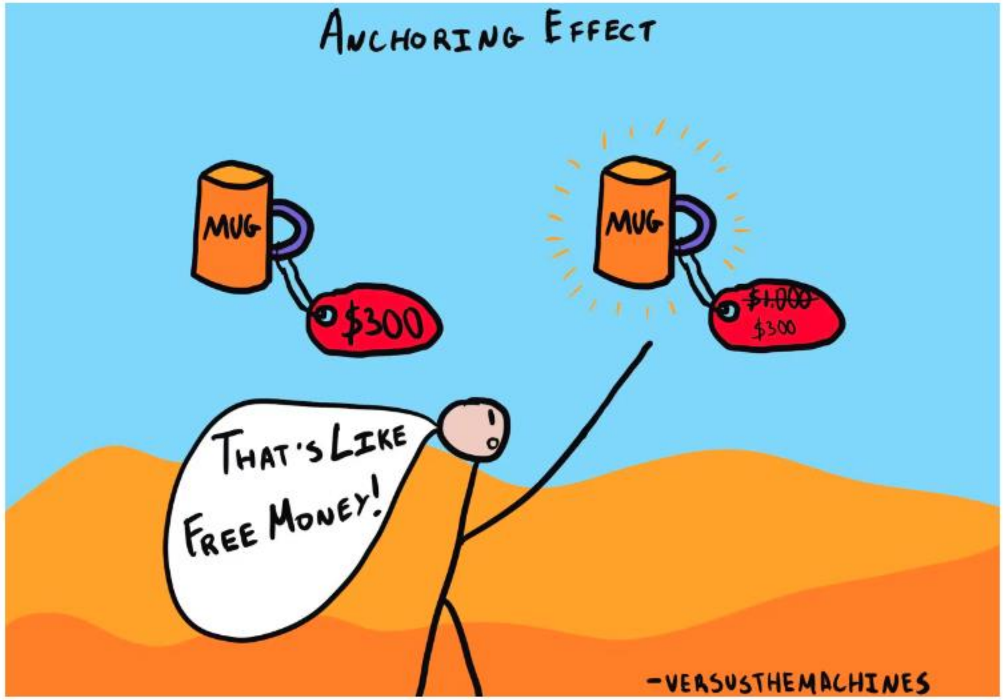
Bias Detection via Signaling

Yiling Chen[1], **Tao Lin**[1], Ariel D. Procaccia[1], Aaditya Ramdas[2], Itai Shapira[1]

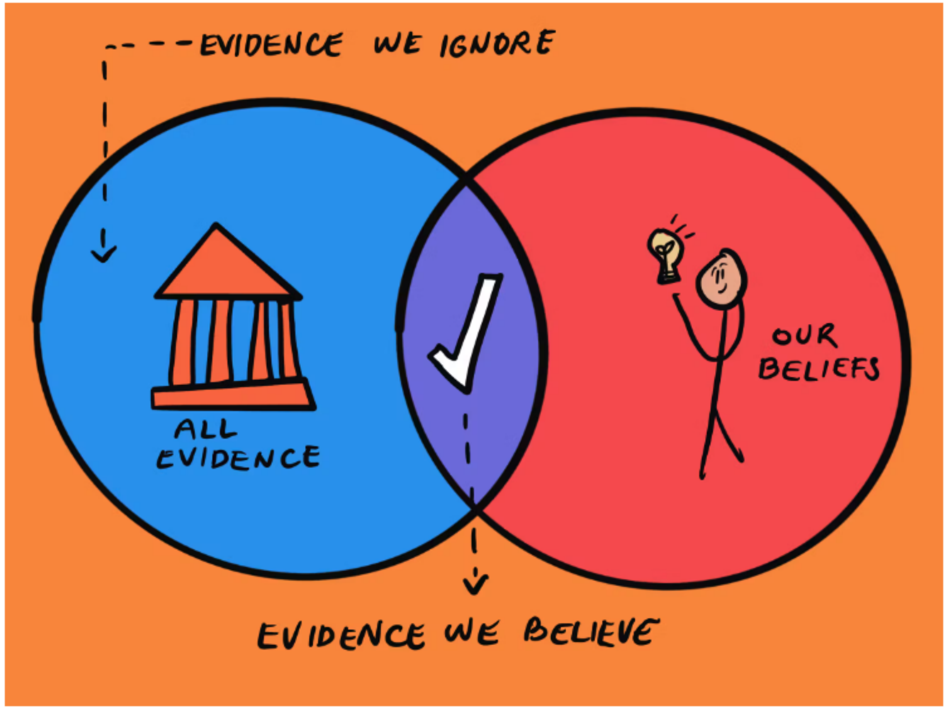
[1]:Harvard University [2]:Carnegie Mellon University

NeurIPS 2024

People (or AI) are biased



Anchoring Bias



Confirmation Bias

Q1: How to *quantify* bias?

We adopt a linear bias model in the context of Bayesian reasoning from economics:

- An agent has to make a decision (choosing an action $a \in A$)
- There is a state of the world $\theta \in \Theta$, distributed according to prior $\mu_0 \in \Delta(\Theta)$
- If the agent was perfectly Bayesian, then whenever receiving a piece of evidence (signal) $s \sim P(s|\theta)$, the posterior $\mu_s \in \Delta(\Theta)$ should be $\mu_s(\theta) = \frac{\mu_0(\theta)P(s|\theta)}{P(s)}$
- The agent has a *bias towards the prior*, measured by $w \in [0, 1]$ (bias level):
$$\nu_s = w\,\mu_0 + (1 - w)\,\mu_s$$
- Based on their biased belief ν_s, the agent makes an optimal decision:
$$a^* \in \mathrm{argmax}_{a \in A}\, \mathbb{E}_{\theta \sim \nu_s}[\, U(a, \theta)\,]$$

Q2: How to *detect* and *measure* bias?

- An agent has to make a decision (choosing an action $a \in A$)
- There is a state of the world $\theta \in \Theta$, distributed according to prior $\mu_0 \in \Delta(\Theta)$
- If the agent was perfectly Bayesian, then whenever receiving a piece of evidence (signal) $s \sim P(s|\theta)$, the posterior $\mu_s \in \Delta(\Theta)$ should be $\mu_s(\theta) = \frac{\mu_0(\theta)P(s|\theta)}{P(s)}$
- The agent has a *bias towards the prior*, measured by $w \in [0, 1]$ (bias level):

$$\nu_s = w\,\mu_0 + (1 - w)\,\mu_s$$

- Based on their biased belief ν_s, the agent makes an optimal decision:

$$a^* \in \mathrm{argmax}_{a \in A}\, \mathbb{E}_{\theta \sim \nu_s}[\,U(a, \theta)\,]$$

We use **information design**:
design the "$P(\cdot \mid \cdot)$" (signaling scheme)
to infer whether $w \geq \tau$ or $w \leq \tau$ (from the agent's actions)

We can design signaling scheme $\pi_t : \Omega \to \Delta(S)$ *adaptively*

Our Results

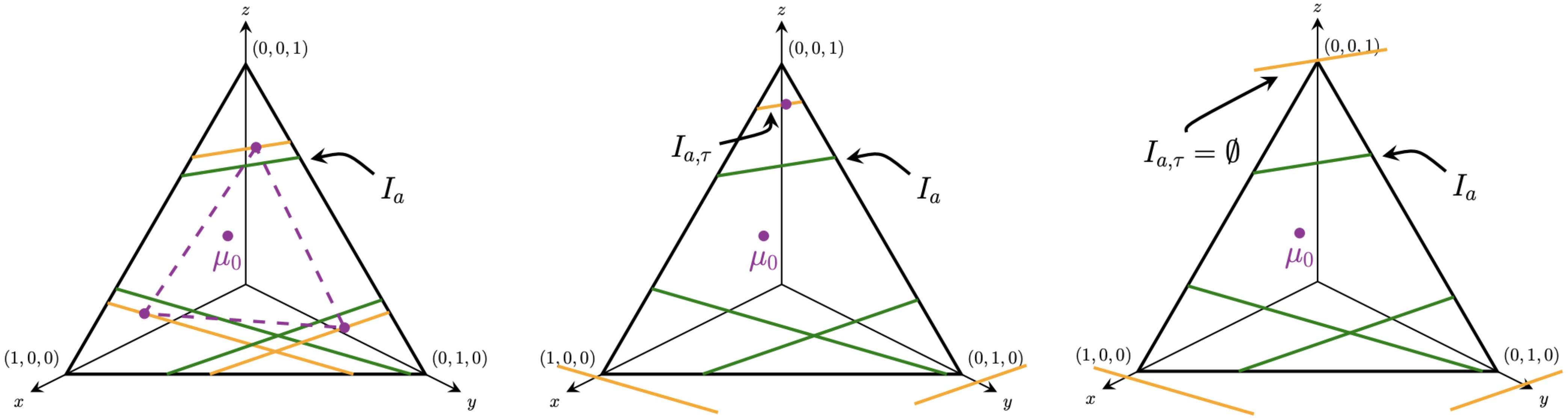
The **Sample Complexity** of an adaptive algorithm for the bias detection problem is:

E[time steps needed for the algorithm to output $w \geq \tau$ or $w \leq \tau$]

Main Theorem:

- *Constant* algorithms are as powerful as adaptive algorithms.
- The optimal constant signaling scheme π^* can be computed by a linear program (with poly size)
- π^* has the following properties:
  - π^* recommends action $a \in A$ to agent ("revelation principle")
  - Let a_0 be the optimal action for the agent at the prior belief μ_0. When π^* recommends action $a \neq a_0$, the agent's actual action reveals $w \geq \tau$ or $w \leq \tau$ (if actual action = a_0, then $w \geq \tau$; if actual action = a or any $a' \neq a_0$, then $w \leq \tau$)

Geometric Characterization



(a) A single sample (b) Finite sample complexity (c) Cannot be solved

Scan here for details:

Bias Detection via Signaling. (NeurIPS 2024)

Yiling Chen, Tao Lin, Ariel D. Procaccia, Aaditya Ramdas, Itai Shapira