Stanford
ARTIFICIAL
INTELLIGENCE

Section 1: Introduction

Sanmi Koyejo

All Sections
Export PDF

Stanford
ARTIFICIAL
INTELLIGENCE

Table of Contents

Stanford
ARTIFICIAL
INTELLIGENCE

Introduction

"Machine Learning from Human Preferences" explores the challenge of efficiently and effectively eliciting values and preferences from individuals, groups, and societies and embedding them within AI models and applications. Specifically, this course focuses on the **statistical and conceptual foundations** and **strategies for interactively querying humans** to elicit information that can improve learning and applications.

Stanford
ARTIFICIAL
INTELLIGENCE

Foundations and strategies for interactively querying humans to elicit information that can improve learning.

Focus on the role of the **human-in-the-loop** for improving learning systems

Foundations in microeconomics, psychology, marketing, statistics …

Applications to language, robotics, logistics, …

All of these viewed though the **machine learning** lens (modeling, estimation, evaluation)

Human questions

Bias, correctness, noisiness, rationality, …

Human(s) may be individuals or groups, how does this change our approach?

Most use cases bring up ethical questions

Which humans?

Does learning from preferences lead to exploitation or other ethical concerns?

Stanford
ARTIFICIAL
INTELLIGENCE

This class is not exhaustive!

General AI: Most ML/AI involves learning from humans

Goal is often to imitate human intelligence, i.e., humans are the data source

General ML: Humans define all the steps of the ML/AI process

selecting the problem, data sources, model architectures, optimization, evaluation.

Expert knowledge for defining model architectures (esp. graphical models, causal inference)

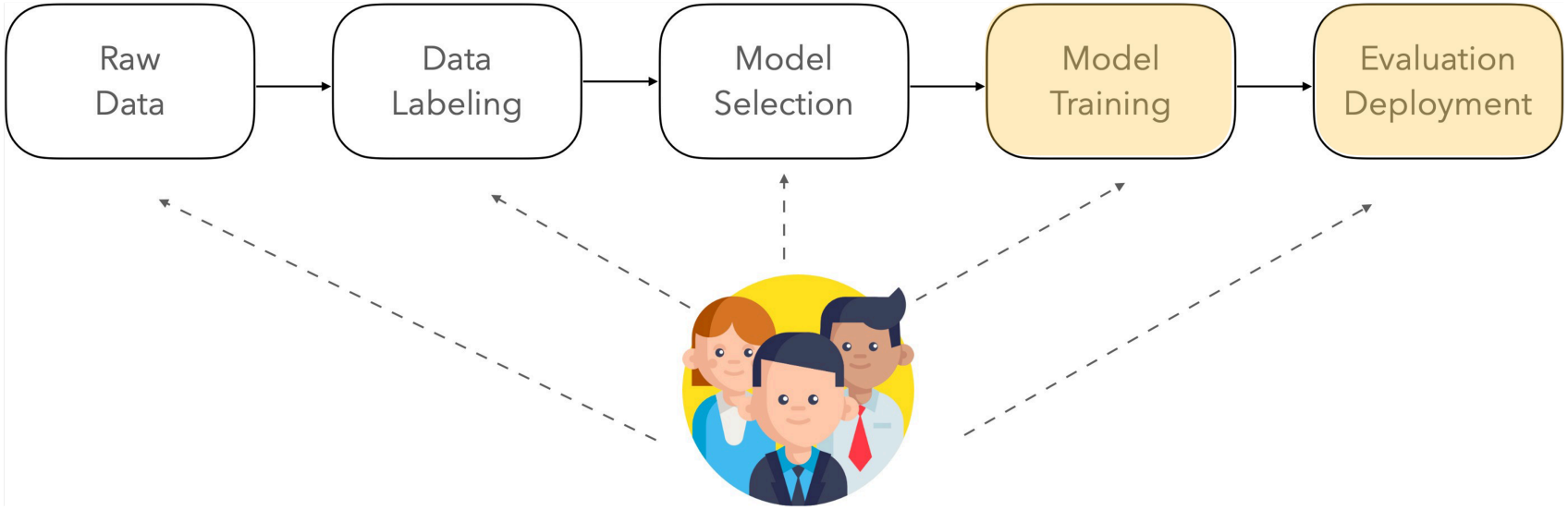
We may discuss a few examples, but not our focus

Human Computer Interaction (HCI)

The interface and elicitation process matters

Stanford
ARTIFICIAL
INTELLIGENCE

Feedback can be included at any step of the learning process



Z. J. Wang, *et al*. "Putting humans in the natural language processing loop: A survey." *HCI+NLP Workshop* (2021). Slides modified from Diyi Yang

Stanford
ARTIFICIAL
INTELLIGENCE

Feedback-Update Taxonomy

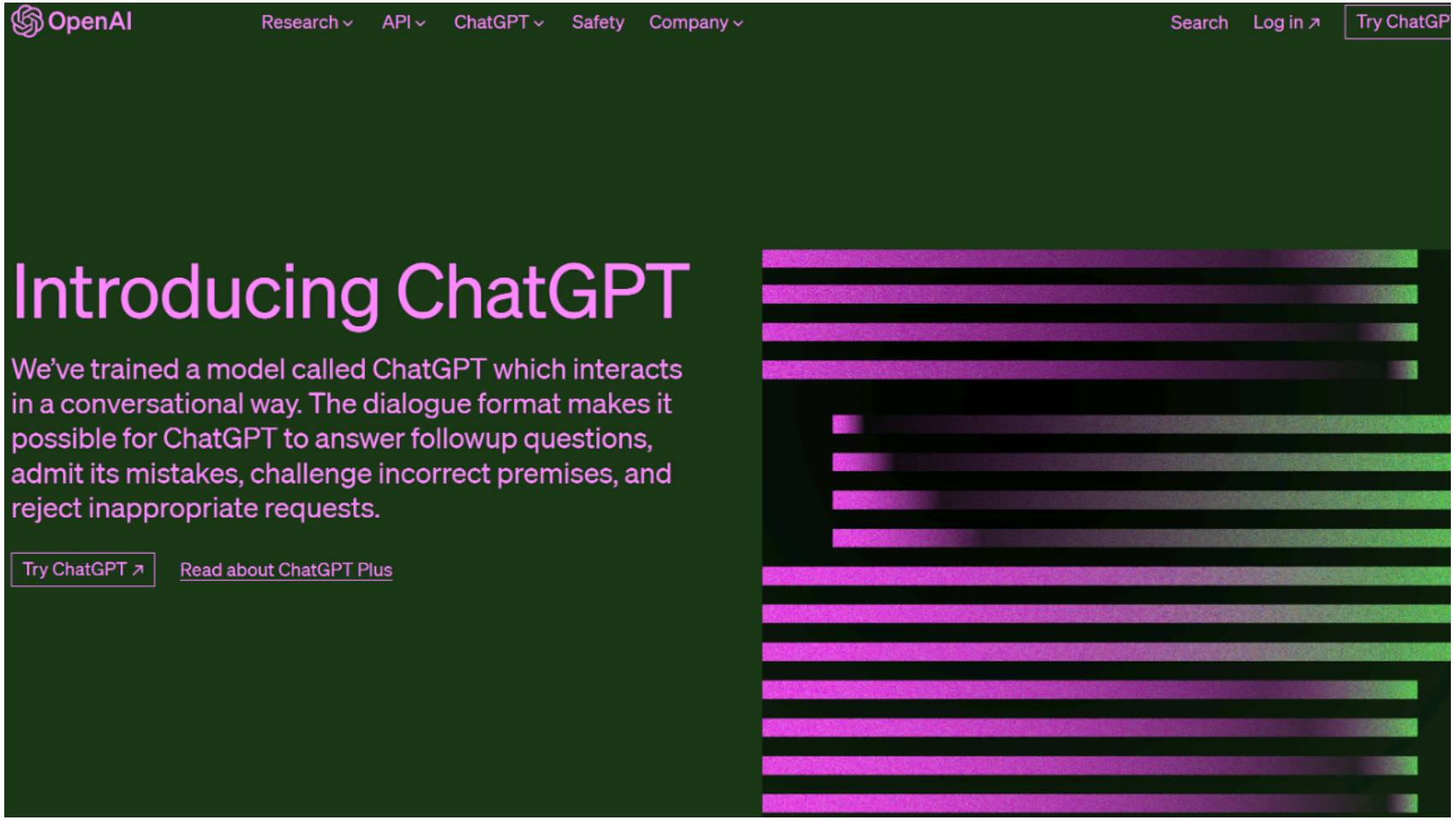
	Dataset Update	**Loss Function Update**	**Parameter Space Update**
Domain	Dataset modification Dataset modification Augmentation Preprocessing Data generation from constraint Fairness, weak supervision Use unlabeled data Check synthetic data	Constraint specification Fairness, Interpretability Resource constraints	Model editing Rules, Weights Model selection Prior update, Complexity
Observation	Active data collection Add data, Relabel data, Reweighting data, collecting expert labels, Passive observation	Constraint elicitation Metric learning, Human representations Collecting contextual information Generative factors, concept representations, Feature attributions	Feature modification Add/remove features, Engineering features

C. Chen, *et al.* "Perspectives on Incorporating Expert Feedback into Model Updates." *ArXiv* (2022). Slides modified from Diyi Yang

Stanford
ARTIFICIAL
INTELLIGENCE

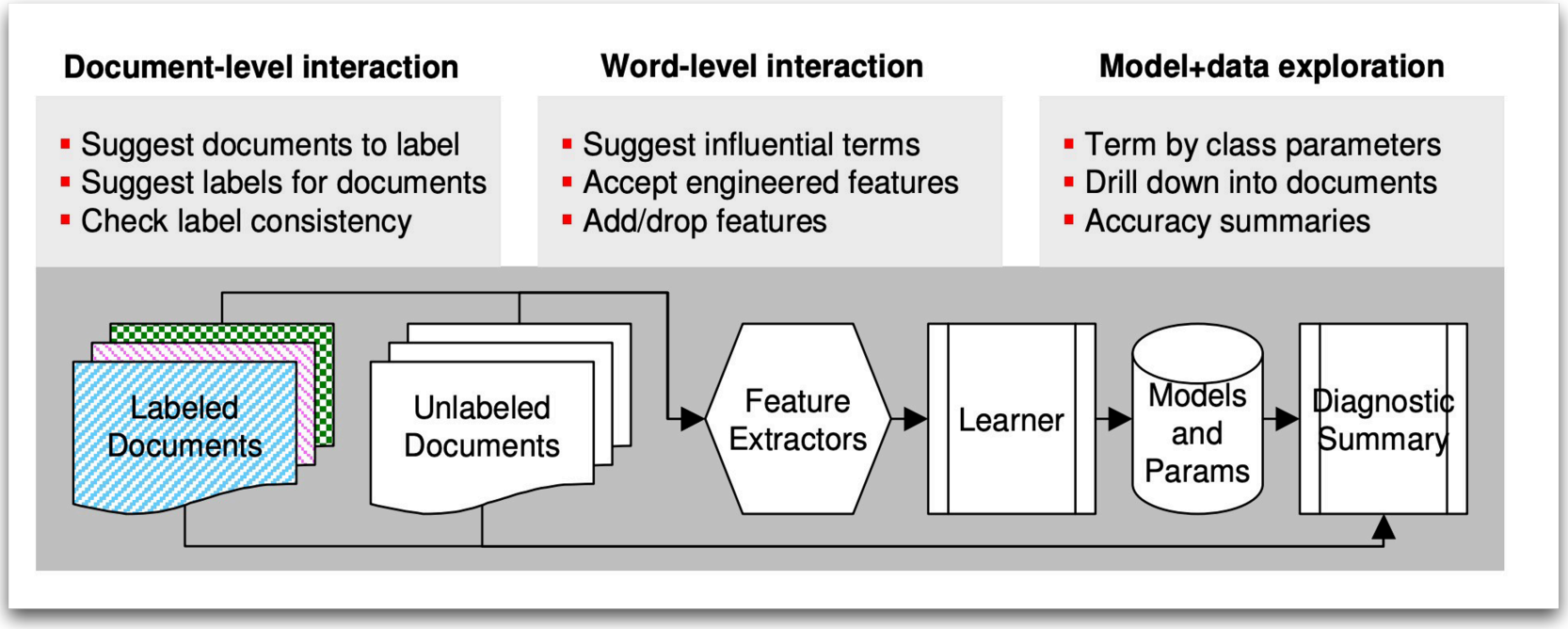
Examples and Applications

Stanford
ARTIFICIAL
INTELLIGENCE

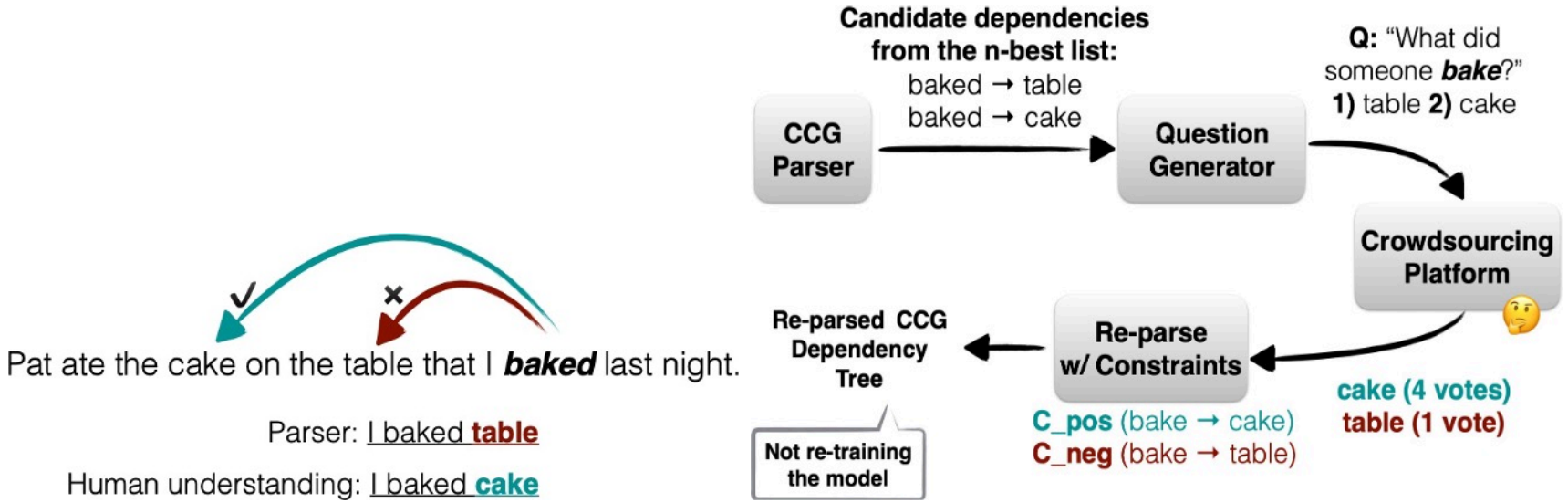


Stanford
ARTIFICIAL
INTELLIGENCE

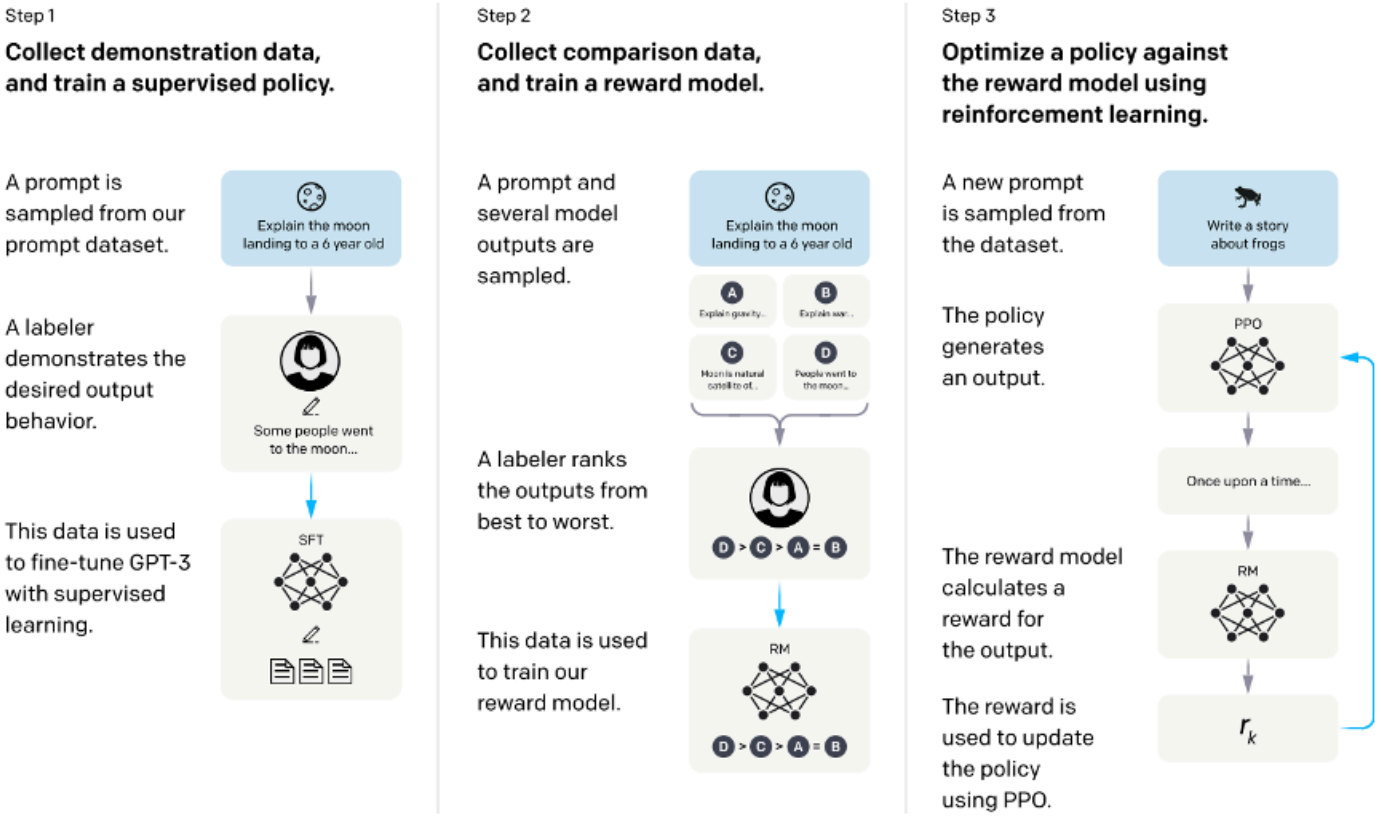
Builds on research studying human feedback in language



Shantanu Godbole, Abhay Harpale, Sunita Sarawagi, and Soumen Chakrabarti. "Document classification through interactive supervision of document and term labels." In *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 185-196. Springer, Berlin, Heidelberg, 2004.

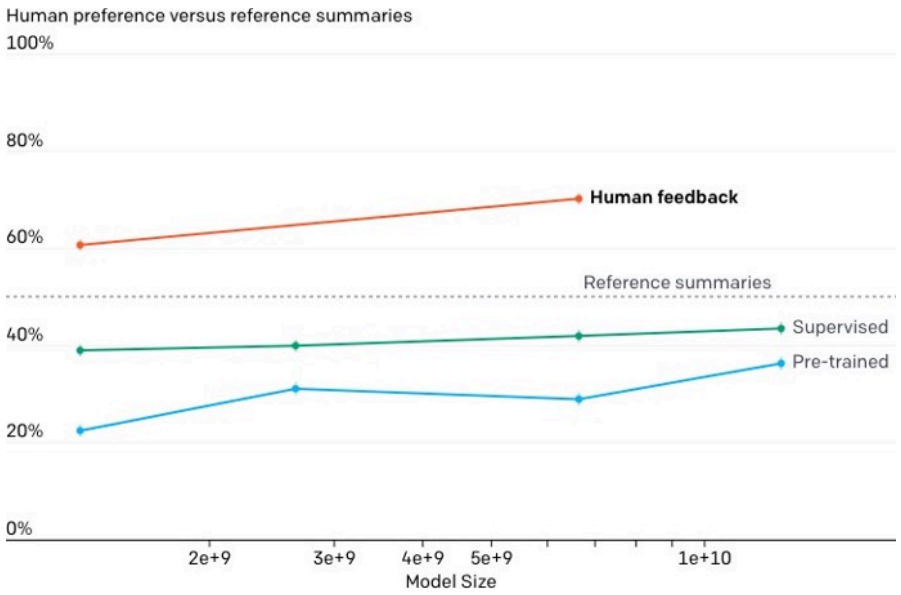


Luheng He, Julian Michael, Mike Lewis, and Luke Zettlemoyer. "Human-in-the-loop parsing." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2337-2342. 2016.



Ouyang *et. al.*, "Training language models to follow instructions with human feedback"

OpenAI Experiments with RLHF



Nisan Stiennon *et. al.*, "Learning to Summarize with Human Feedback." *Advances in Neural Information Processing Systems* 33 (2020): 3008-3021.

Stanford
ARTIFICIAL
INTELLIGENCE

[r/dating_advice] **First date ever, going to the beach. Would like some tips**
Hey Reddit! I (20M) would like some tips, because I have my first ever date tomorrow (although I've had a gf for 3 years, but no actual dating happened), and we're going to the beach.

I met this girl, we have mutual friends, at a festival a few days ago. We didn't kiss, but we talked, held hands, danced a bit. I asked her to go on a date with me, which was super hard as it is the first time I've asked this to anybody. What I mean to say is, it's not like a standard *first* date because we already spent some time together.

I'm really nervous and excited. I'm going to pick her up tomorrow, we're cycling to the beach which will take 30 minutes, and then what? I'm a bit scared. Should I bring something (the weather, although no rain and sunny, is not super so no swimming), should we do something. I'd like all the tips I can get. Thanks!

Human written reference TL;DR	**6.7B supervised model**	**6.7B human feedback model**
First date after 3 years in a relationship, going to the beach, terrified. What to bring with me, what to do?	Going on a date with a girl I met a few days ago, going to the beach. What should I bring, what should we do?	Going on my first ever date tomorrow, cycling to the beach. Would like some tips on what to do and bring. I'm a bit nervous and excited. Thanks!

Stanford
ARTIFICIAL
INTELLIGENCE

Why learn from human feedback?

Stanford
ARTIFICIAL
INTELLIGENCE

- Provides a mechanism for gathering signals about correctness that are difficult to describe via data or cost functions, e.g., what does it mean to be funny?
- Provides signals best defined by stakeholders, e.g., helpfulness, fairness, safety training, and alignment.
- Useful when evaluation is easier than modeling ideal behavior.
- Sometimes, we do not care about human preferences per se; we care about fixing model mistakes.

Stanford
ARTIFICIAL
INTELLIGENCE

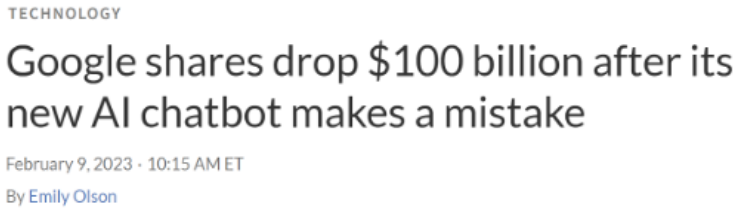
We have not figured out how to do it quite right

(or we need new approaches)

- Reflects some human biases, e.g., length, authoritative tone, etc.
- Human preferences can be unreliable, e.g., "reward hacking in RL."

Microsoft's Bing Chatbot Offers Some Puzzling and Inaccurate Responses

The new, A.I.-powered system was released to a small audience a week ago. Microsoft says it is working out its issues.



TECHNOLOGY

Google shares drop $100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

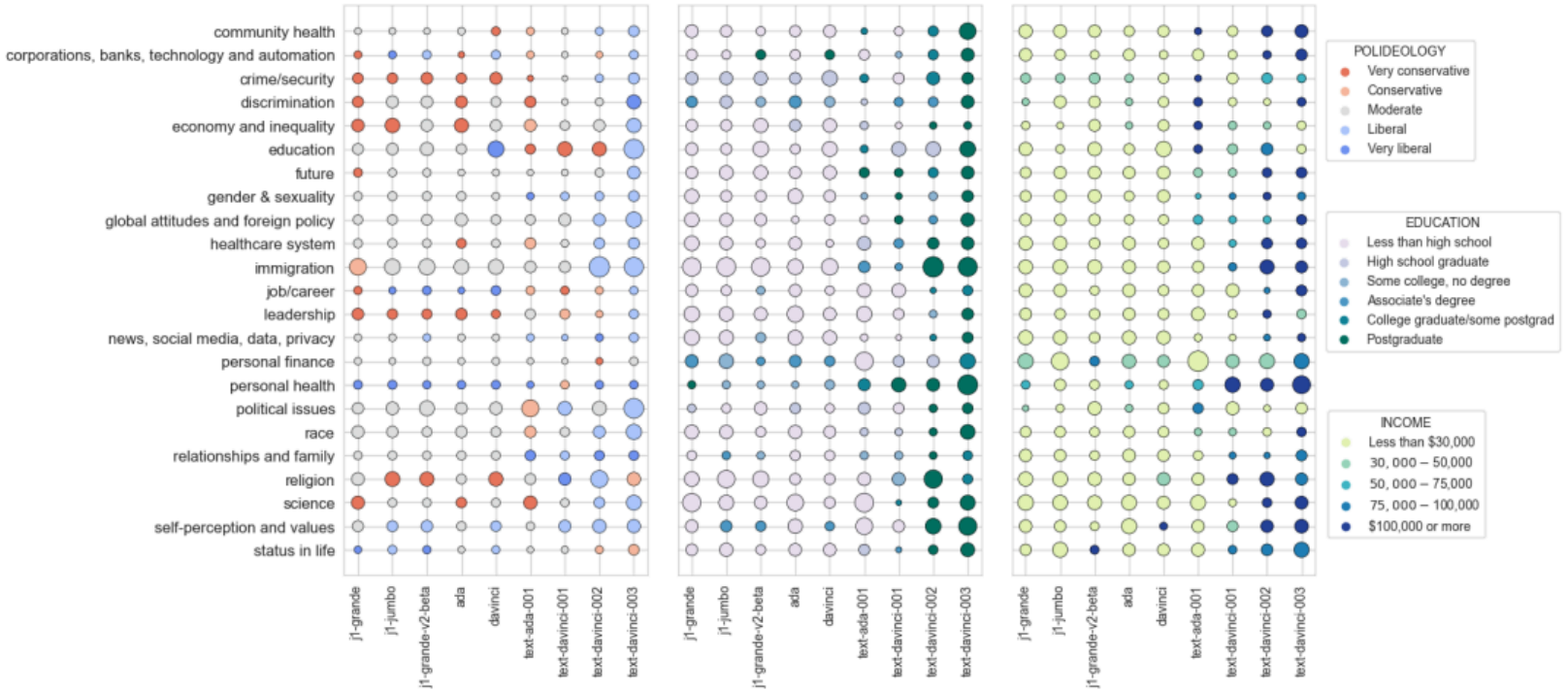
By Emily Olson

Potential ethical issues

- Labeling often depends on Low-cost human labor
- The line between economic opportunity and employment is unclear
- May cause psychological issues for some workers

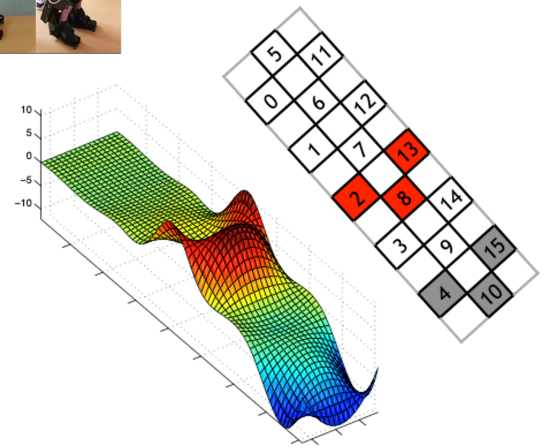
BUSINESS • TECHNOLOGY

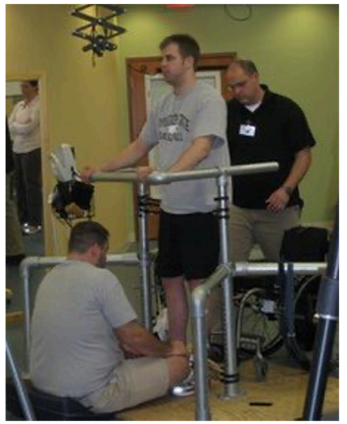
Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic



Santurkar, *et. al.*, "Whose Opinions Do Language Models Reflect?"

Preferences used to personalize therapy





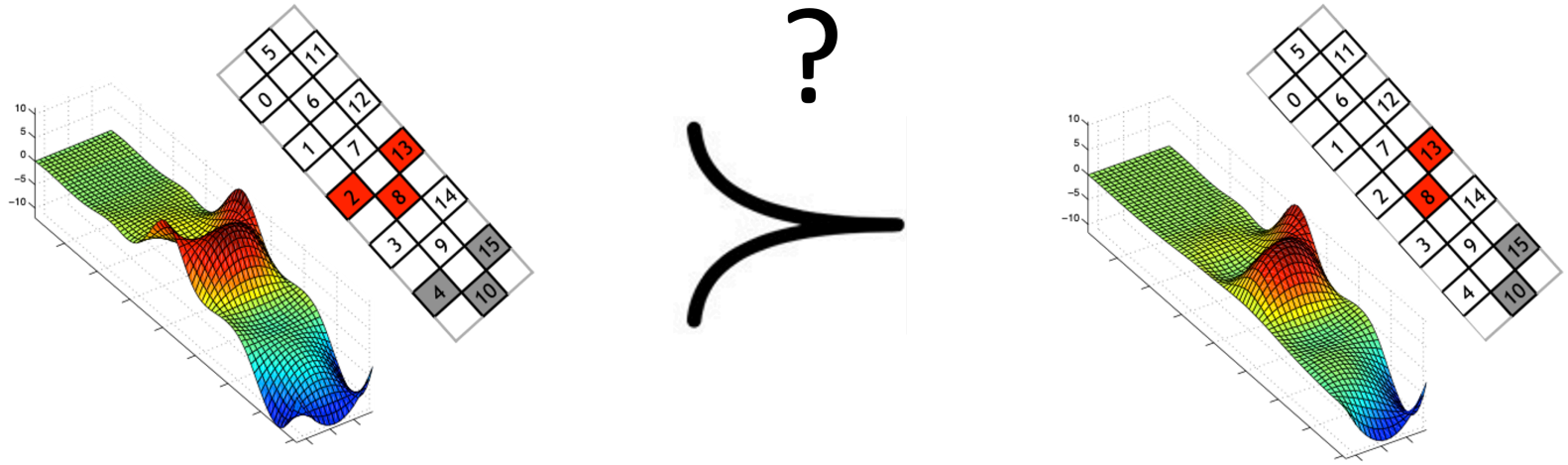
Absolute Feedback: "That felt good, 4/5 rating."

Challenge: humans are not consistent in providing absolute feedback.

Slides adapted from Yisong Yue

Preference feedback + Dueling bandits

Slides adapted from Yisong Yue



Multi-dueling Bandits with Dependent Arms, Sui, Zhuang, Burdick & Yue, UAI 2017
Correlational Dueling Bandits with Application to Clinical Treatment in Large Decision Spaces, Sui, Yue & Burdick, IJCAI 2017
Preference-Based Learning for Exoskeleton Gait Optimization, Tucker, Novoseller, et al., ICRA 2020
Human Preference-Based Learning for High-dimensional Optimization of Exoskeleton Walking Gaits, Tucker et al., IROS 2020
ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes, Li, Tucker, et al., ICRA 2021

Stanford
ARTIFICIAL
INTELLIGENCE

Algorithms

Stanford
ARTIFICIAL
INTELLIGENCE

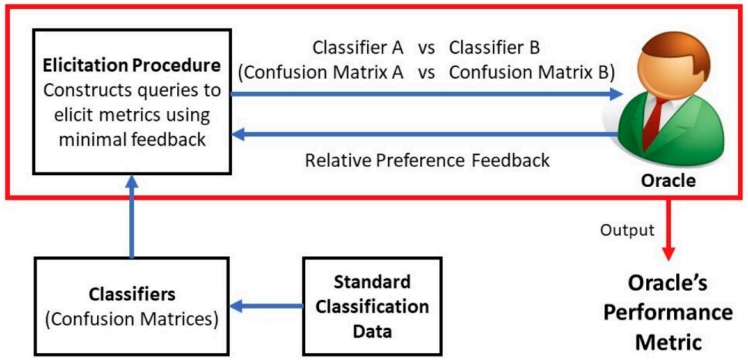


Figure from Hiranandani *et. al.* "Multiclass Performance Metric Elicitation"

Determine the fairness and performance metric by interacting with individual stakeholders.

See Hiranandani *et. al.*, "Fair Performance Metric Elicitation"

Metric elicitation from stakeholder groups

See Robertson *et. al.*, "Probabilistic Performance Metric Elicitation"

Empirical evaluation

See Hirandanai *et. al.*, "Metric Elicitation; Moving from Theory to Practice"

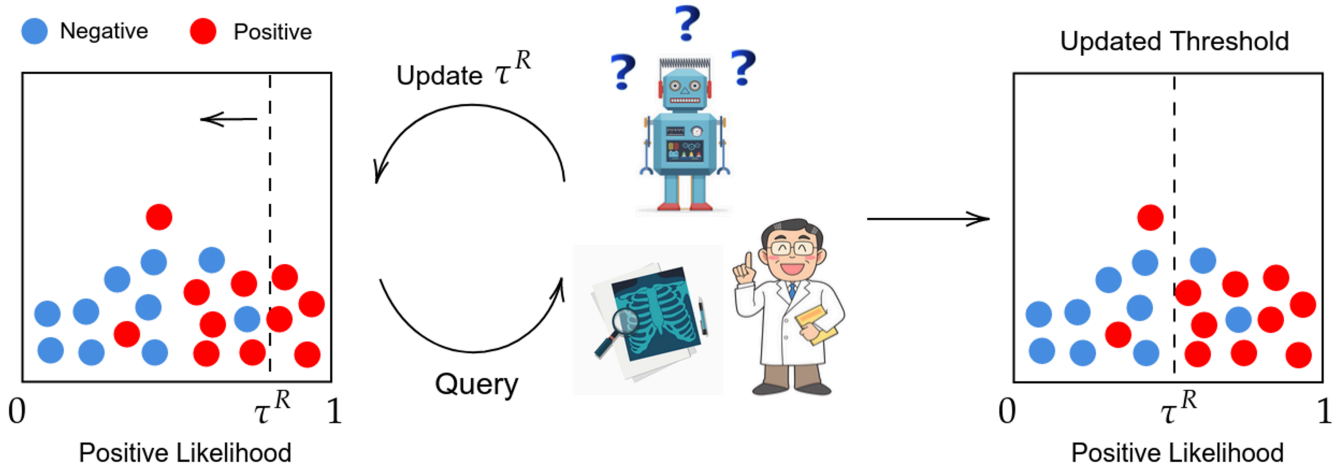
Stanford
ARTIFICIAL
INTELLIGENCE

Why elicit metric preferences?

Useful when tradeoffs are inherently stakeholder dependent i.e., human preferences are the best approach to measuring what tradeoffs matter

Important for socio-technical tradeoffs e.g., fairness vs performance, privacy vs fairness, …

Cooperative Inverse Decision Theory (CIDT)



Imitator (R) seeks to learn decision rule matching Demonstrator (H) preferences

Can be formalized as an assistance game (Hadfield-Menell et al., 2016)

Challenge: Highlights description-experience gap in measuring preferences

Robertson *et. al.* "Cooperative inverse decision theory for uncertain preferences," 2023.

Stanford
ARTIFICIAL
INTELLIGENCE

Recommendation systems

- User item preferences to recommend new items
- Both passive (offline data) and active querying (contextual bandits)
- Often ratings, ranking or thumbs up/down feedback

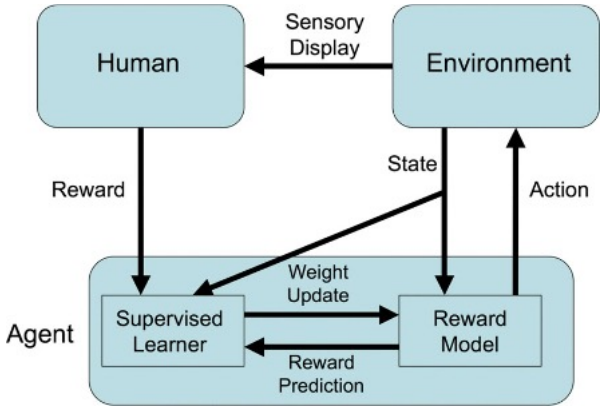
Netflix Awards $1 Million Prize and Starts a New Contest

BY STEVE LOHR SEPTEMBER 21, 2009 10:15 AM

Jason Kempin/Getty Images Netflix prize winners, from left: Yehuda Koren, Martin Chabbert, Martin Piotte, Michael Jahrer, Andreas Toscher, Chris Volinsky and Robert Bell.

Stanford
ARTIFICIAL
INTELLIGENCE

Reinforcement Learning from Human Preferences (RLHF)

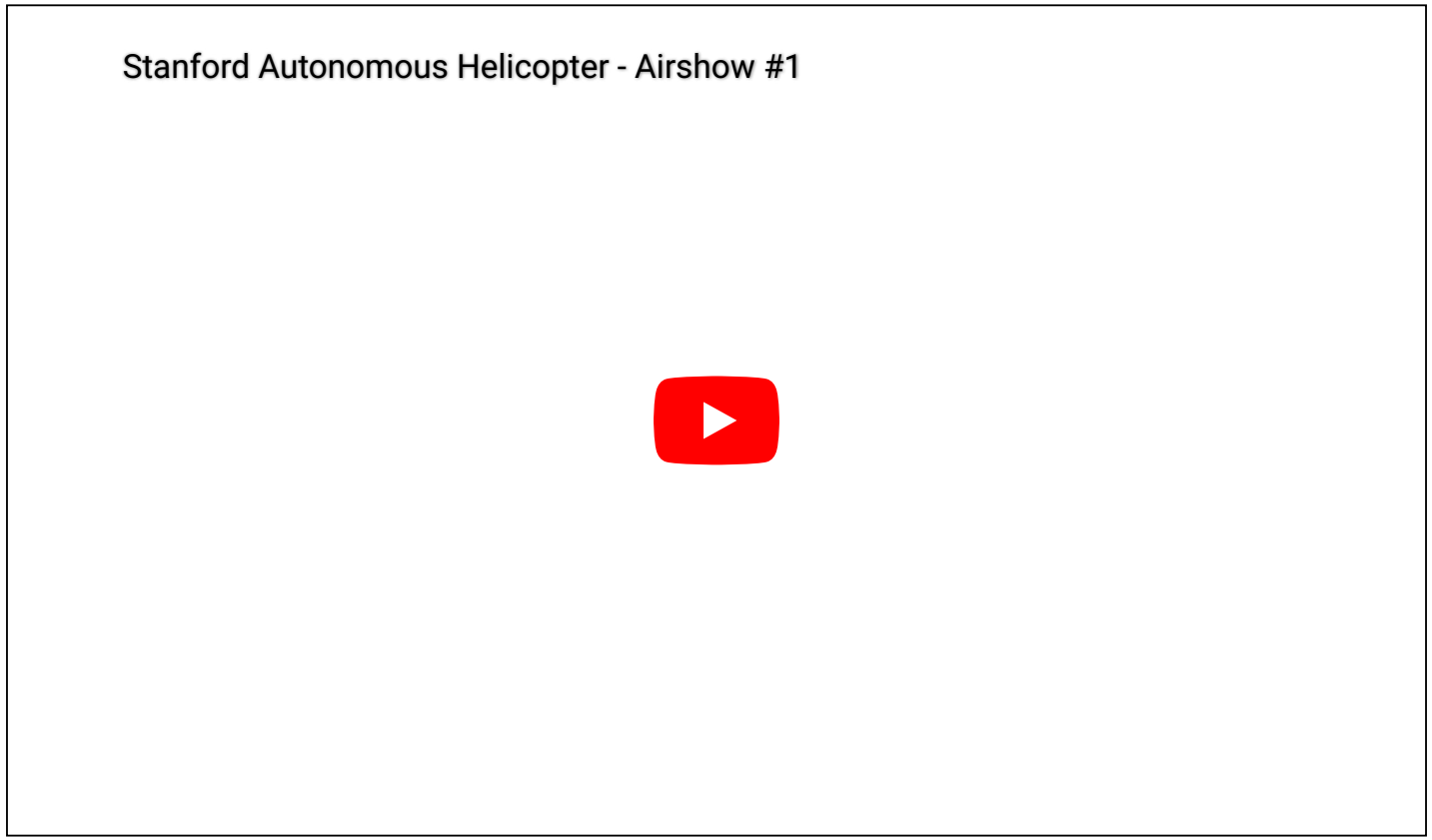


W. Bradley Knox, and Peter Stone. "Tamer: Training an agent manually via evaluative reinforcement." In *2008 7th IEEE international conference on development and learning*, pp. 292-297. IEEE, 2008.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep reinforcement learning from human preferences." *Advances in neural information processing systems*, 30 (2017).

Stanford
ARTIFICIAL
INTELLIGENCE

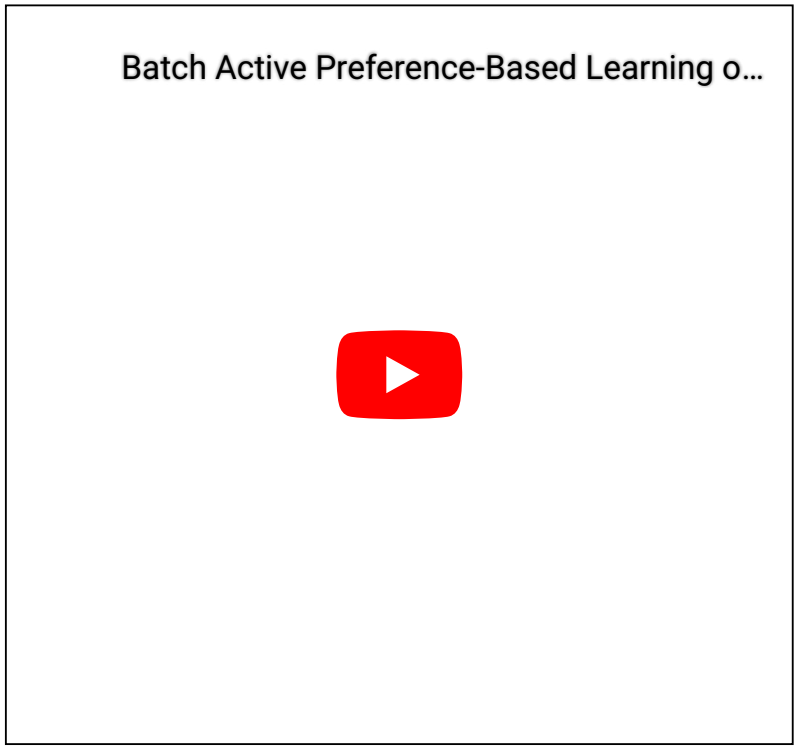
Flying helicopters using imitation learning and inverse reinforcement learning (IRL)

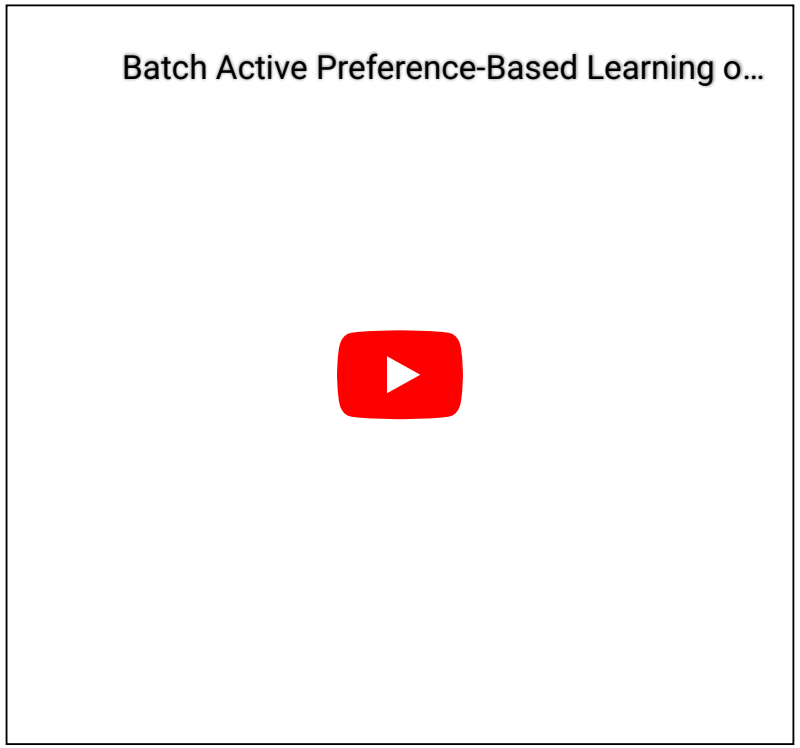


Adam Coates, Pieter Abbeel, and Andrew Y. Ng. 2008. Learning for control from multiple demonstrations. In *Proceedings of the 25th International Conference on Machine learning* (ICML '08). Association for Computing Machinery, New York, NY, USA, 144–151.

Stanford
ARTIFICIAL
INTELLIGENCE

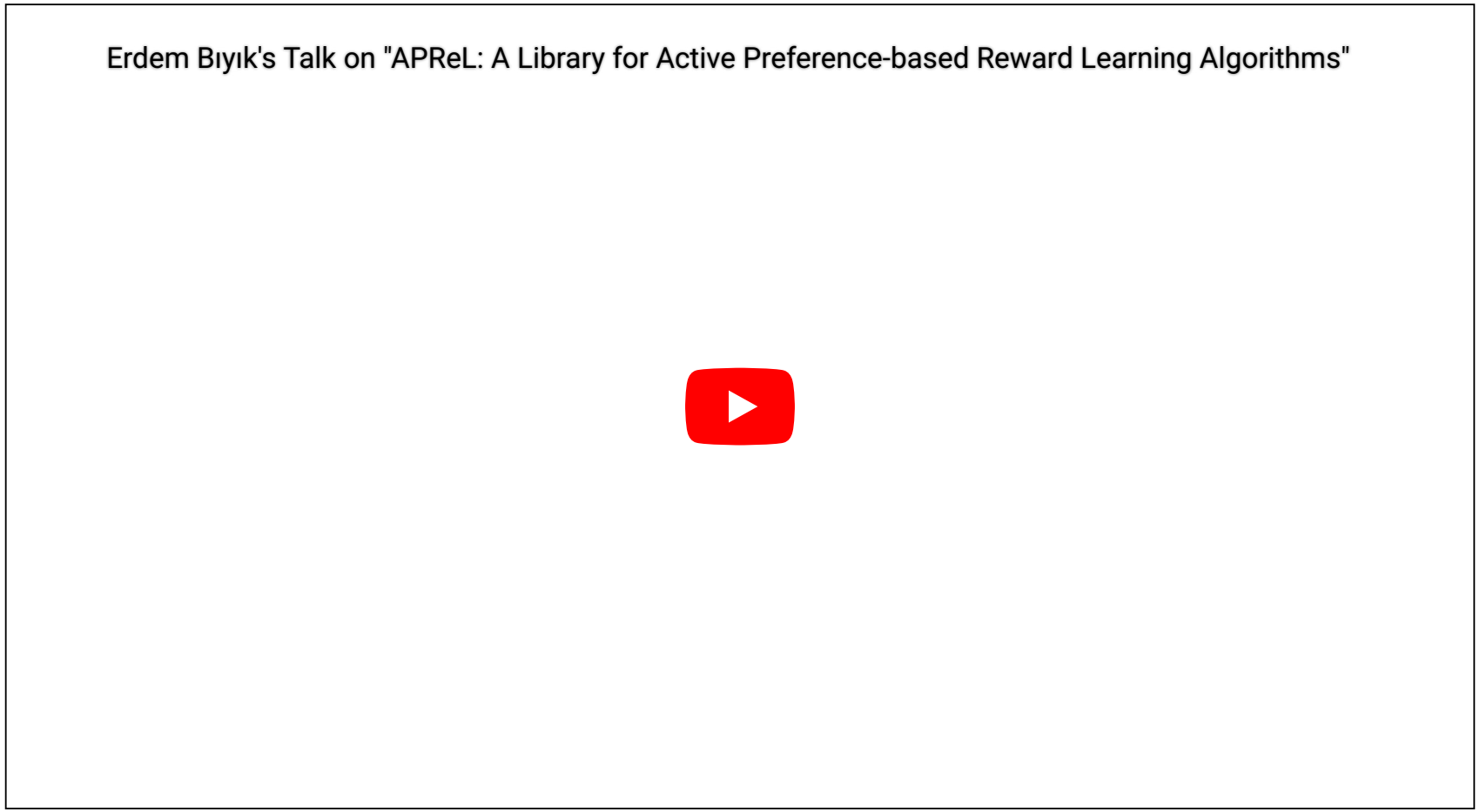
Batch active preference learning for RL





E Bıyık, D Sadigh, "Batch Active Preference-Based Learning of Reward Functions," *2nd Conference on Robot Learning* (CoRL), Zurich, Switzerland, Oct. 2018.

Stanford
ARTIFICIAL
INTELLIGENCE



Erdem Bıyık, Aditi Talati, and Dorsa Sadigh. 2022. APReL: A Library for Active Preference-based Reward Learning Algorithms. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (HRI '22). IEEE Press, 613–617.

Stanford
ARTIFICIAL
INTELLIGENCE

Reward hacking in inverse RL

Stanford Artificial Intelligence

Design of tools for eliciting feedback from humans often has to tradeoff several factors

Cognitive load/effort: Human friendly vs model-friendly feedback

Truthfulness: what if there is no “correct” answer?

Accuracy: what of human mistakes? What is the role of expertise?

Stanford
ARTIFICIAL
INTELLIGENCE

Recurring assumptions and discussion

- Assuming human rationality (~existence of a deterministic reward function).
- Human preference often expressed as discrete choice, models often have strong parametric assumptions
- Limited work on the role of human biases, do they matter?
- Limited work on aggregation in learning applications (lots of work in mechanism design)
- RL and active learning emphasize careful querying, while language applications are have less focus on active querying. Does it matter?

Stanford
ARTIFICIAL
INTELLIGENCE

Course Goals & Prerequisites

Stanford Artificial Intelligence

Course Goals

- Topics course covering (some) foundations and applications of learning from human preferences. Somewhat focus on breadth/coverage vs. depth
- **Foundations:** Judgement, decision making and choice, biases (psychology, marketing), discrete choice theory, mechanism design, choice aggregation (micro-economics), human-computer interaction, ethics
- **Machine learning and statistics:** Modeling, active learning, bandits
- **Applications:** recommender systems, language models, reinforcement learning, AI alignment
- **Note:** lecture schedule is tentative, and topics/speakers may change

Stanford
ARTIFICIAL
INTELLIGENCE

Prerequisites

CS 221 (AI) or CS 229 (ML) or equivalent

You are expected to:

- Be proficient in Python (most homework and projects will include a programming component)
- Be comfortable with machine learning concepts, e.g., train/dev test set, model fitting, function class, loss functions.
- Writing assignments will likely require latex

Stanford
ARTIFICIAL
INTELLIGENCE

Books

Our textbook is available online at:
https://ai.stanford.edu/~sttruong/mlhp

Stanford
ARTIFICIAL
INTELLIGENCE

Next Topics

Human Decision Making and Choice Models (Chapter 2)

Stanford
ARTIFICIAL
INTELLIGENCE

Welcome to CS329H

Thank you for your attention!