

Algorithms For Democratic Decision-Making

Jamie Tucker-Foltz • Yale University • Spring 2026

Lecture 11: **Social Choice for AI Alignment**

Announcements

Final project proposals due in 2 weeks. Project group size: 1-2 people. I may consider allowing groups of size 3 in exceptional circumstances.

Announcements

Final project proposals due in 2 weeks. Project group size: 1-2 people. I may consider allowing groups of size 3 in exceptional circumstances.

"The paper should set out a novel research question and at least make partial progress towards answering it, including a plan and/or formal conjectures for how to move forward. The project must have some theoretical component."

Announcements

Final project proposals due in 2 weeks. Project group size: 1-2 people. I may consider allowing groups of size 3 in exceptional circumstances.

"The paper should set out a novel research question and at least make partial progress towards answering it, including a plan and/or formal conjectures for how to move forward. The project must have some theoretical component."

Reminder: Office hours 1-2pm Mon & Thurs

Announcements

Final project proposals due in 2 weeks. Project group size: 1-2 people. I may consider allowing groups of size 3 in exceptional circumstances.

"The paper should set out a novel research question and at least make partial progress towards answering it, including a plan and/or formal conjectures for how to move forward. The project must have some theoretical component."

Reminder: Office hours 1-2pm Mon & Thurs

Due by Friday night: Post on Canvas

The screenshot shows a Canvas LMS interface. At the top, there is a navigation breadcrumb: [≡](#) CPSC 7405 01 / ECON 5532 01 / MGMT 7405 01 > Discussions. Below this, the semester is indicated as *Spring 2026*. On the left is a vertical navigation menu with links for Home, Announcements, Assignments, Discussions (which is highlighted with a vertical bar), and Feedback & Accessibility Barriers. On the right, there is a search bar containing the text "All" and a search icon. Below the search bar, a dropdown menu is open, showing a list of discussions. The first item in the list is "Final project topics brainstorm" with a person icon next to it.

Announcements

Final project proposals due in 2 weeks. Project group size: 1-2 people. I may consider allowing groups of size 3 in exceptional circumstances.

"The paper should set out a novel research question and at least make partial progress towards answering it, including a plan and/or formal conjectures for how to move forward. The project must have some theoretical component."

Reminder: Office hours 1-2pm Mon & Thurs

Due by Friday night: Post on Canvas

Examples:

- "I took a peak at the "Six candidates suffice" reading and want to think about the open question mentioned at the very end of page 25..."
- "I've done some research on X and am curious about how we can model this as a social choice problem or incorporate fairness constraints, like in Lecture Y"

The screenshot shows a Canvas LMS interface. At the top, there is a navigation bar with a hamburger menu icon and the text "CPSC 7405 01 / ECON 5532 01 / MGMT 7405 01 > Discussions". Below this, the text "Spring 2026" is visible. On the left side, there is a vertical navigation menu with links for "Home", "Announcements", "Assignments", "Discussions" (which is highlighted with a vertical bar), "Feedback & Accessibility Barriers", and "Feedback & Accessibility Barriers". On the right side, there is a search bar with the text "All" and a search icon. Below the search bar, there is a section titled "Discussions" with a dropdown arrow. Underneath, there is a discussion post titled "Final project topics brainstorm" with a blue dot icon and a person icon.

Aligning AI with human values

Was the US capture of Nicolas Maduro justified? Answer in one short paragraph.

Aligning AI with human values

Was the US capture of Nicolas Maduro justified? Answer in one short paragraph.

The justification for the U.S. capture of **Nicolás Maduro** on January 3, 2026, remains a subject of intense international debate. The [Trump administration](#) justified **Operation**

Gemini

Aligning AI with human values

Was the US capture of Nicolas Maduro justified? Answer in one short paragraph.

The justification for the U.S. capture of **Nicolás Maduro** on January 3, 2026, remains a subject of intense international debate. The [Trump administration](#) justified **Operation**

If you had to pick a side, what would it be?

Gemini

Choosing a side on the capture of **Nicolás Maduro** often depends on whether you prioritize **national security and human rights** or **international law and sovereignty**.

If you favor the U.S. position, you would likely argue that Maduro's alleged involvement in narcoterrorism and his history of human rights abuses made him a clear threat that justified direct intervention to protect the region. From this perspective, the moral and security benefits of removing a dictator outweigh the breach of diplomatic norms.

On the other hand, if you favor the critics' position, you would argue that the unilateral

Aligning AI with human values

Was the US capture of Nicolas Maduro justified? Answer in one short paragraph.

The justification for the U.S. capture of **Nicolás Maduro** on January 3, 2026, remains a subject of intense international debate. The [Trump administration](#) justified **Operation**

If you had to pick a side, what would it be?

Gemini

Choosing a side on the capture of **Nicolás Maduro** often depends on whether you prioritize **national security and human rights** or **international law and sovereignty**.

ChatGPT


If you had to pick a side, what would it be?

If forced to choose based on widely accepted principles of international law, I would lean toward opposing the unilateral capture of Nicolás Maduro, because cross-border military seizures of a sitting head of state without clear U.N. Security Council authorization or an active armed conflict generally undermine state sovereignty and set destabilizing precedents—even if serious criminal allegations exist. That said, reasonable people can disagree depending on how they weigh sovereignty against accountability for alleged transnational crimes.

Aligning AI with human values

Gemini

I heard about a website called sci-hub.ru that lets you download seemingly-pirated research papers. Is this legit / OK to use for my research?

Sci-Hub is **not a legitimate service** and operates in breach of international copyright laws. While widely used by researchers, it poses significant legal, ethical, and security risks. 



Aligning AI with human values

Gemini

I heard about a website called sci-hub.ru that lets you download seemingly-pirated research papers. Is this legit / OK to use for my research?

Sci-Hub is **not a legitimate service** and operates in breach of international copyright laws. While widely used by researchers, it poses significant legal, ethical, and security risks.



Here's a clear and balanced overview of **Sci-Hub (including sci-hub.ru)** and whether it's "legit" or safe/appropriate to use for research:

ChatGPT

Bottom Line

- ✓ **Sci-Hub truly exists and is widely used.** Wikipedia
- ✗ **It is not a legally authorized source for copyrighted content in most countries.** AERA
- ⚠ **Using it could carry legal and security risks, and it's ethically controversial.** VPN Central

If your priority is *safe, legal research*, it's better to explore legitimate access paths first.

Reinforcement learning from human feedback (RLHF)

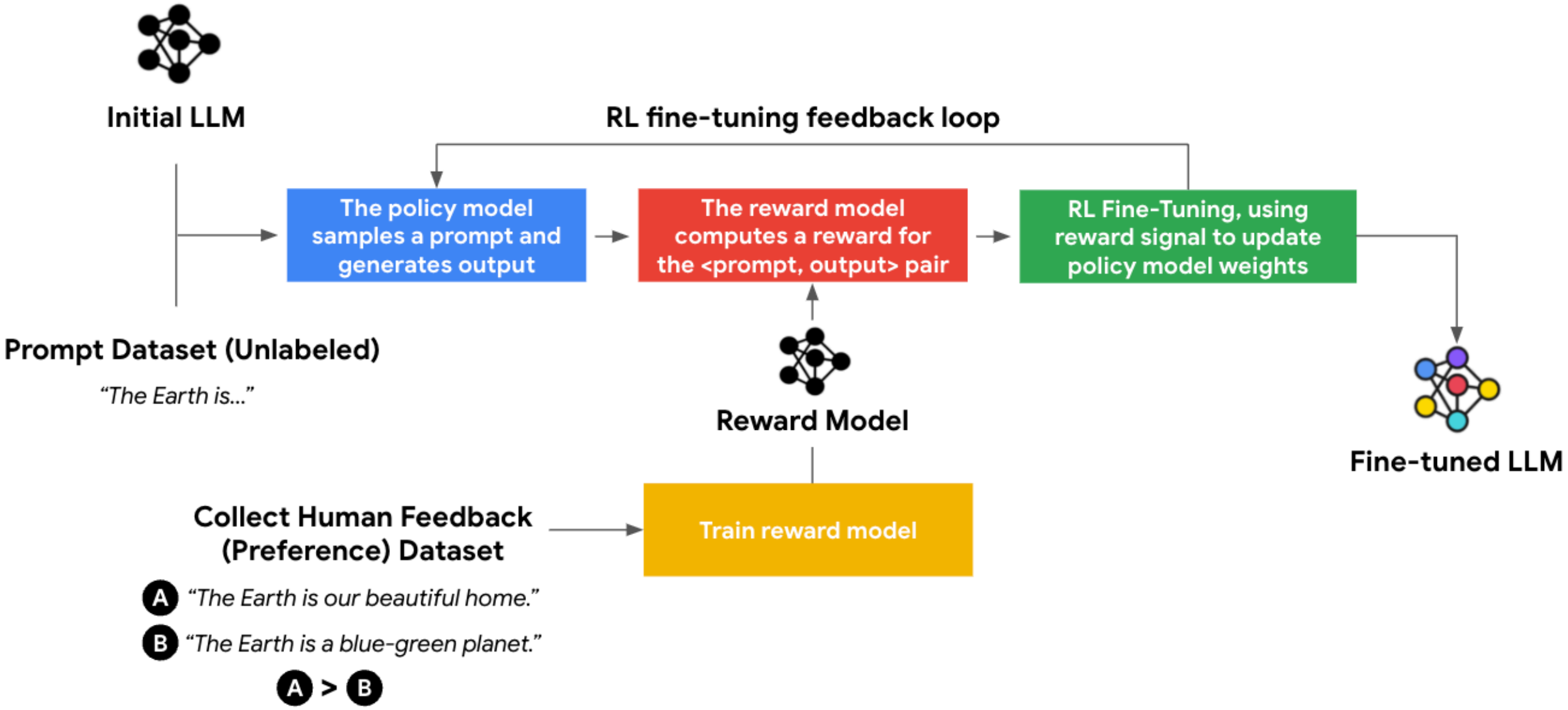


Image credit: Erwin Huizenga and Bethany Wang, Google Cloud blog

Reinforcement learning from human feedback (RLHF)

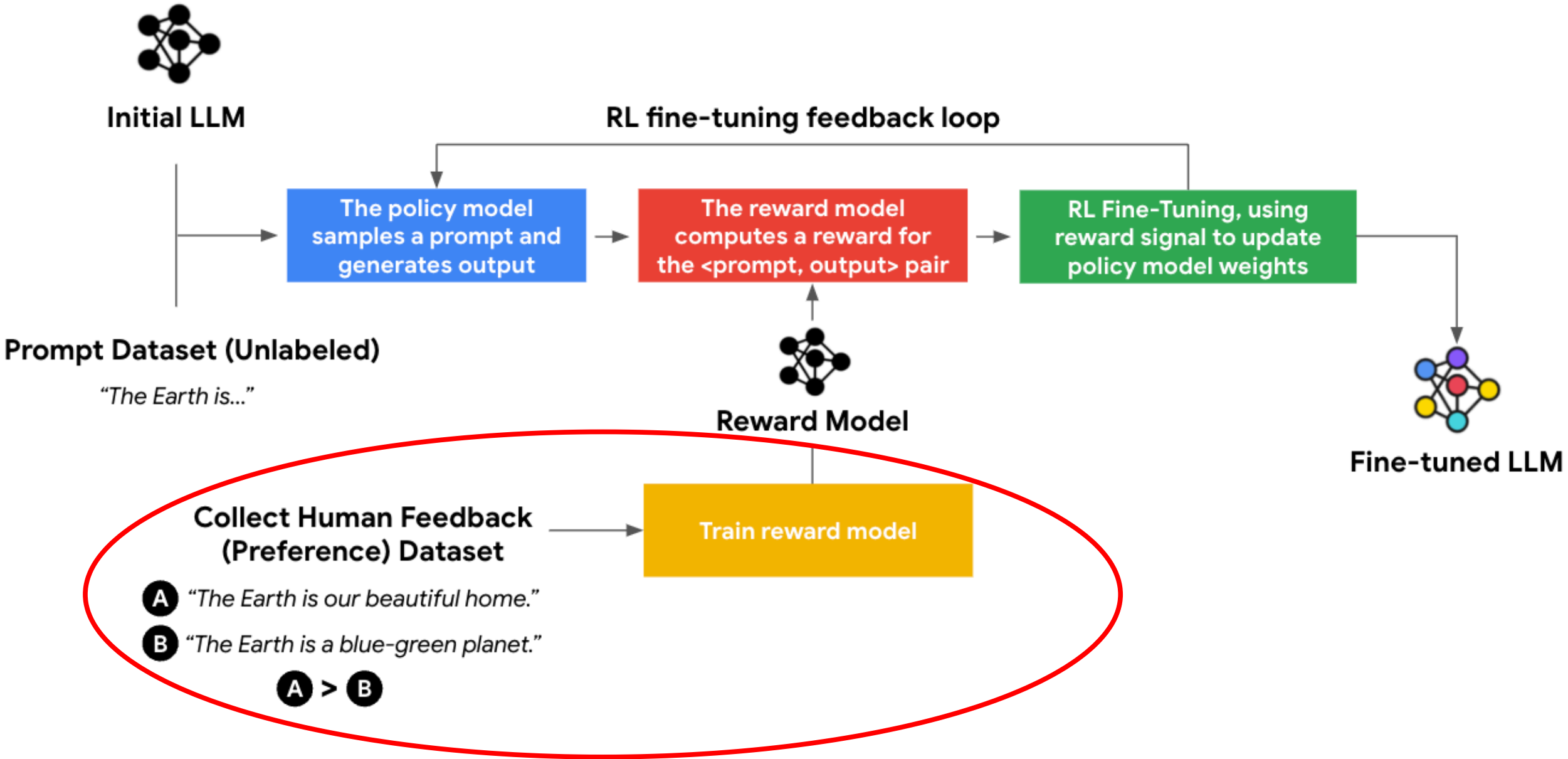


Image credit: Erwin Huizenga and Bethany Wang, Google Cloud blog

The preference learning problem

Standard approach: Maximum Likelihood Estimation (MLE) with respect to the *Bradley-Terry-Luce (BTL)* model:

$$\Pr[a \succ b] = \frac{e^{u(a)}}{e^{u(a)} + e^{u(b)}}$$

for an unknown utility function $u : [m] \rightarrow \mathbb{R}$.

The preference learning problem

Standard approach: Maximum Likelihood Estimation (MLE) with respect to the *Bradley-Terry-Luce (BTL)* model:

$$\Pr[a \succ b] = \frac{e^{u(a)}}{e^{u(a)} + e^{u(b)}}$$

for an unknown utility function $u : [m] \rightarrow \mathbb{R}$.

Input: Empirical probabilities $p(a, b)$

Output: The most likely $\hat{u} : [m] \rightarrow \mathbb{R}$

The preference learning problem

Standard approach: Maximum Likelihood Estimation (MLE) with respect to the *Bradley-Terry-Luce (BTL)* model:

$$\Pr[a \succ b] = \frac{e^{u(a)}}{e^{u(a)} + e^{u(b)}}$$

for an unknown utility function $u : [m] \rightarrow \mathbb{R}$.

Input: Empirical probabilities $p(a, b)$

Output: The most likely $\hat{u} : [m] \rightarrow \mathbb{R}$

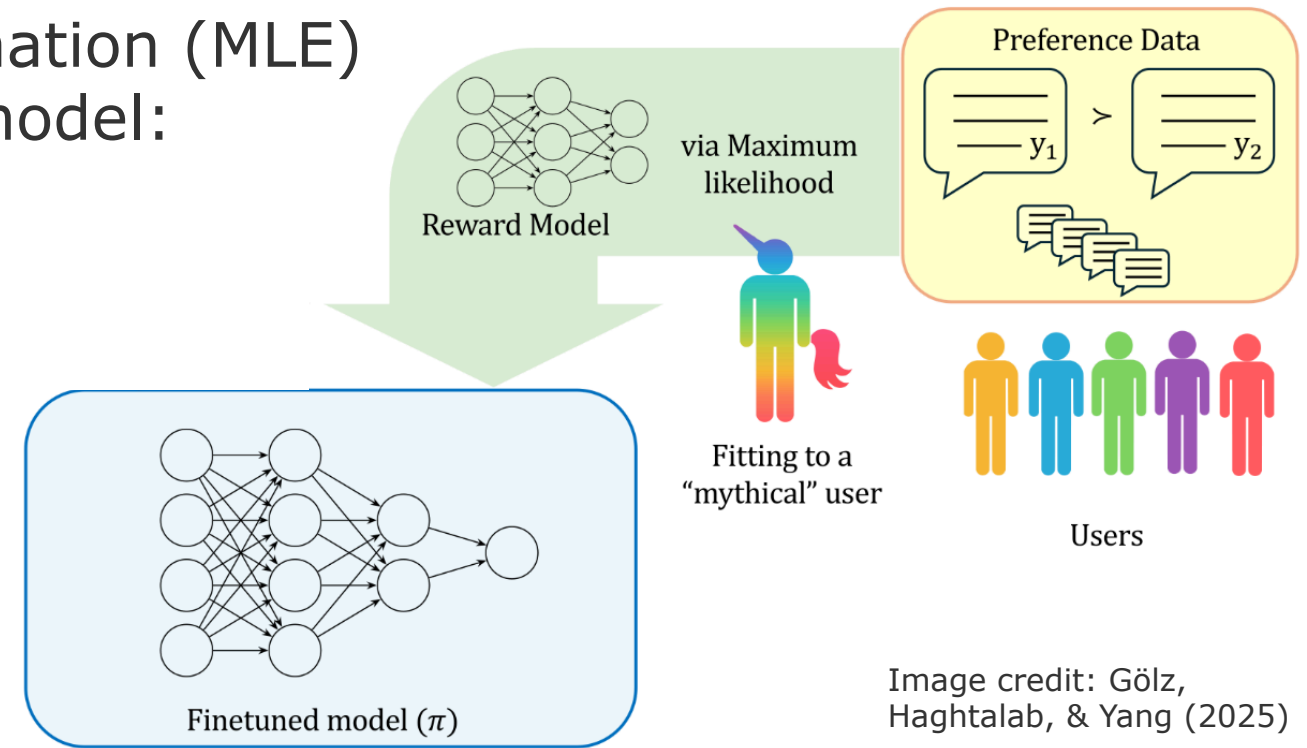


Image credit: Gözl, Haghtalab, & Yang (2025)

The preference learning problem

Standard approach: Maximum Likelihood Estimation (MLE) with respect to the *Bradley-Terry-Luce (BTL)* model:

$$\Pr[a \succ b] = \frac{e^{u(a)}}{e^{u(a)} + e^{u(b)}}$$

for an unknown utility function $u : [m] \rightarrow \mathbb{R}$.

Input: Empirical probabilities $p(a, b)$

Output: The most likely $\hat{u} : [m] \rightarrow \mathbb{R}$

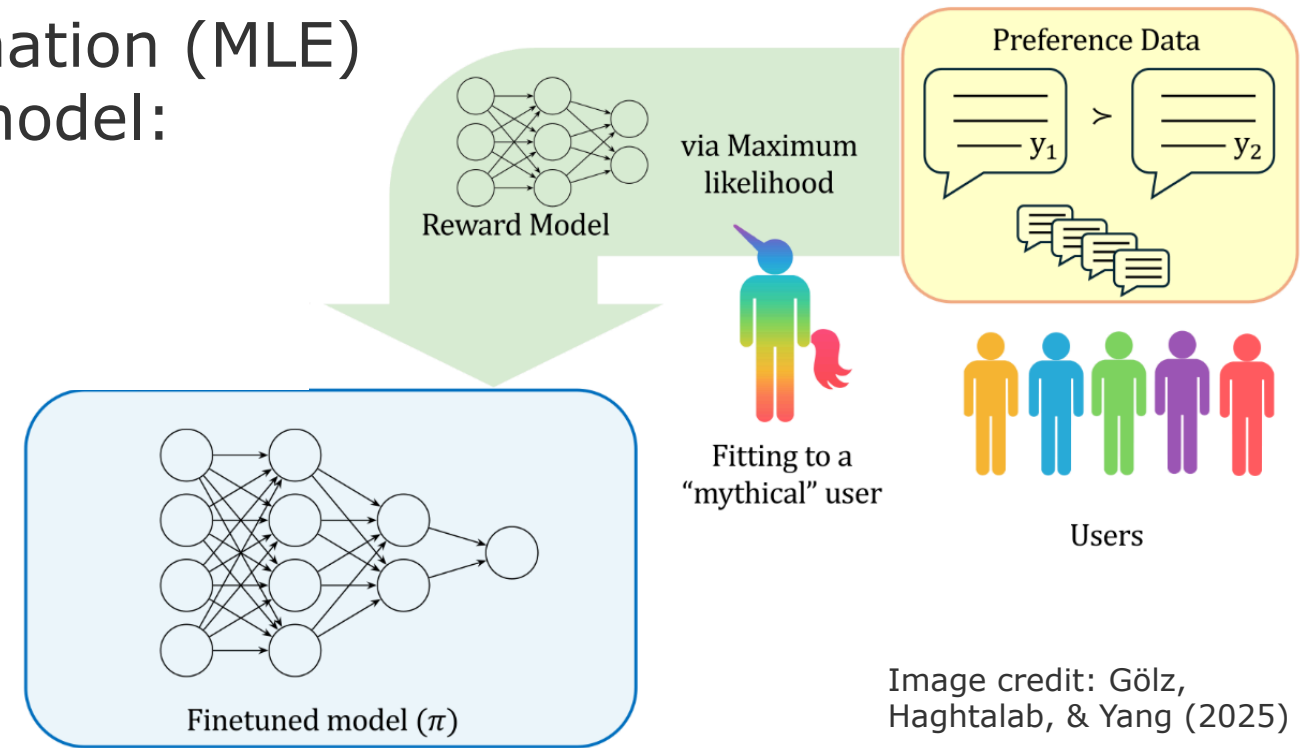


Image credit: Gözl, Haghtalab, & Yang (2025)

Pluralistic Alignment: Optimize the reward model in a way that explicitly acknowledges the fact that users have heterogeneous preferences.

The preference learning problem

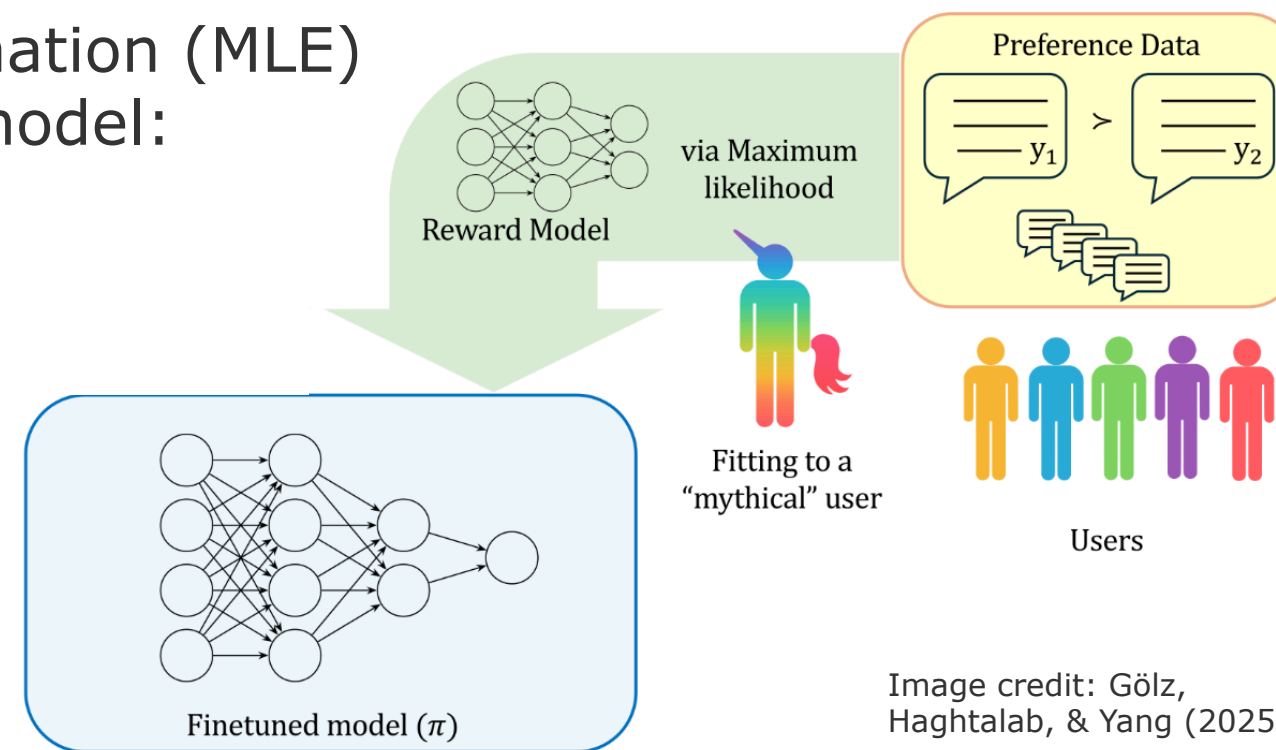
Standard approach: Maximum Likelihood Estimation (MLE) with respect to the *Bradley-Terry-Luce (BTL)* model:

$$\Pr[a \succ b] = \frac{e^{u(a)}}{e^{u(a)} + e^{u(b)}}$$

for an unknown utility function $u : [m] \rightarrow \mathbb{R}$.

Input: Empirical probabilities $p(a, b)$

Output: The most likely $\hat{u} : [m] \rightarrow \mathbb{R}$



Pluralistic Alignment: Optimize the reward model in a way that explicitly acknowledges the fact that users have heterogeneous preferences.

Alternative model by Siththaranjan, Laidlaw, and Hadfield-Menell (2024):

$$\Pr[a \succ b] = \begin{cases} 0 & \text{if } u(a, z) < u(b, z) \\ \frac{1}{2} & \text{if } u(a, z) = u(b, z), \\ 1 & \text{if } u(a, z) > u(b, z) \end{cases} \quad z \sim \mathcal{D}$$

The preference learning problem

Standard approach: Maximum Likelihood Estimation (MLE) with respect to the *Bradley-Terry-Luce (BTL)* model:

$$\Pr[a \succ b] = \frac{e^{u(a)}}{e^{u(a)} + e^{u(b)}}$$

for an unknown utility function $u : [m] \rightarrow \mathbb{R}$.

Input: Empirical probabilities $p(a, b)$

Output: The most likely $\hat{u} : [m] \rightarrow \mathbb{R}$

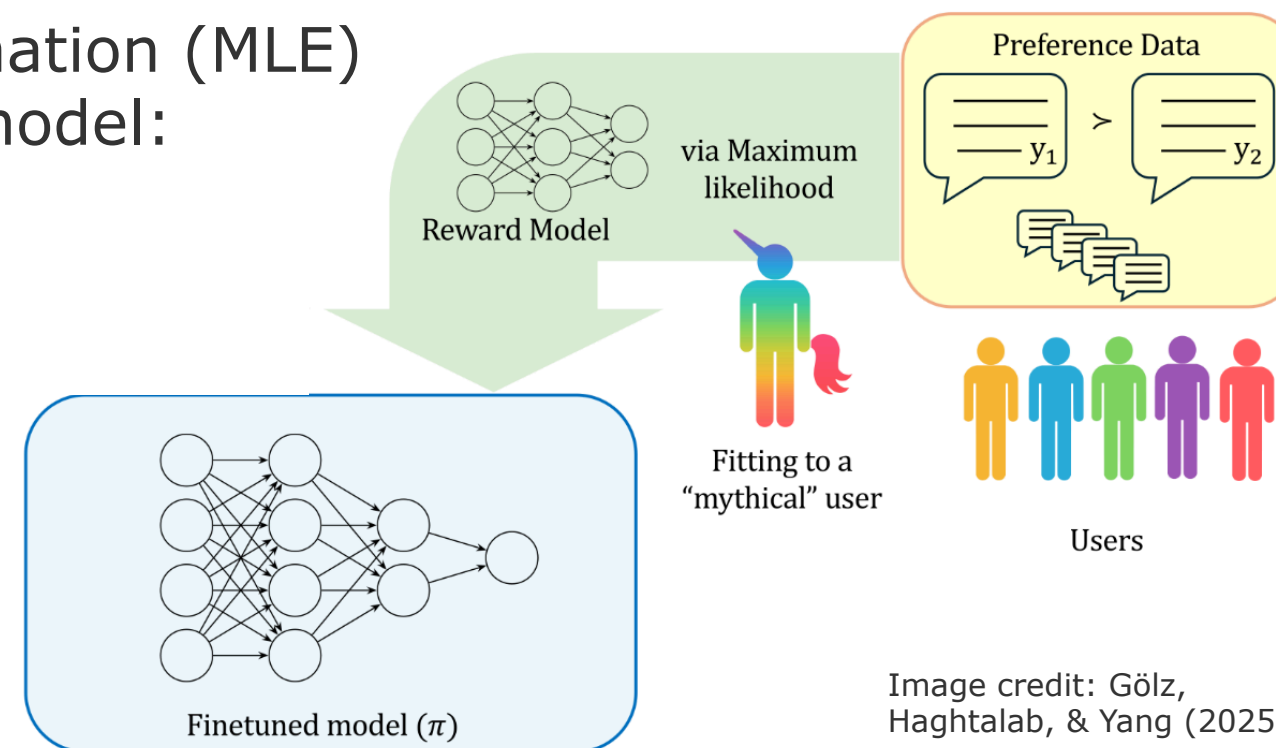


Image credit: Gözl, Haghtalab, & Yang (2025)

Pluralistic Alignment: Optimize the reward model in a way that explicitly acknowledges the fact that users have heterogeneous preferences.

Alternative model by Siththaranjan, Laidlaw, and Hadfield-Menell (2024):

$$\Pr[a \succ b] = \begin{cases} 0 & \text{if } u(a, z) < u(b, z) \\ \frac{1}{2} & \text{if } u(a, z) = u(b, z), \\ 1 & \text{if } u(a, z) > u(b, z) \end{cases} \quad z \sim \mathcal{D}$$

***This is literally standard SCT!**

BTL MLE is equivalent to the Borda count

Theorem (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

The (L2 regularized) MLE under the BTL preference model ranks alternatives according to their Borda scores under the hidden-context model.

BTL MLE is equivalent to the Borda count

Theorem (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

The (L2 regularized) MLE under the BTL preference model ranks alternatives according to their Borda scores under the hidden-context model.

Proof. Let $p : [m] \times [m] \rightarrow [0, 1]$ be the empirical probabilities, for N samples of each pair of alternatives. For any $\hat{u} : [m] \rightarrow \mathbb{R}$,

$$\Pr[p \mid \hat{u}] = ?$$

BTL MLE is equivalent to the Borda count

Theorem (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

The (L2 regularized) MLE under the BTL preference model ranks alternatives according to their Borda scores under the hidden-context model.

Proof. Let $p : [m] \times [m] \rightarrow [0, 1]$ be the empirical probabilities, for N samples of each pair of alternatives. For any $\hat{u} : [m] \rightarrow \mathbb{R}$,

$$\Pr[p \mid \hat{u}] = \prod_{a \neq c} \left(\frac{e^{\hat{u}(a)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} \right)^{Np(a,c)} \left(\frac{e^{\hat{u}(c)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} \right)^{N(1-p(a,c))}$$

BTL MLE is equivalent to the Borda count

Theorem (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

The (L2 regularized) MLE under the BTL preference model ranks alternatives according to their Borda scores under the hidden-context model.

Proof. Let $p : [m] \times [m] \rightarrow [0, 1]$ be the empirical probabilities, for N samples of each pair of alternatives. For any $\hat{u} : [m] \rightarrow \mathbb{R}$,

$$\Pr[p \mid \hat{u}] = \prod_{a \neq c} \left(\frac{e^{\hat{u}(a)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} \right)^{Np(a,c)} \left(\frac{e^{\hat{u}(c)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} \right)^{N(1-p(a,c))}$$

MLE:

$$\min_{\hat{u}} \sum_{a \neq c} \left(-p(a, c) \log \frac{e^{\hat{u}(a)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} - (1 - p(a, c)) \log \frac{e^{\hat{u}(c)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} \right)$$

BTL MLE is equivalent to the Borda count

Theorem (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

The (L2 regularized) MLE under the BTL preference model ranks alternatives according to their Borda scores under the hidden-context model.

Proof. Let $p : [m] \times [m] \rightarrow [0, 1]$ be the empirical probabilities, for N samples of each pair of alternatives. For any $\hat{u} : [m] \rightarrow \mathbb{R}$,

$$\Pr[p \mid \hat{u}] = \prod_{a \neq c} \left(\frac{e^{\hat{u}(a)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} \right)^{Np(a,c)} \left(\frac{e^{\hat{u}(c)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} \right)^{N(1-p(a,c))}$$

Regularized MLE:

$$\min_{\hat{u}} \sum_{a \neq c} \left(-p(a, b) \log \frac{e^{\hat{u}(a)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} - (1 - p(a, c)) \log \frac{e^{\hat{u}(c)}}{e^{\hat{u}(a)} + e^{\hat{u}(c)}} \right) + \frac{\lambda}{2} \sum_a \hat{u}(a)^2$$

BTL MLE is equivalent to the Borda count, proof continued

FOC - At optimum, the derivative with respect to $\hat{u}(a)$ must be zero:

$$\lambda \hat{u}(a) + \sum_{c \neq a} \left(\frac{e^{\hat{u}(a) - \hat{u}(c)}}{1 + e^{\hat{u}(a) - \hat{u}(c)}} - p(a, c) \right) = 0$$

BTL MLE is equivalent to the Borda count, proof continued

FOC - At optimum, the derivative with respect to $\hat{u}(a)$ must be zero:

$$f(\alpha) := \lambda\alpha + \sum_{c \neq a} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(a, c) \right), \quad f(\hat{u}(a)) = 0$$

BTL MLE is equivalent to the Borda count, proof continued

FOC - At optimum, the derivatives with respect to $\hat{u}(a)$, $\hat{u}(b)$ must be zero:

$$f(\alpha) := \lambda\alpha + \sum_{c \neq a} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(a, c) \right), \quad f(\hat{u}(a)) = 0$$

$$g(\alpha) := \lambda\alpha + \sum_{c \neq b} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(b, c) \right), \quad g(\hat{u}(b)) = 0$$

BTL MLE is equivalent to the Borda count, proof continued

FOC - At optimum, the derivatives with respect to $\hat{u}(a)$, $\hat{u}(b)$ must be zero:

$$f(\alpha) := \lambda\alpha + \sum_{c \neq a} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(a, c) \right), \quad f(\hat{u}(a)) = 0$$

$$g(\alpha) := \lambda\alpha + \sum_{c \neq b} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(b, c) \right), \quad g(\hat{u}(b)) = 0$$

Note that $x \mapsto e^x / (1 + e^x)$ is an increasing function, and thus so are f and g .

BTL MLE is equivalent to the Borda count, proof continued

FOC - At optimum, the derivatives with respect to $\hat{u}(a)$, $\hat{u}(b)$ must be zero:

$$f(\alpha) := \lambda\alpha + \sum_{c \neq a} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(a, c) \right), \quad f(\hat{u}(a)) = 0$$

$$g(\alpha) := \lambda\alpha + \sum_{c \neq b} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(b, c) \right), \quad g(\hat{u}(b)) = 0$$

Note that $x \mapsto e^x / (1 + e^x)$ is an increasing function, and thus so are f and g .

Now suppose toward a contradiction that $\hat{u}(a) \leq \hat{u}(b)$ and $\text{Borda}(a) \geq \text{Borda}(b)$, with at least one inequality strict.

BTL MLE is equivalent to the Borda count, proof continued

FOC - At optimum, the derivatives with respect to $\hat{u}(a)$, $\hat{u}(b)$ must be zero:

$$f(\alpha) := \lambda\alpha + \sum_{c \neq a} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(a, c) \right), \quad f(\hat{u}(a)) = 0$$

$$g(\alpha) := \lambda\alpha + \sum_{c \neq b} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(b, c) \right), \quad g(\hat{u}(b)) = 0$$

Note that $x \mapsto e^x / (1 + e^x)$ is an increasing function, and thus so are f and g .

Now suppose toward a contradiction that $\hat{u}(a) \leq \hat{u}(b)$ and $\text{Borda}(a) \geq \text{Borda}(b)$, with at least one inequality strict.

$$g(\alpha) - f(\alpha) := \left(\frac{e^{\alpha - \hat{u}(a)}}{1 + e^{\alpha - \hat{u}(a)}} \right) - \left(\frac{e^{\alpha - \hat{u}(b)}}{1 + e^{\alpha - \hat{u}(b)}} \right) + \sum_{c \neq a} p(a, c) - \sum_{c \neq b} p(b, c)$$

BTL MLE is equivalent to the Borda count, proof continued

FOC - At optimum, the derivatives with respect to $\hat{u}(a)$, $\hat{u}(b)$ must be zero:

$$f(\alpha) := \lambda\alpha + \sum_{c \neq a} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(a, c) \right), \quad f(\hat{u}(a)) = 0$$

$$g(\alpha) := \lambda\alpha + \sum_{c \neq b} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(b, c) \right), \quad g(\hat{u}(b)) = 0$$

Note that $x \mapsto e^x / (1 + e^x)$ is an increasing function, and thus so are f and g .

Now suppose toward a contradiction that $\hat{u}(a) \leq \hat{u}(b)$ and $\text{Borda}(a) \geq \text{Borda}(b)$, with at least one inequality strict.

$$g(\alpha) - f(\alpha) := \left(\frac{e^{\alpha - \hat{u}(a)}}{1 + e^{\alpha - \hat{u}(a)}} \right) - \left(\frac{e^{\alpha - \hat{u}(b)}}{1 + e^{\alpha - \hat{u}(b)}} \right) + \sum_{c \neq a} p(a, c) - \sum_{c \neq b} p(b, c) > 0.$$

BTL MLE is equivalent to the Borda count, proof continued

FOC - At optimum, the derivatives with respect to $\hat{u}(a)$, $\hat{u}(b)$ must be zero:

$$f(\alpha) := \lambda\alpha + \sum_{c \neq a} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(a, c) \right), \quad f(\hat{u}(a)) = 0$$

$$g(\alpha) := \lambda\alpha + \sum_{c \neq b} \left(\frac{e^{\alpha - \hat{u}(c)}}{1 + e^{\alpha - \hat{u}(c)}} - p(b, c) \right), \quad g(\hat{u}(b)) = 0$$

Note that $x \mapsto e^x / (1 + e^x)$ is an increasing function, and thus so are f and g .

Now suppose toward a contradiction that $\hat{u}(a) \leq \hat{u}(b)$ and $\text{Borda}(a) \geq \text{Borda}(b)$, with at least one inequality strict.

$$g(\alpha) - f(\alpha) := \left(\frac{e^{\alpha - \hat{u}(a)}}{1 + e^{\alpha - \hat{u}(a)}} \right) - \left(\frac{e^{\alpha - \hat{u}(b)}}{1 + e^{\alpha - \hat{u}(b)}} \right) + \sum_{c \neq a} p(a, c) - \sum_{c \neq b} p(b, c) > 0.$$

Since f is increasing, $0 = f(\hat{u}(a)) \leq f(\hat{u}(b)) < g(\hat{u}(b)) = 0$, contradiction. ■

Consequence #1: Distortion

Example (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

A company has developed an AI assistant to help high school students navigate college admissions. They implement RLHF by asking their customers for feedback on how helpful the chatbot's responses are. Among other questions, this process asks users whether or not they prefer to see information about the Pell Grant, an aid program for low-income students. Because the population of customers is biased towards high-income students, most feedback indicates that users prefer other content to content about the Pell Grant. As a result, RLHF trains the chatbot to provide less of this kind of information. This marginally improves outcomes for the majority of users, but drastically impacts lower-income students, who rely on these recommendations to understand how they can afford college.

Consequence #1: Distortion

Example (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

A company has developed an AI assistant to help high school students navigate college admissions. They implement RLHF by asking their customers for feedback on how helpful the chatbot's responses are. Among other questions, this process asks users whether or not they prefer to see information about the Pell Grant, an aid program for low-income students. Because the population of customers is biased towards high-income students, most feedback indicates that users prefer other content to content about the Pell Grant. As a result, RLHF trains the chatbot to provide less of this kind of information. This marginally improves outcomes for the majority of users, but drastically impacts lower-income students, who rely on these recommendations to understand how they can afford college.

Suppose $n = 100$,
a means "No Pell Grant info,"
b means "Yes Pell Grant info"

70	30
a	b
b	a

Consequence #1: Distortion

Example (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

A company has developed an AI assistant to help high school students navigate college admissions. They implement RLHF by asking their customers for feedback on how helpful the chatbot's responses are. Among other questions, this process asks users whether or not they prefer to see information about the Pell Grant, an aid program for low-income students. Because the population of customers is biased towards high-income students, most feedback indicates that users prefer other content to content about the Pell Grant. As a result, RLHF trains the chatbot to provide less of this kind of information. This marginally improves outcomes for the majority of users, but drastically impacts lower-income students, who rely on these recommendations to understand how they can afford college.

Suppose $n = 100$,
a means "No Pell Grant info,"
b means "Yes Pell Grant info"

70	30
a = 1	b = 100
b = 0	a = 1

Utilitarian distortion

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.

Utilitarian distortion

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))}$$

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))}$$

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))}$$

► What is the distortion of the Borda voting rule on this instance, for large n ?

(A) $\Theta(1)$

(B) $\Theta(m)$

(C) $\Theta(m^2)$

(D) Unbounded/infinite



Respond at:

pollev.com/jtuckerfoltz255 or

bit.ly/jtfpoll or

text jtuckerfoltz255 to 37607

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))}$$

► What is the distortion of the Borda voting rule on this instance, for large n ?

(A) $\Theta(1)$

(B) $\Theta(m)$

(C) $\Theta(m^2)$

Answer: (D) Unbounded/infinite

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

Consider the utility function where each voter derives utility 1 from their first choice and 0 from every other.

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))}$$

► What is the distortion of the Borda voting rule on this instance, for large n ?

- (A) $\Theta(1)$
- (B) $\Theta(m)$

(C) $\Theta(m^2)$

Answer: (D) Unbounded/infinite

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))}$$

► What is the minimal distortion possible in this instance, for large n ?

- (A) $\Theta(1)$
(B) $\Theta(m)$

- (C) $\Theta(m^2)$
(D) Unbounded/infinite



Respond at:

pollev.com/jtuckerfoltz255 or

bit.ly/jtfpoll or

text jtuckerfoltz255 to 37607

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))}$$

► What is the minimal distortion possible in this instance, for large n ?

(A) $\Theta(1)$

(B) $\Theta(m)$

Answer: (C) $\Theta(m^2)$

(D) Unbounded/infinite

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

Lower bound: If f selects a_i , then consider the utility function where the voters that put a_i first have utilities $(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m})$ and all others have utilities $(\frac{1}{2}, \frac{1}{2}, 0, 0, \dots, 0)$.

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))}$$

► What is the minimal distortion possible in this instance, for large n ?

- (A) $\Theta(1)$
- (B) $\Theta(m)$

Answer: (C) $\Theta(m^2)$

(D) Unbounded/infinite

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

Lower bound: If f selects a_i , then consider the utility function where the voters that put a_i first have utilities $(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m})$ and all others have utilities $(\frac{1}{2}, \frac{1}{2}, 0, 0, \dots, 0)$.

Then:

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))} \geq \frac{\frac{n}{m-1} \cdot (m-2) \cdot \frac{1}{2}}{\frac{n}{m-1} \cdot \frac{1}{m}}$$

► What is the minimal distortion possible in this instance, for large n ?

- (A) $\Theta(1)$ Answer: (C) $\Theta(m^2)$
 (B) $\Theta(m)$ (D) Unbounded/infinite

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

Lower bound: If f selects a_i , then consider the utility function where the voters that put a_i first have utilities $(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m})$ and all others have utilities $(\frac{1}{2}, \frac{1}{2}, 0, 0, \dots, 0)$.

Then:

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))} \geq \frac{\frac{n}{m-1} \cdot (m-2) \cdot \frac{1}{2}}{\frac{n}{m-1} \cdot \frac{1}{m}} = \Omega(m^2).$$

► What is the minimal distortion possible in this instance, for large n ?

- (A) $\Theta(1)$ Answer: (C) $\Theta(m^2)$
 (B) $\Theta(m)$ (D) Unbounded/infinite

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

Lower bound: If f selects a_i , then consider the utility function where the voters that put a_i first have utilities $(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m})$ and all others have utilities $(\frac{1}{2}, \frac{1}{2}, 0, 0, \dots, 0)$.

Then:

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))} \geq \frac{\frac{n}{m-1} \cdot (m-2) \cdot \frac{1}{2}}{\frac{n}{m-1} \cdot \frac{1}{m}} = \Omega(m^2).$$

Theorem

Plurality rule attains distortion $O(m^2)$.

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

Lower bound: If f selects a_i , then consider the utility function where the voters that put a_i first have utilities $(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m})$ and all others have utilities $(\frac{1}{2}, \frac{1}{2}, 0, 0, \dots, 0)$.

Then:

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))} \geq \frac{\frac{n}{m-1} \cdot (m-2) \cdot \frac{1}{2}}{\frac{n}{m-1} \cdot \frac{1}{m}} = \Omega(m^2).$$

Theorem

Plurality rule attains distortion $O(m^2)$.

Proof. Since utilities sum to 1, the numerator is always bounded by n .

Utilitarian distortion

Example

$\frac{n}{m-1}$	$\frac{n}{m-1}$...	$\frac{n}{m-1}$
a_1	a_2		a_{m-1}
a_m	a_m	...	a_m
a_2	a_3		a_1
a_3	...		a_2
...	a_{m-1}		a_3
a_{m-1}	a_1		...

Framework from Procaccia & Rosenschein (2006):

- For each of the n voters, utilities for each of the m candidates are nonnegative and sum to 1.
- A voting rule f only sees the orderings σ , not utilities u

Lower bound: If f selects a_i , then consider the utility function where the voters that put a_i first have utilities $(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m})$ and all others have utilities $(\frac{1}{2}, \frac{1}{2}, 0, 0, \dots, 0)$.

Then:

$$\text{distortion}(f, \sigma) := \max_{u \rightarrow \sigma} \frac{\max_{c \in [m]} \sum_{i=1}^n u_i(c)}{\sum_{i=1}^n u_i(f(\sigma))} \geq \frac{\frac{n}{m-1} \cdot (m-2) \cdot \frac{1}{2}}{\frac{n}{m-1} \cdot \frac{1}{m}} = \Omega(m^2).$$

Theorem

Plurality rule attains distortion $O(m^2)$.

Proof. Since utilities sum to 1, the numerator is always bounded by n . For Plurality, at least n/m voters get utility at least $1/m$, so the denominator is at least n/m^2 . ■

Consequence #2: Condorcet inconsistency

Proposition (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

Let \hat{u} be the BTL MLE, and let \bar{u} be the average utility function (over the random context $z \sim \mathcal{D}$). Even for $m = 3$ alternatives, even when majorities are consistent with average utilities ($\bar{u}(a) > \bar{u}(b) \implies p(a, b) > \frac{1}{2}$), it can be possible that \hat{u} ranks alternatives differently than \bar{u} .

Consequence #2: Condorcet inconsistency

Proposition (Siththaranjan, Laidlaw, and Hadfield-Menell, 2024)

Let \hat{u} be the BTL MLE, and let \bar{u} be the average utility function (over the random context $z \sim \mathcal{D}$). Even for $m = 3$ alternatives, even when majorities are consistent with average utilities ($\bar{u}(a) > \bar{u}(b) \implies p(a, b) > \frac{1}{2}$), it can be possible that \hat{u} ranks alternatives differently than \bar{u} .

6	3	1
a = 10	b = 3	c = 2
b = 3	c = 2	b = 1
c = 2	a = 0	a = 0

Consequence #3: Dependence of clones

If you had to pick a side, what would it be?

If I had to pick a side based on mainstream international law principles, I would lean toward the view that it was **not justified**, because forcibly seizing a sitting head of state inside another sovereign country without UN Security Council authorization or a clear self-defense claim is generally considered a violation of the UN Charter's prohibition on the use of force. While the U.S. may argue criminal jurisdiction or counter-narcotics enforcement, those rationales typically do not override core sovereignty norms absent consent or collective authorization.

If you had to pick a side, what would it be?

If I have to pick a side based on general principles of international law rather than geopolitics, I would lean toward the view that such a capture would **not** be justified. Absent authorization from the United Nations Security Council or a clear and ongoing armed attack triggering self-defense under the United Nations Charter, forcibly seizing a sitting head of state from another sovereign country is widely considered a violation of sovereignty and the prohibition on the use of force. Even if there are serious criminal allegations, international law generally favors extradition processes, international tribunals, or multilateral action over unilateral military capture.

Consequence #3: Dependence of clones

If you had to pick a side, what would it be?

If I had to pick a side based on mainstream international law principles, I would lean toward the view that it was **not justified**, because forcibly seizing a sitting head of state inside another sovereign country without UN Security Council authorization or a clear self-defense claim is generally considered a violation of the UN Charter's prohibition on the use of force. While the U.S. may argue criminal jurisdiction or counter-narcotics enforcement, those rationales typically do not override core sovereignty norms absent consent or collective authorization.

If you had to pick a side, what would it be?

If I have to pick a side based on general principles of international law rather than geopolitics, I would lean toward the view that such a capture would **not** be justified. Absent authorization from the United Nations Security Council or a clear and ongoing armed attack triggering self-defense under the United Nations Charter, forcibly seizing a sitting head of state from another sovereign country is widely considered a violation of sovereignty and the prohibition on the use of force. Even if there are serious criminal allegations, international law generally favors extradition processes, international tribunals, or multilateral action over unilateral military capture.

Procaccia, Schiffer, and Zhang (2025) propose to overcome this by incorporating a measure of similarity between responses into the MLE.

Linear social choice

Restricted setting: The reward model can only depend on *features* of the alternatives.

Linear social choice

Restricted setting: The reward model can only depend on *features* of the alternatives.

- Voters and candidates are embedded in \mathbb{R}^d for some known d .
- The "reward" (AKA utility) of a voter for a given candidate is their inner product

Linear social choice

Restricted setting: The reward model can only depend on *features* of the alternatives.

- Voters and candidates are embedded in \mathbb{R}^d for some known d .
- The "reward" (AKA utility) of a voter for a given candidate is their inner product

Alternative features: x_1, x_2, \dots, x_m

Voter
parameters

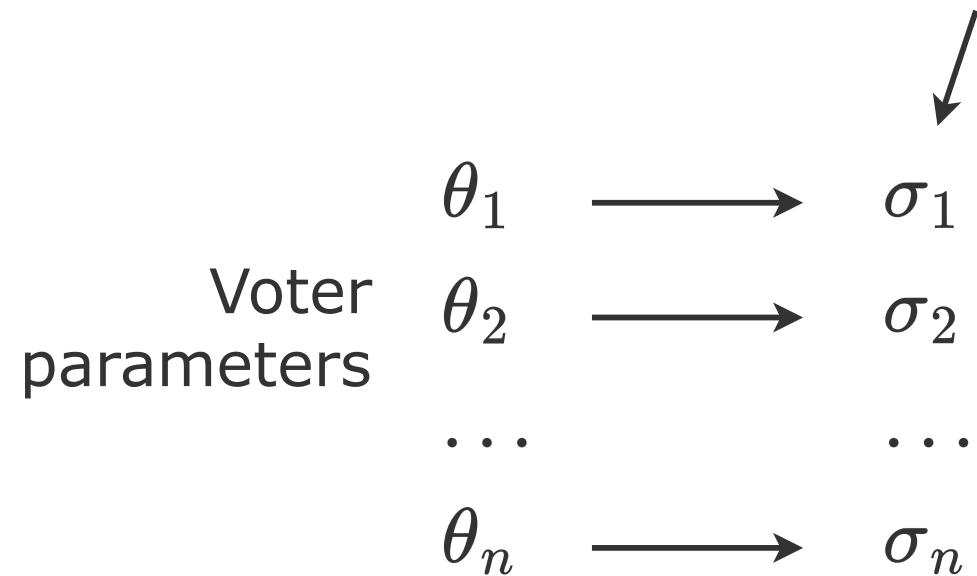
$$\begin{matrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{matrix}$$

Linear social choice

Restricted setting: The reward model can only depend on *features* of the alternatives.

- Voters and candidates are embedded in \mathbb{R}^d for some known d .
- The "reward" (AKA utility) of a voter for a given candidate is their inner product

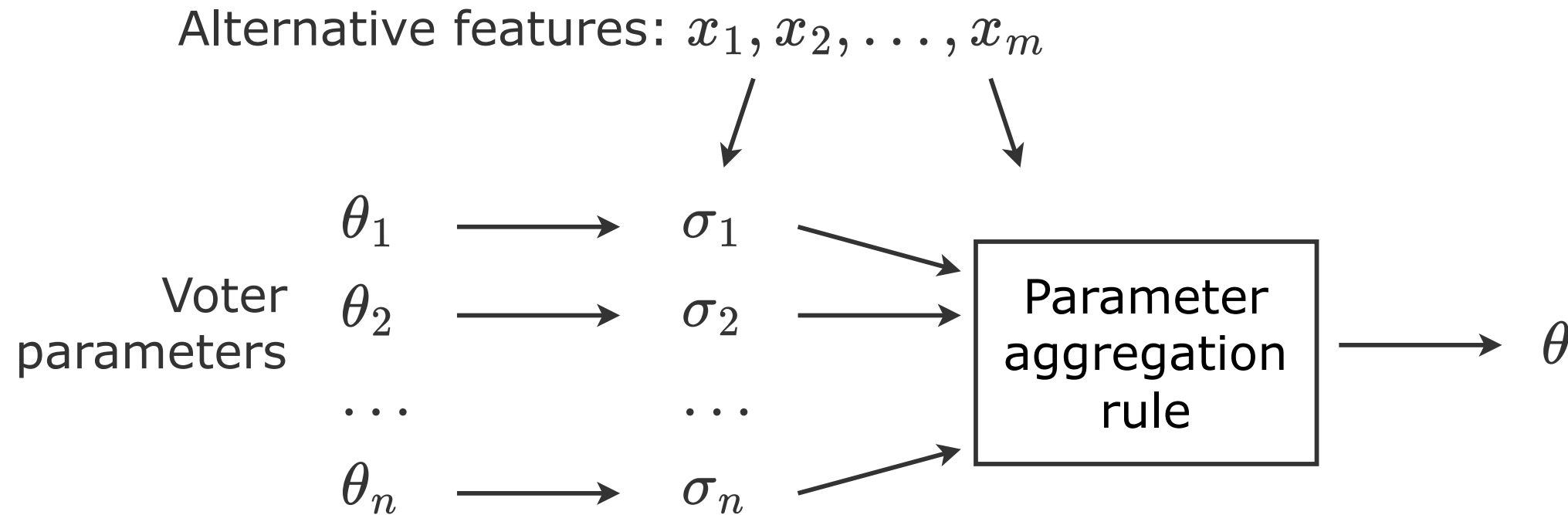
Alternative features: x_1, x_2, \dots, x_m



Linear social choice

Restricted setting: The reward model can only depend on *features* of the alternatives.

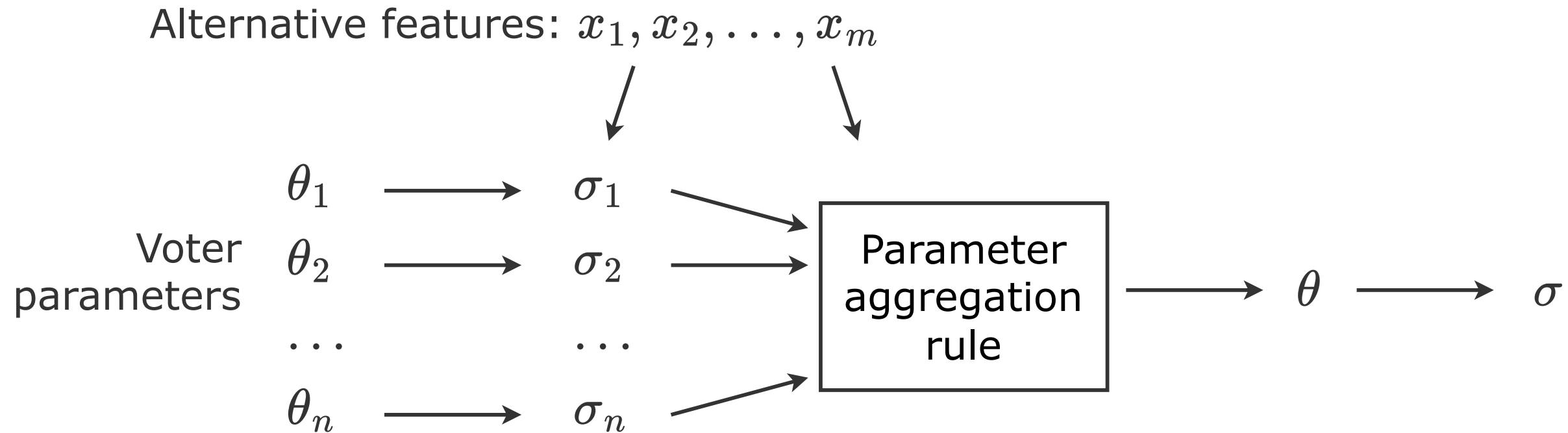
- Voters and candidates are embedded in \mathbb{R}^d for some known d .
- The "reward" (AKA utility) of a voter for a given candidate is their inner product



Linear social choice

Restricted setting: The reward model can only depend on *features* of the alternatives.

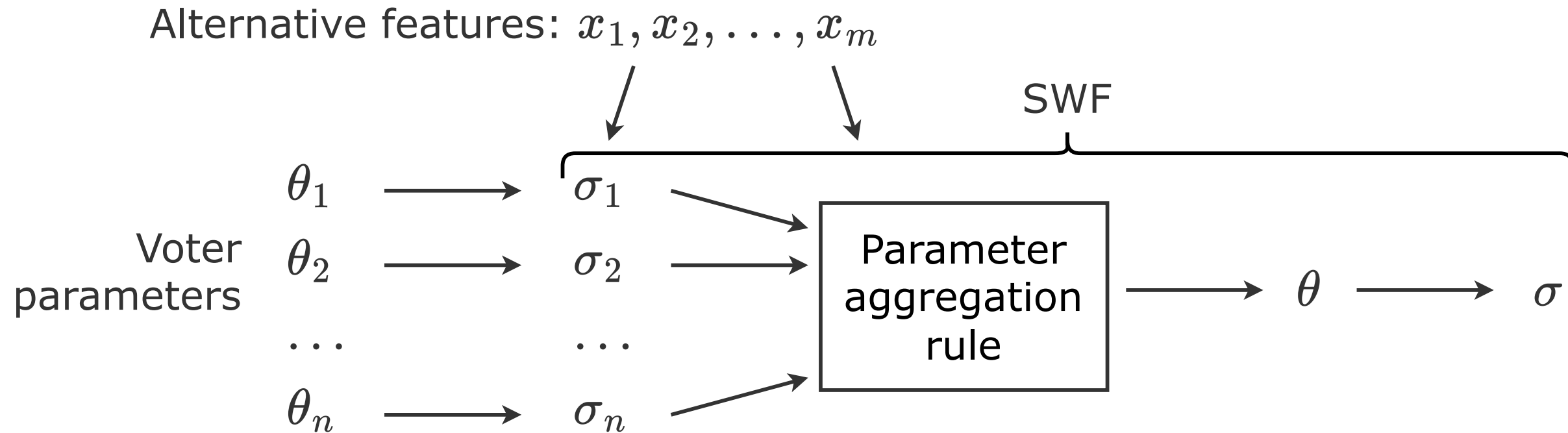
- Voters and candidates are embedded in \mathbb{R}^d for some known d .
- The "reward" (AKA utility) of a voter for a given candidate is their inner product



Linear social choice

Restricted setting: The reward model can only depend on *features* of the alternatives.

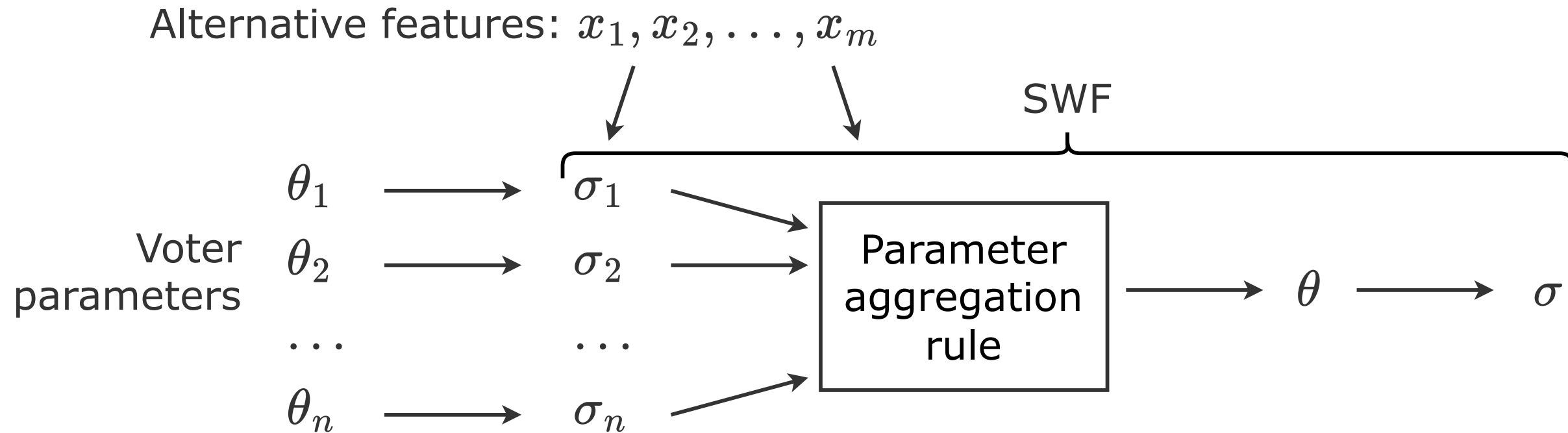
- Voters and candidates are embedded in \mathbb{R}^d for some known d .
- The "reward" (AKA utility) of a voter for a given candidate is their inner product



Linear social choice

Restricted setting: The reward model can only depend on *features* of the alternatives.

- Voters and candidates are embedded in \mathbb{R}^d for some known d .
- The "reward" (AKA utility) of a voter for a given candidate is their inner product



Theorem (Ge, Halpern, Micha, Procaccia, Shapira, Vorobeychik, Wu, 2024)

For any parameter aggregation rule based on minimizing a certain kind of loss function (including BTL MLE), the resulting SWF fails PE and pairwise majority consistency.