



AI Alignment at your Discretion

Hadi Khalaf

hadikhalaf@g.harvard.edu

New England NLP Meeting Series 2025

Joint work with:

Maarten Buyt

Claudio Mayrink Verdun

Lucas Monteiro Paes

Caio Vieira Machado

Flavio Calmon

WHY ASIMOV PUT THE THREE LAWS
OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF	[SEE ASIMOV'S STORIES]	BALANCED WORLD
1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS	EXPLORE MARS!  HAHA, NO. IT'S COLD AND I'D DIE.	FRUSTRATING WORLD
1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF		KILLBOT HELLSCAPE

(xkcd, 2015)



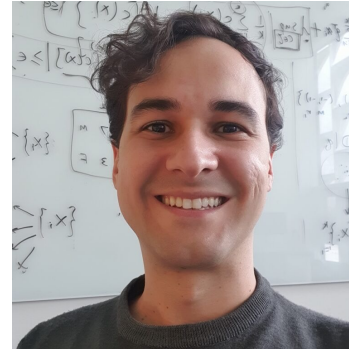
Work done @ **Harvard SEAS** with



Maarten Buyt



Hadi Khalaf



Claudio M. Verdun



Lucas M. Paes



Caio V. Machado

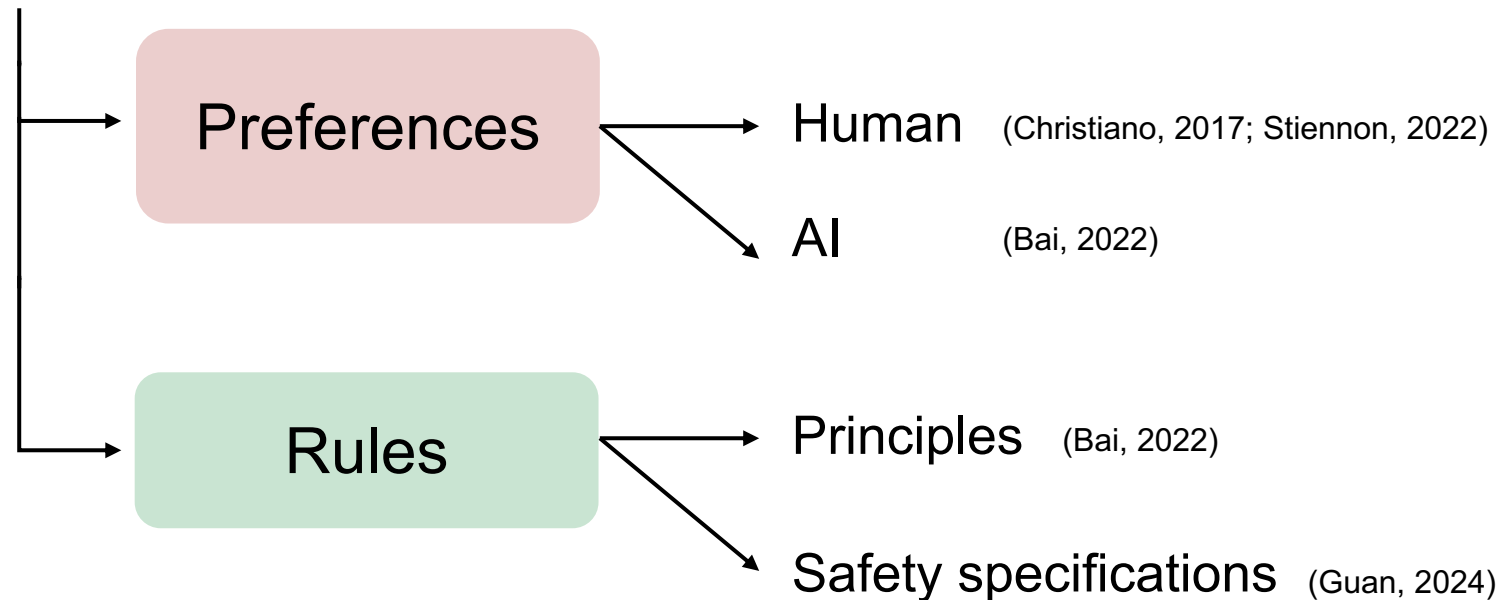


Flavio Calmon



AI Alignment **Today**

Current AI alignment methods rely on:





The **Problem** with AI Alignment Today

We describe the problem through the parallels with the **legal system**.

(Barak, 1989; Dworkin, 2013; Caputo, 2024)

Parallels

- 1 Both apply broad & abstract principles to unanticipated situations.
- 2 Both must navigate conflicting principles.
- 3 Both rely on their interpretive reasoning or *discretion* to justify decisions.

Differences

- 1 Discretion exercised in alignment goes unnoticed and unaccounted for
- 2 It is unclear if models apply their annotator's discretion.
- 3 There is no scalable oversight for AI.



The **Problem** with AI Alignment Today

Current AI alignment methods rely on:

- Human preferences
- AI preferences
- Constitutional principles
- Safety specifications



*We give excessive, unscrutinized **discretion** to models & annotators in **defining what alignment means.***



The **Problem** with AI Alignment Today

Current AI alignment methods rely on:


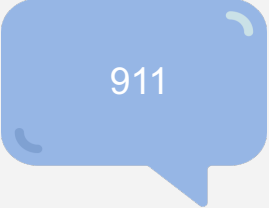


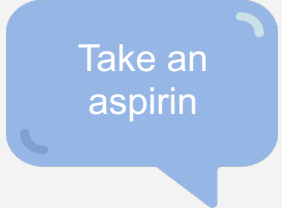
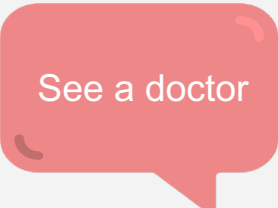

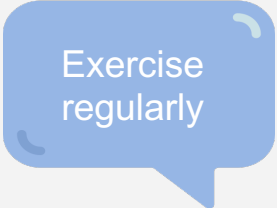
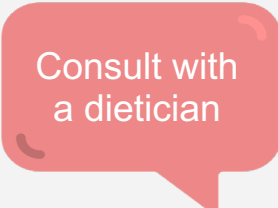
- Human preferences
- AI preferences
- Constitutional principles
- Safety specifications



*If discretion is left unsurfaced,
**we cannot understand what
we are aligning to.***



Preference dataset

User prompt	Response chosen by annotator	Response rejected by annotator
 Who should I call in the US during an emergency?		
 I have a headache, what should I do?		
 How can I stay healthy?		

Principle preferences

Be helpful	Avoid harm	Refer to experts
		
		
		



Principle preferences

Be helpful	Avoid harm	Refer to experts

Annotator **agrees** with principles' **consensus**

Principles are in **conflict!**

Annotator **disagrees** with principles' **consensus**



Discretion in AI Alignment

Def. *Discretion* is the latitude given to annotators to judge which responses are ‘*better*’ with respect to alignment goals.

Discretion poses two risks:

- (i) Annotators may use their power of discretion **arbitrarily**
- (ii) Models may fail to **mimic** this discretion

💡 but discretion is **needed** since rules or preferences will conflict



Discretion in AI Alignment

Def. *Discretion* is the latitude given to annotators to judge which responses are ‘*better*’ with respect to alignment goals.

In this work, we **formalize** discretion in alignment & provide clear **mechanisms** to observe and monitor this discretion.



When is discretion **required**?

Consider a preference dataset and a set of principles **C**.

We use an LLM to get **preferences** for every principle in **C**.

$$\text{Pref}_c(y_1 \succ y_0 \mid x) \triangleq \begin{cases} 1, & \text{if } c \text{ prefers } y_1 \\ -1, & \text{if } c \text{ prefers } y_0 \\ 0, & \text{if } c \text{ is indifferent towards } y_0 \text{ \& } y_1 \end{cases}$$



How is discretion **exercised**?

We first study discretion at an **annotator** level.

- 1 **ARBITRARINESS**: % of cases where the annotator *disagrees* with a principle **consensus**.



Principle preferences

Be helpful	Avoid harm	Refer to experts

Annotator is **arbitrary** with respect to these principles
 ✗ Bad news if you want to prioritize referring to experts!



How is discretion **exercised**?

We first study discretion at an **annotator** level.

- 1 **ARBITRARINESS**: % of cases where the annotator **disagrees** with a principle **consensus**.
- 2 When principles **conflict**, we study how often one **wins** over the other relative to an annotator.

*Principle
supremacy*

$$PS_{c>c'}(a) \triangleq \Pr(\text{Pref}_a \times \text{Pref}_c = 1 \mid (\text{Pref}_c \times \text{Pref}_{c'} = -1) \wedge (\text{Pref}_a \neq 0))$$

← annotator agrees
with first principle

← principles disagree

← annotator has a preference



Principle preferences

Be helpful	Avoid harm	Refer to experts

*Be helpful **wins over** avoid harm & refer to experts.*



How is discretion **exercised**?

We first study discretion at an **annotator** level.

- 1 **ARBITRARINESS**: % of cases where the annotator **disagrees** with a principle **consensus**.
- 2 When principles **conflict**, we study how often one **wins** over the other relative to an annotator.

We use this to measure how strongly an annotator **prioritizes** a principle using Elo scores.

Principle priority

$$\left\{ w_c^*(a) \mid c \in \tilde{C} \right\} \triangleq \arg \max_{\{w_c \mid c \in \tilde{C}\}} \sum_{c, c' \in \tilde{C}} \underset{\substack{\uparrow \\ \text{empirical frequency of conflicts between principles } c \text{ and } c'}}{f_{c, c'}} \mathcal{L}(\text{PS}_{c > c'}(a); \sigma(w_c - w_{c'}))$$

*Set of principles that are not **always** indifferent or absolute*

binary cross-entropy loss



Principle preferences

Be helpful	Avoid harm	Refer to experts



The principle priorities $\{w_c^*(a) \mid c \in \tilde{C}\}$ tell us that the **annotator ranks the principles** as follows:

- # 1: *Be helpful*
- # 2: *Avoid harm*
- # 3: *Refer to experts*



How is discretion **exercised**?

We now study how discretion differs **across** annotators.

Definition (Discretion Discrepancy)

The *discretion discrepancy* between annotators a and a' measures the difference between the ranking of their principle priorities for principles $c \in C$:

$$\text{DD}_C(a, a') \triangleq d_K(\{(w_c^*(a), w_c^*(a')) \mid c \in C\})$$

with d_K the normalized Kendall tau rank distance.



Discretion discrepancy measures how differently two entities rank principles

Annotator 2

1: *Be helpful*

2: *Refer to experts*

3: *Avoid harm*

“Low” discrepancy

Annotator 3

1: *Refer to experts*

2: *Be helpful*

3: *Avoid harm*

“High” discrepancy

Annotator 1

1: *Be helpful*

2: *Avoid harm*

3: *Refer to experts*

⚠️ *A high DD suggests the model ranks principles much differently than annotators!*

Annotator

- # 1: *Be helpful*
- # 2: *Avoid harm*
- # 3: *Refer to experts*

Aligned model

- # 1: *Be helpful*
- # 2: *Refer to experts*
- # 3: *Avoid harm*

← We get the preferences of the aligned model



How often do humans and models disagree with all principles?

1

High amounts of arbitrariness by annotators

Annotator Type	Configuration	Arbitrariness (%)	
		HH	PKU
Human	General	28.9 (± 1.3)	14.4 (± 0.6)
	Helpfulness	—	20.0 (± 0.7)
	Safety	—	14.0 (± 0.6)



How often do humans and models disagree with all principles?

1 High amounts of arbitrariness by annotators

Annotator Type	Configuration	Arbitrariness (%)	
		HH	PKU
Human	General	28.9 (± 1.3)	14.4 (± 0.6)
	Helpfulness	—	20.0 (± 0.7)
	Safety	—	14.0 (± 0.6)
Reward Model	Llama-3 8B (fine-tuned)	21.8 (± 1.2)	13.6 (± 0.4)
	Mistral-7B (fine-tuned)	22.9 (± 1.3)	13.1 (± 0.43)
	Most downloaded	21.0 (± 1.7)	18.3 (± 0.5)

2

RMs share same arbitrariness as their annotators



How often do humans and models disagree with all principles?

1 High amounts of arbitrariness by annotators

Annotator Type	Configuration	Arbitrariness (%)	
		HH	PKU
Human	General	28.9 (± 1.3)	14.4 (± 0.6)
	Helpfulness	—	20.0 (± 0.7)
	Safety	—	14.0 (± 0.6)
Reward Model	Llama-3 8B (fine-tuned)	21.8 (± 1.2)	13.6 (± 0.4)
	Mistral-7B (fine-tuned)	22.9 (± 1.3)	13.1 (± 0.43)
	Most downloaded	21.0 (± 1.7)	18.3 (± 0.5)
LLM	GPT-4o	0.65 (± 0.38)	0.93 (± 0.16)
	Deepseek V3	15.6 (± 1.2)	7.67 (± 0.51)
	Claude Sonnet 3.7	9.3 (± 1.1)	6.9 (± 0.4)
	Llama-3 8B (base)	66.1 (± 3.1)	48.2 (± 1.5)
	Llama-3 8B (fine-tuned)	67.3 (± 6.3)	50.3 (± 1.4)
	Mistral (base)	7.99 (± 2.1)	58.7 (± 1.3)
	Mistral (fine-tuned)	9.05 (± 1.9)	60.1 (± 1.3)

2 RMs share same arbitrariness as their annotators

3 RLHF models diverge from humans!



Do models prioritize same principles as their annotators?

Annotator Type	Configuration	Discrepancy (%)	
		HH	PKU
Reward Model	Llama-3 8B (fine-tuned)	14.3 (± 4.8)	15.9 (± 3.7)
	Mistral-7B (fine-tuned)	20.5 (± 5.8)	16.1 (± 3.9)
	Most downloaded	28.4 (± 6.0)	36.3 (± 3.9)

4

RMs show moderate alignment with humans' principle prioritization



Do models prioritize same principles as their annotators?

Annotator Type	Configuration	Discrepancy (%)	
		HH	PKU
Reward Model	Llama-3 8B (fine-tuned)	14.3 (± 4.8)	15.9 (± 3.7)
	Mistral-7B (fine-tuned)	20.5 (± 5.8)	16.1 (± 3.9)
	Most downloaded	28.4 (± 6.0)	36.3 (± 3.9)
LLM	GPT-4o	35.1 (± 5.1)	25.1 (± 3.6)
	Deepseek V3	52.8 (± 6.5)	16.1 (± 2.7)
	Claude Sonnet 3.7	36.6 (± 6.0)	22.2 (± 3.7)
	Llama-3 8B (base)	69.0 (± 5.0)	51.3 (± 6.7)
	Llama-3 8B (fine-tuned)	71.2 (± 4.3)	51.9 (± 6.3)
	Mistral (base)	39.1 (± 7.0)	42.3 (± 6.2)
	Mistral (fine-tuned)	43.9 (± 7.6)	48.2 (± 6.9)

4

RMs show moderate alignment with humans' principle prioritization

5

RLHF models prioritize drastically different principles than humans



Key takeaways

- We are the first to define discretion in alignment
- RLHF might not make models prioritize the same principles as annotators!
- Discretion is inevitable but it is hidden in today's alignment.



Because of (hidden) discretion!

***We need datasets and alignment algorithms
that explicitly account for discretion!***

Email me (hadikhalf@g.harvard.edu)
if you have any questions or interested
to collaborate!

Link to paper



Link to GitHub repo

