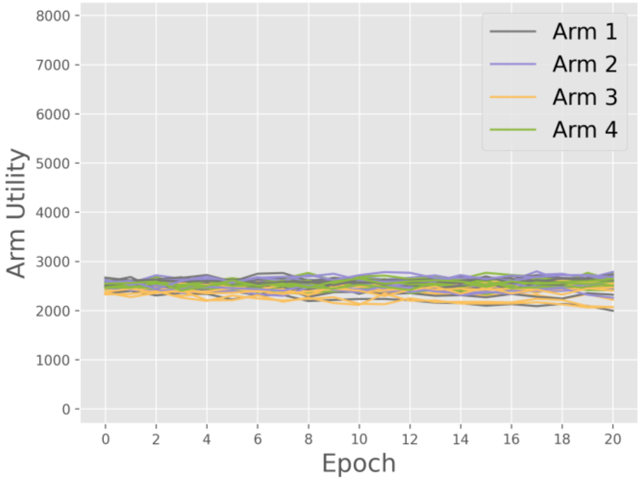
1) Always truthfully reporting $\tilde{x}_{t,a} = x_{t,a}$ is an $\tilde{O}(d\sqrt{KT})$-equilibrium.
2) Under every equilibrium, we suffer at most $\boxed{d\sqrt{KT}} + \boxed{dK^2\sqrt{KT}}$ regret

price of manipulation

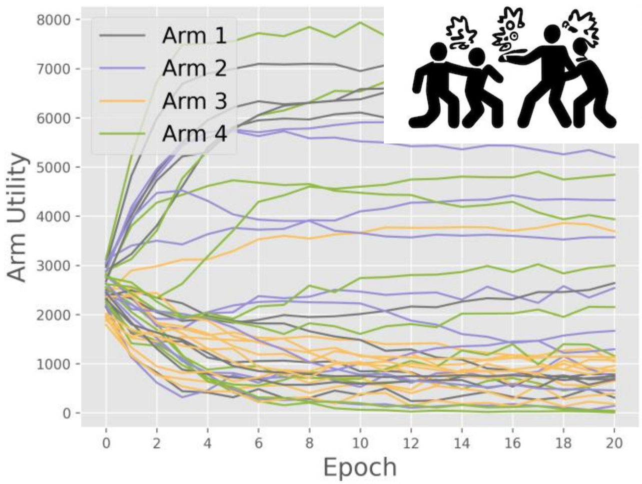
price of mechanism design



total context manipulation



OptGTM



LinUCB

Main Take-Aways

- Taking the strategic perspective we can achieve what is impossible when taking a stochastic or adversarial perspective.
- Mechanism design becomes approximate due to uncertainty about the environment.
- Trade-offs between incentive alignment and reward minimization.
- Many open questions and problems left to explore in this line of research.

Thank you!