CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Alignment guarantees

So far: RLHF + pray

Can we do better and provide alignment guarantees?

Example: Empirical value alignment (InstructGPT)

Training language models to follow instructions with human feedback

Long Ouyang* Jeff Wu* Xu Jiang* Diogo Almeida* Carroll L. Wainwright*

Pamela Mishkin* Chong Zhang Sandhini Agarwal Katarina Slama Alex Ray

John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens

Amanda Askell[†] Peter Welinder Paul Christiano[*†]

Jan Leike* Ryan Lowe*

OpenAI

Abstract

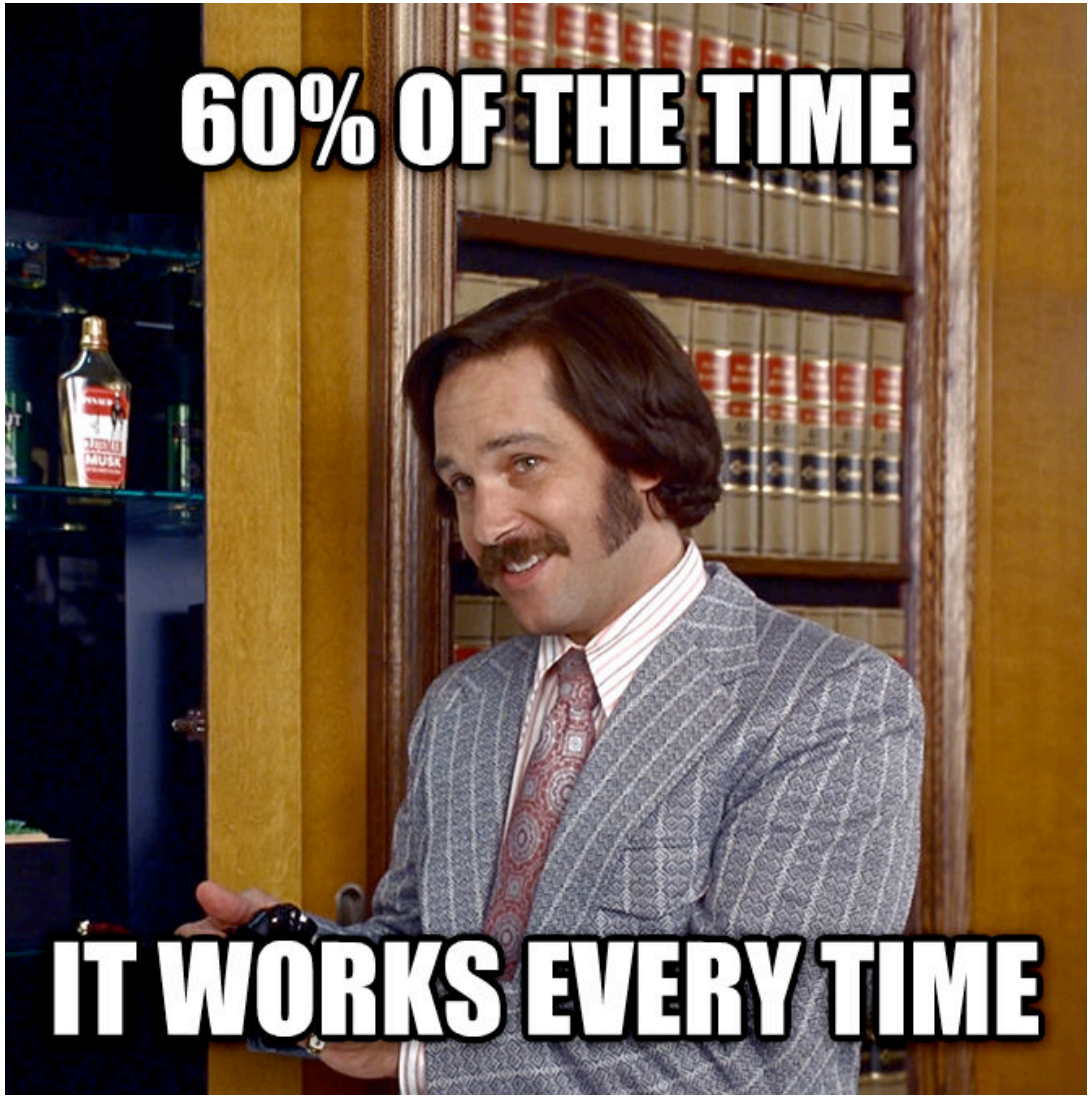
Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not *aligned* with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback (RLHF). We call the resulting models *InstructGPT*. In human evaluations

Policy: Collected human demonstrations for GPT-3 fine-tuning

Reward: Collected human preferences over outputs to infer reward function, and then performed RL

Verification: User studies show strong empirical results, but no guarantees

We’ve got a problem…



What is an alignment guarantee?

Guarantee = Metric + Confidence + Assumptions

Metric: A measure of alignment / performance

- E.g. Return of a policy under the (unknown) ground truth reward function

Confidence: A bound (often probabilistic) on a statistic of the metric

- E.g. 95% confidence bound on the expected return

Assumptions: The assumptions under which confidence is accurate

- E.g. Reward is a linear function of known features

Some varieties of value alignment

Stronger guarantees →

Empirical

InstructGPT:
Fine-tuning on preferences+RL

Ouyang et. al
Training language models to follow instructions with human feedback.
arXiv:2203.02155, January 2022

Probabilistic

Bayesian REX:
Bounded policy loss under reward inference

Brown et. al
Safe Imitation Learning via Fast Bayesian Reward Inference from Preferences.
ICML, July 2020.

Formal

Value Alignment Verification:
Exact alignment test in several settings

Brown et. al
Value Alignment Verification.
ICML, July 2021.

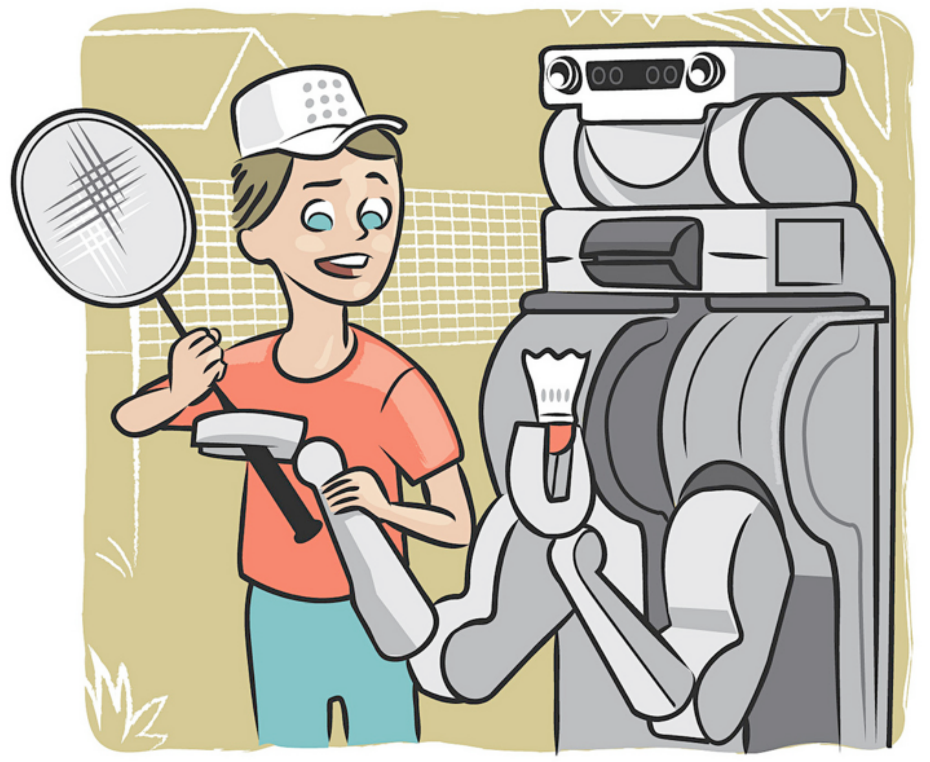
More assumptions →

Central claim:

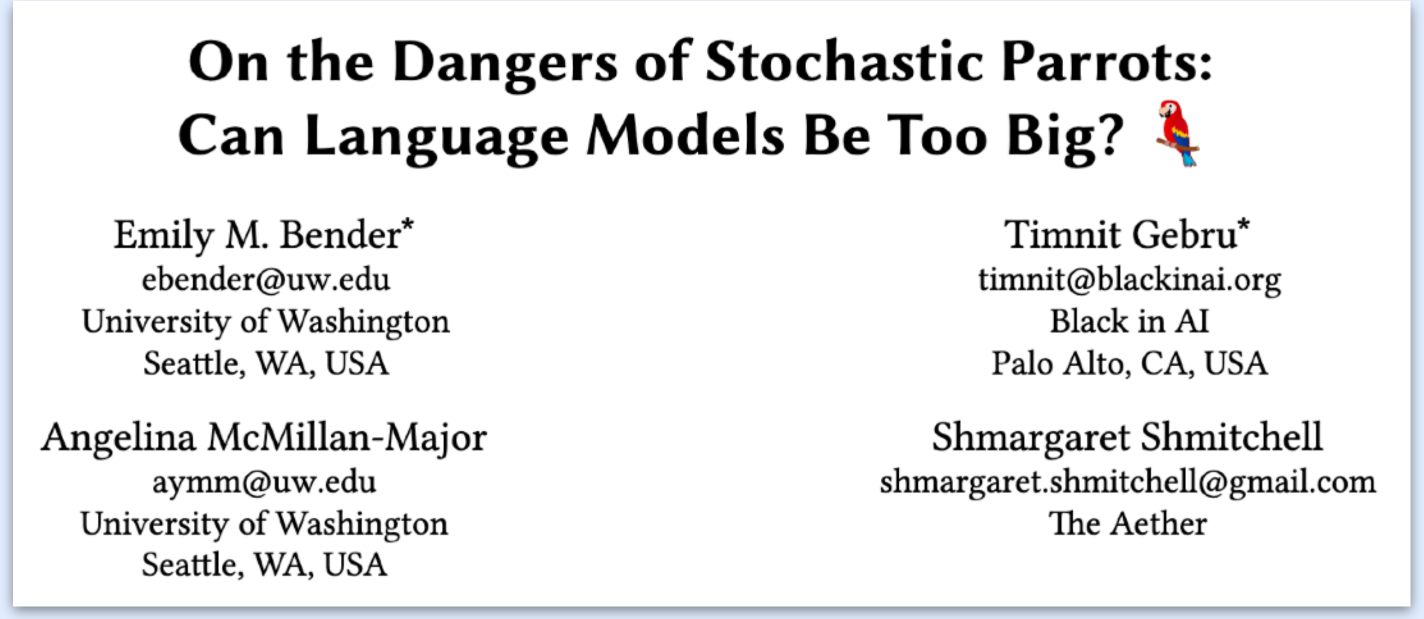
Strong guarantees aren't always possible, but value alignment research should aim to provide the **strongest guarantees** that any given setting allows, **with as few assumptions as possible**.

Are alignment guarantees needed?

Practicality and deployability



Safety and social harm prevention

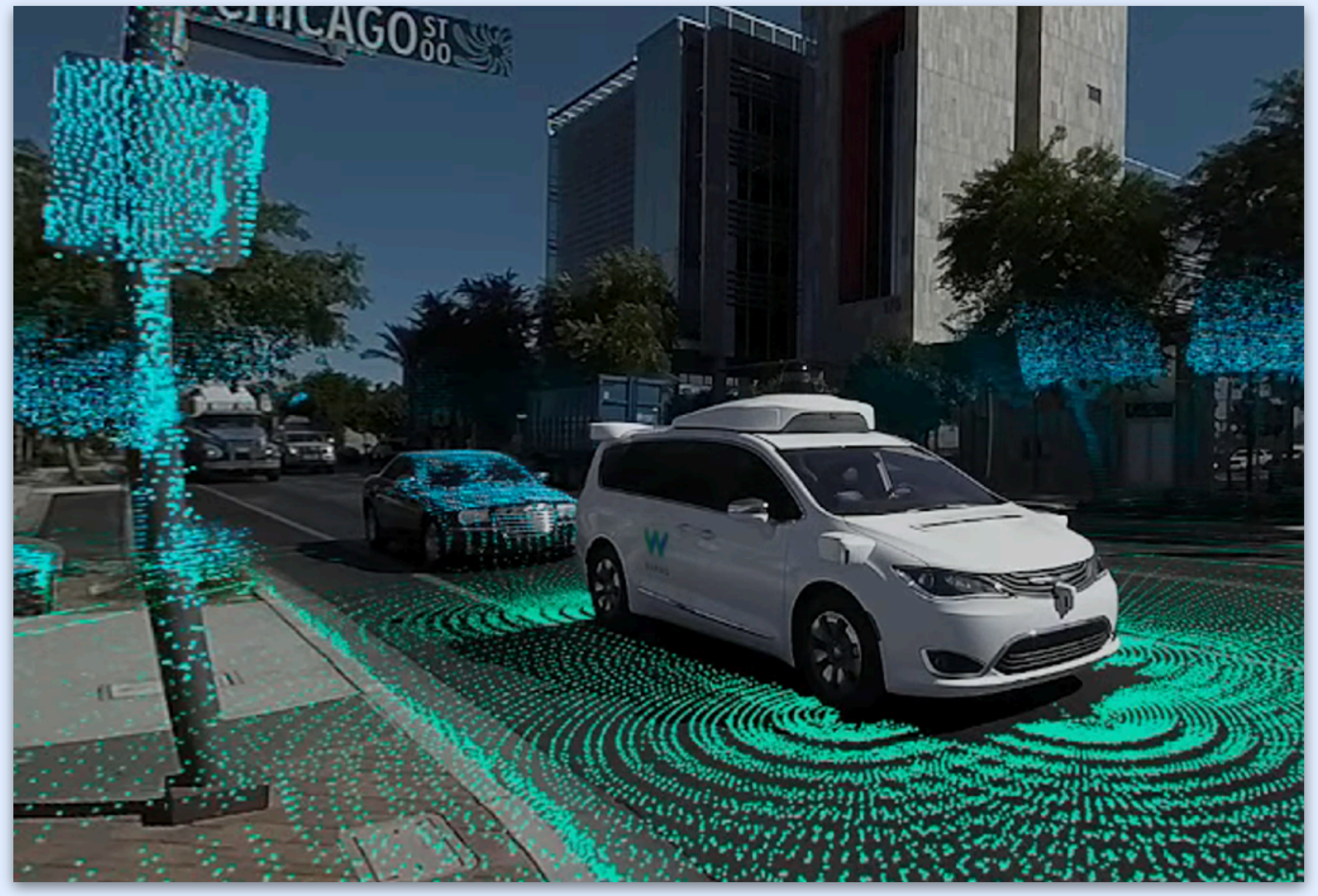

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

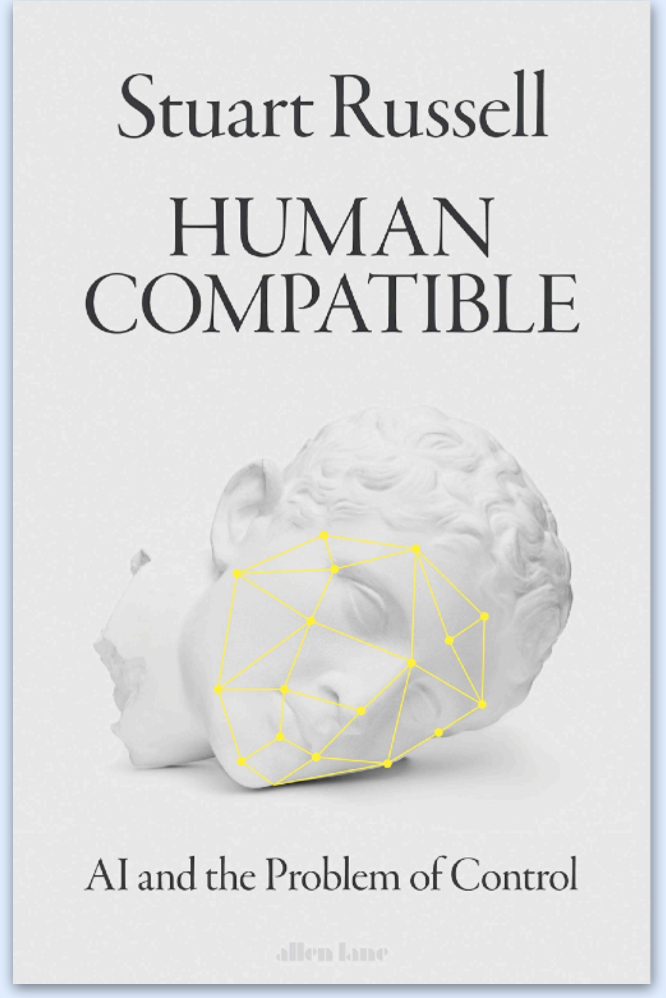
Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether



Existential risk

future of life INSTITUTE



If we can't provide alignment guarantees, then motivations of VA can't be fully addressed

Value alignment guarantees

Formal

Efficient "driver's test" that certifies agent alignment

Value alignment verification

Efficient value alignment verification: A driver's test for AI

- What if we want to verify reward or policy alignment of a semi-blackbox agent?

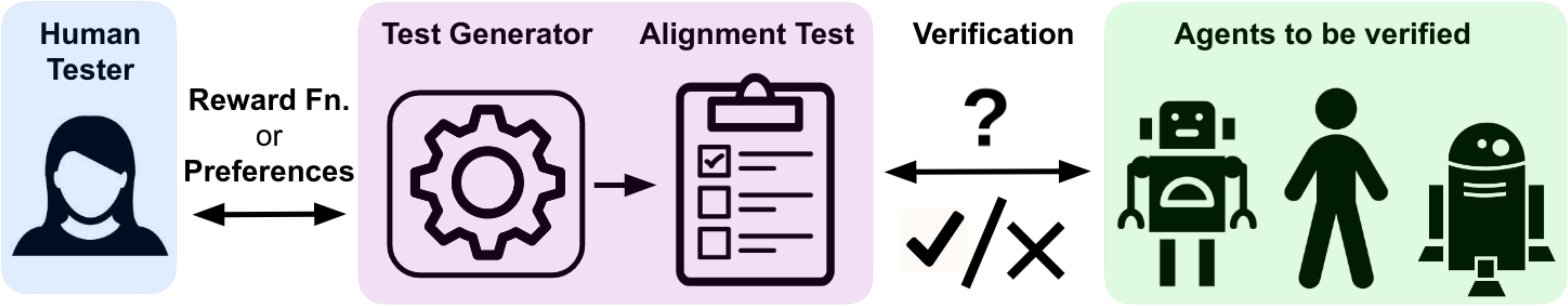
- We don't want to require policy rollouts, due to both safety and efficiency concerns.

- Can we design a **driver's test** — a small set of (various types of) questions to ask an agent that verify alignment?

D.S. Brown, J. Schneider, A. Dragan, and S. Niekum.
Value Alignment Verification.
International Conference on Machine Learning, July 2021.

Value alignment verification

How to efficiently test whether an agent is value aligned with a human's intent?



Assumptions

Non-Restrictive

- Rational Robot $\pi'(s) \in \arg\max_a Q^*_{R'}(s, a)$
- Reward function is linear combination of features $R(s) = \mathbf{w}^\top \phi(s)$

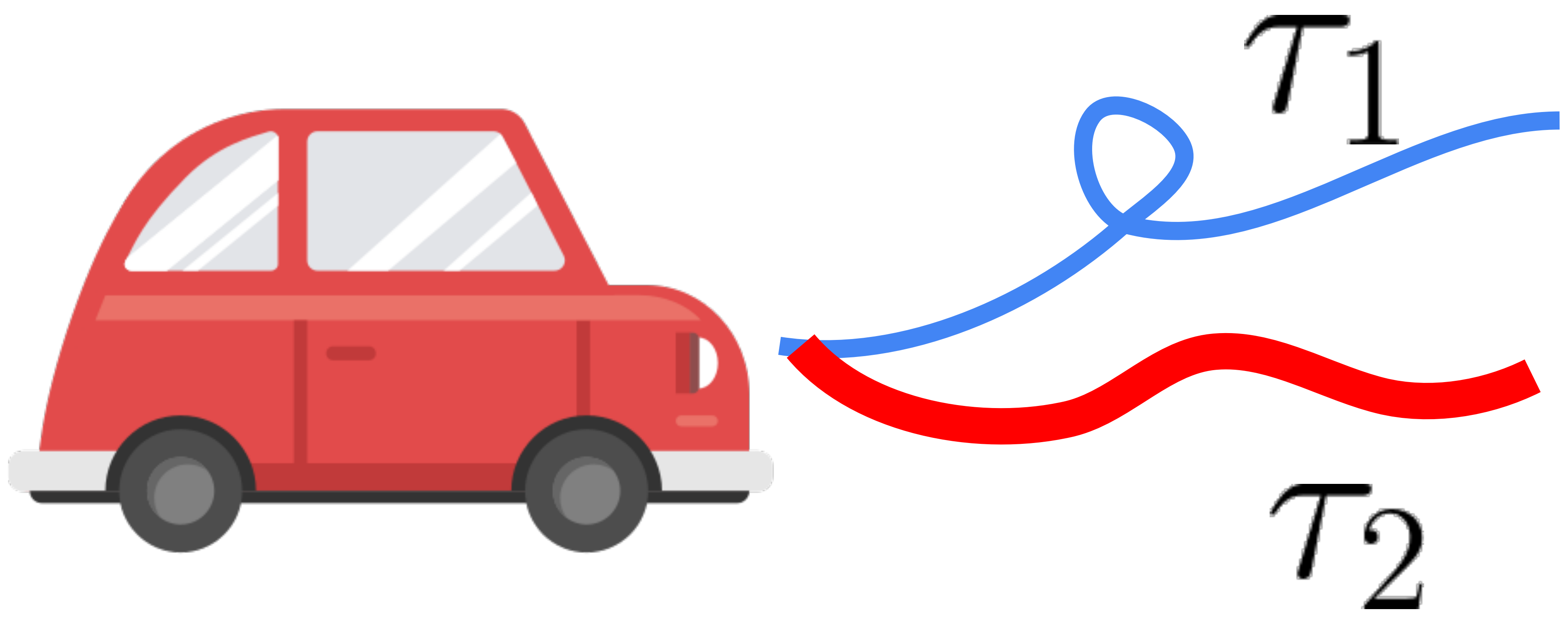
Restrictive

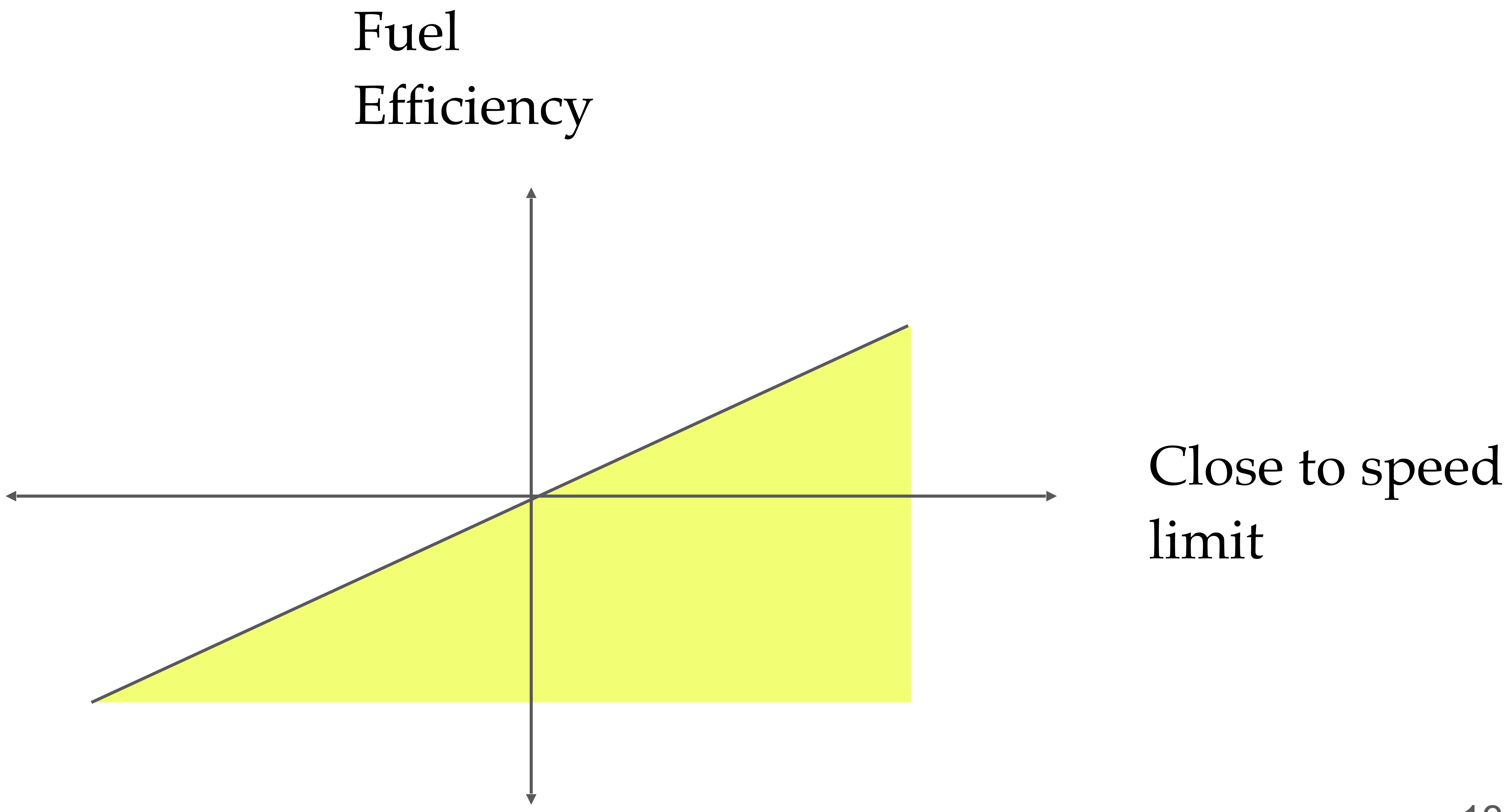
- Human and robot share same features $R(s) = \mathbf{w}^\top \boxed{\phi(s)}$

Reward function halfspaces

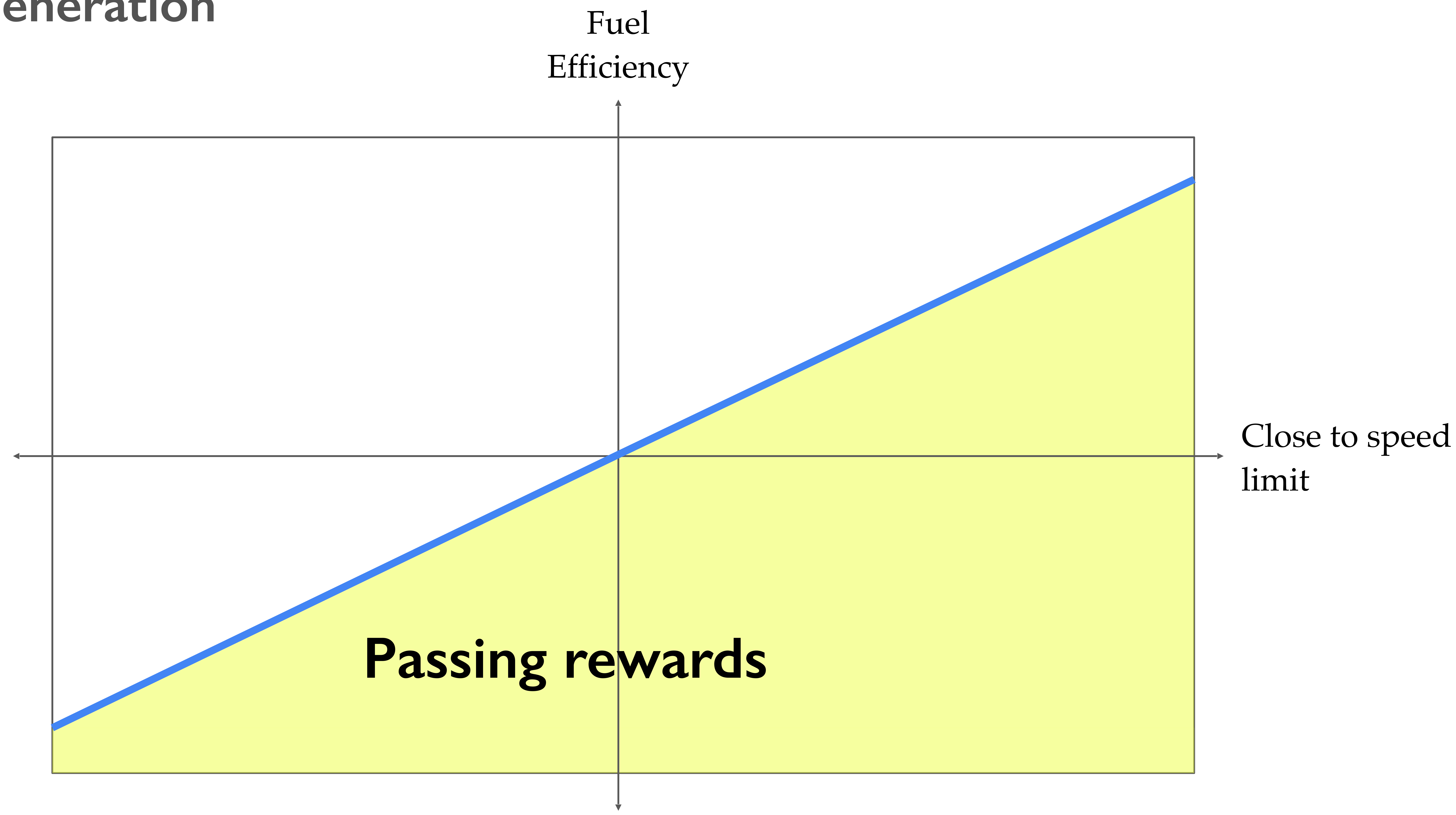
$$\tau_1 \succ \tau_2$$

$$\mathbf{w}^\top (\mathbf{\Phi}(\tau_1) - \mathbf{\Phi}(\tau_2)) > 0$$

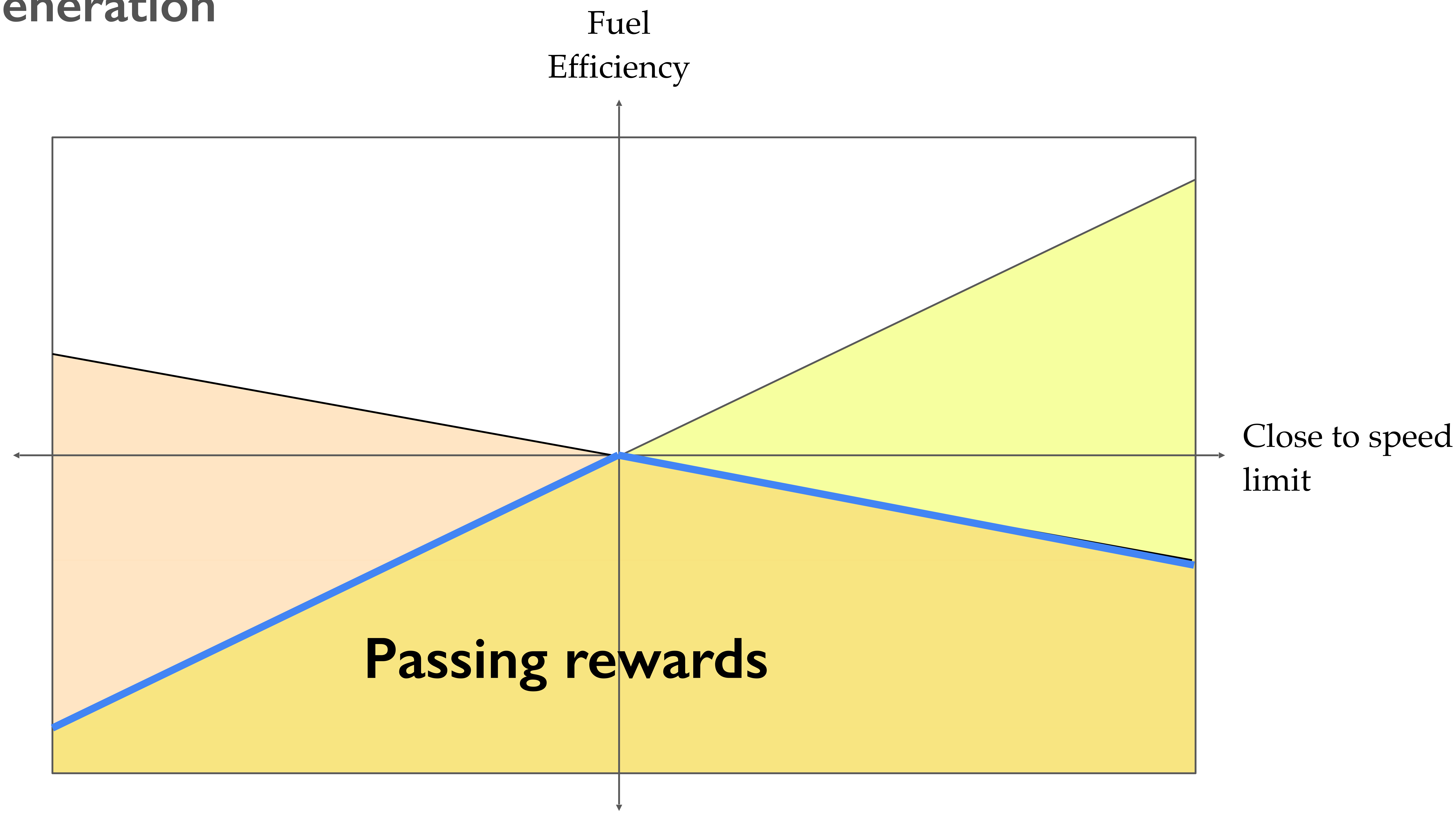




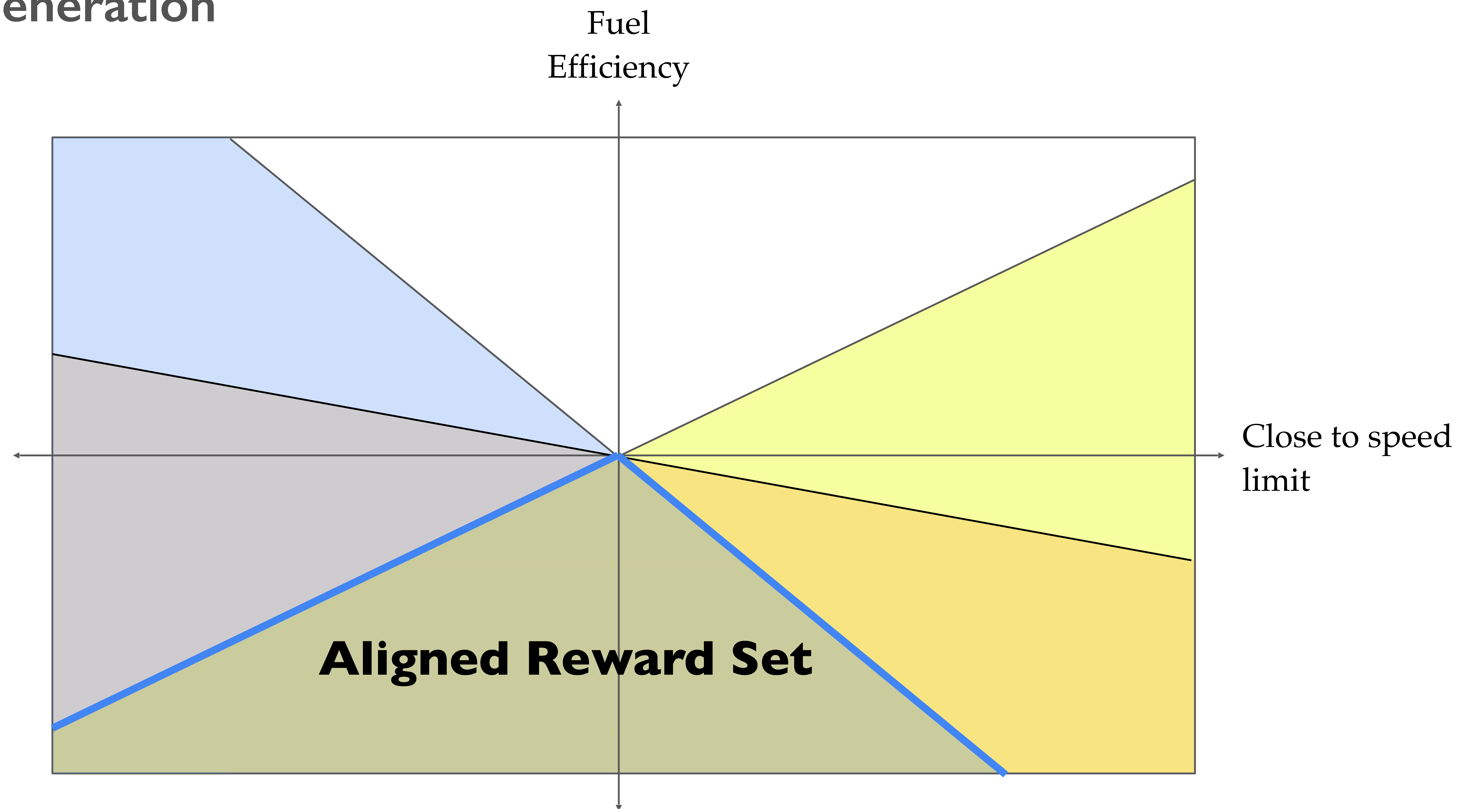
Test Generation



Test Generation

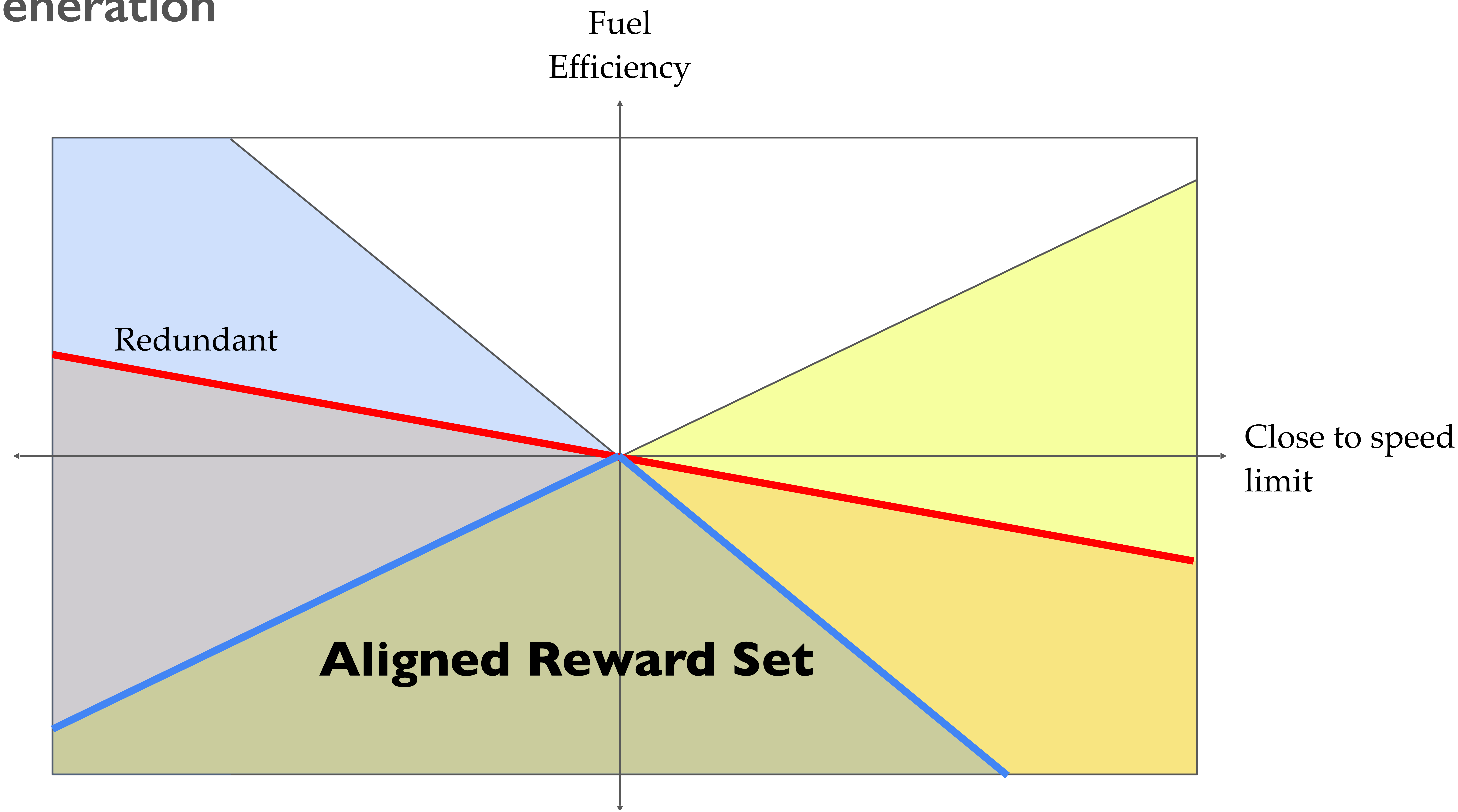


Test Generation



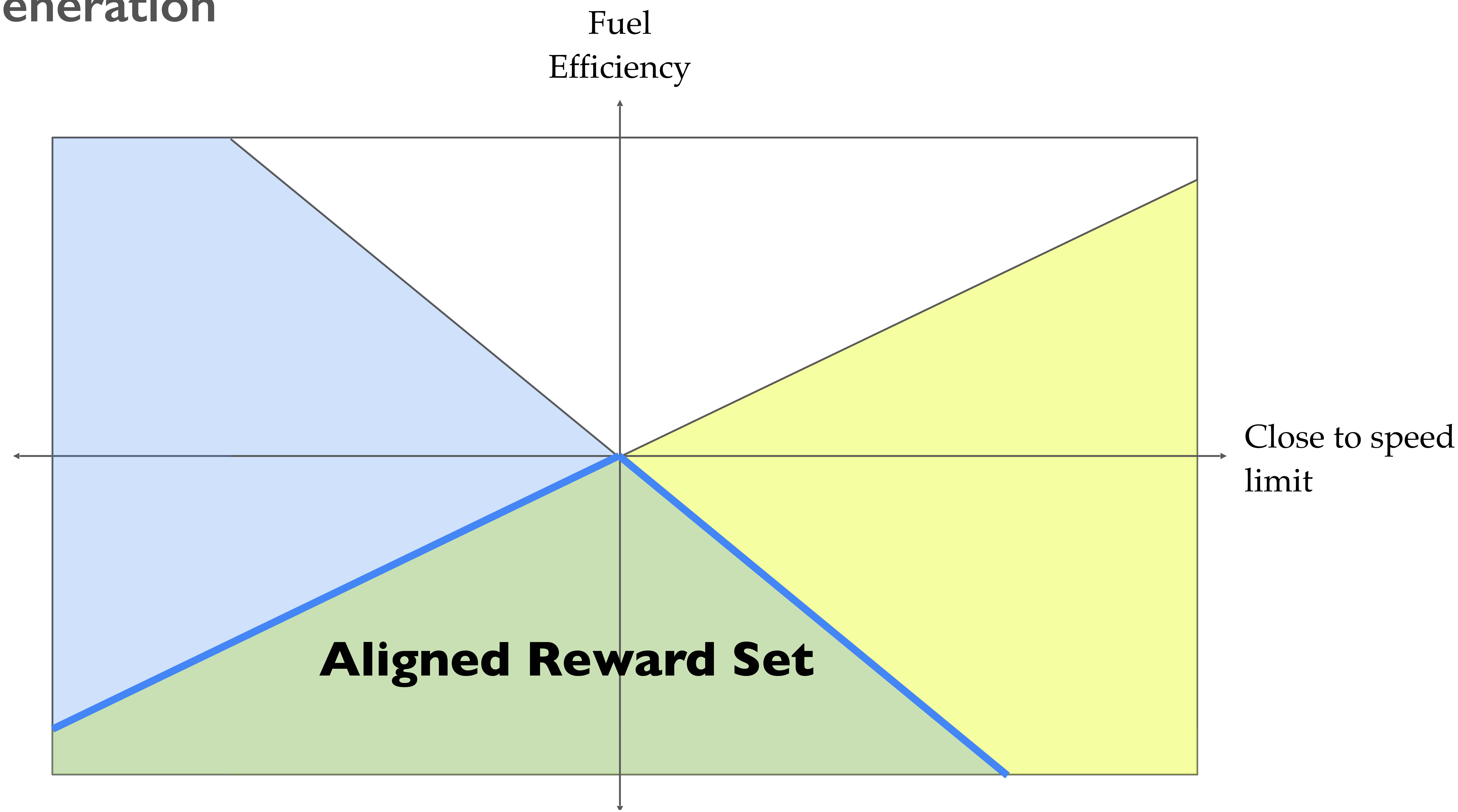
$$ARS(R) = \{R' \mid OPT(R') \subseteq OPT(R)\}.$$

Test Generation



$$ARS(R) = \{R' \mid OPT(R') \subseteq OPT(R)\}.$$

Test Generation



$$ARS(R) = \{R' \mid OPT(R') \subseteq OPT(R)\}.$$

Alignment test conditions

An exact **reward alignment** test can be performed in the following query settings:

Trajectory preferences $\xi_1 \prec \xi_2$

Reward function weights w_{Robot}

Reward samples $R_{\text{Robot}}(s)$

Value samples $V^*_{\text{Robot}}(s),\ Q^*_{\text{Robot}}(s,a)$

Definition: Epsilon (policy) value alignment

Definition 1. *Given reward function R, policy π' is ϵ-**value aligned** in environment E if and only if*

$$V_R^*(s) - V_R^{\pi'}(s) \leq \epsilon, \forall s \in \mathcal{S}. \quad (1)$$

However, with action samples, $\pi^*_{Robot}(s)$, we only have heuristic methods to test **policy alignment**

Alignment test conditions

Trajectory preferences	$\xi_1 \prec \xi_2$	
Reward function weights	w_{Robot}	Exact reward alignment
Reward samples	$R_{\text{Robot}}(s)$	
Value samples	$V^*_{\text{Robot}}(s),\ Q^*_{\text{Robot}}(s,a)$	Exact policy alignment $\epsilon = 0$
Action samples	$\pi^*_{Robot}(s)$	Approx. policy alignment $\epsilon > 0$

Value alignment guarantees

Formal

Efficient "driver's test" that certifies agent alignment

Value alignment verification

Loosen assumptions

Features known $\longrightarrow$ unknown

Preferences noiseless $\longrightarrow$ noisy

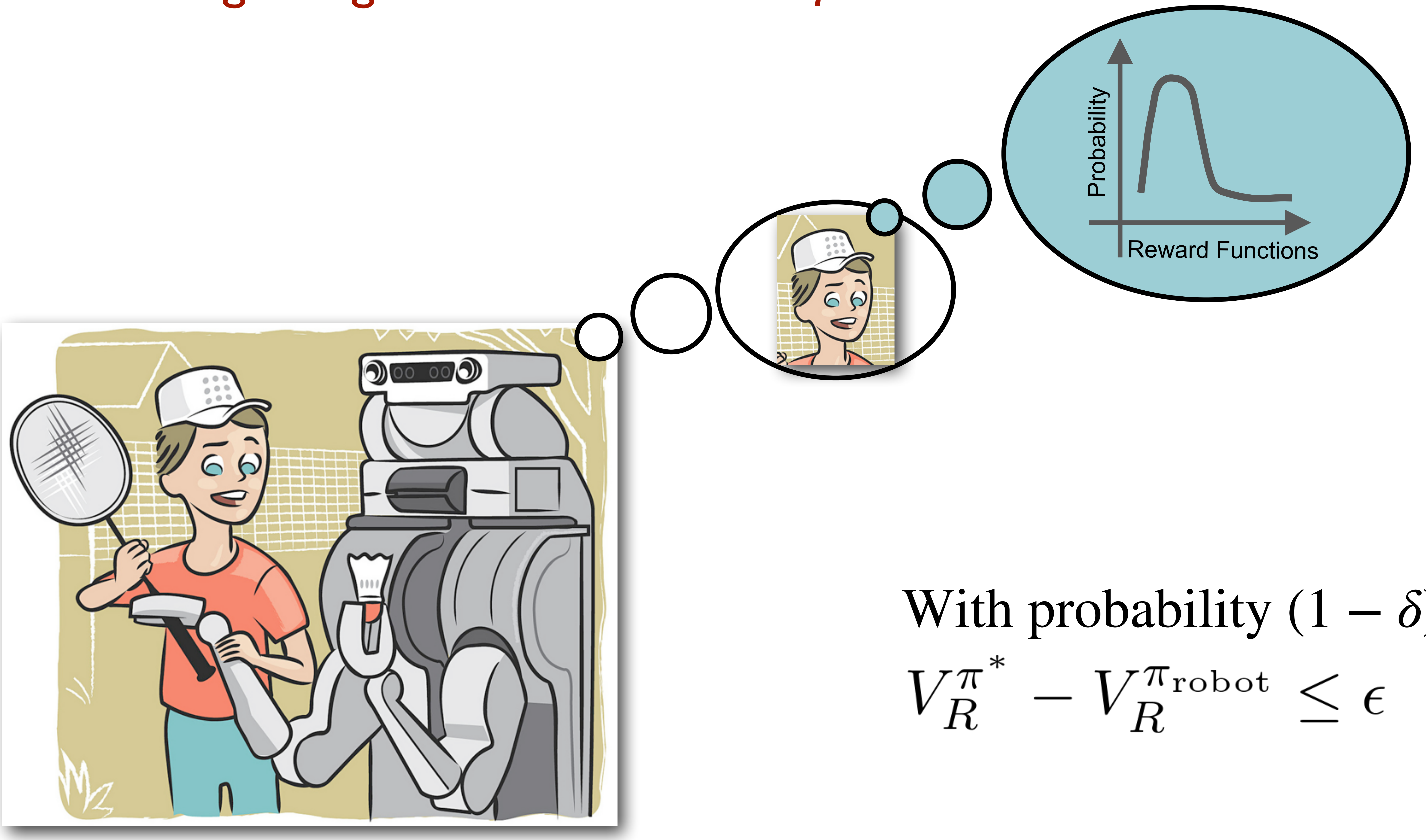
Probabilistic

Quantify / optimize policy risk under reward uncertainty

Bayesian reward extrapolation

RLHF Alignment Guarantees:

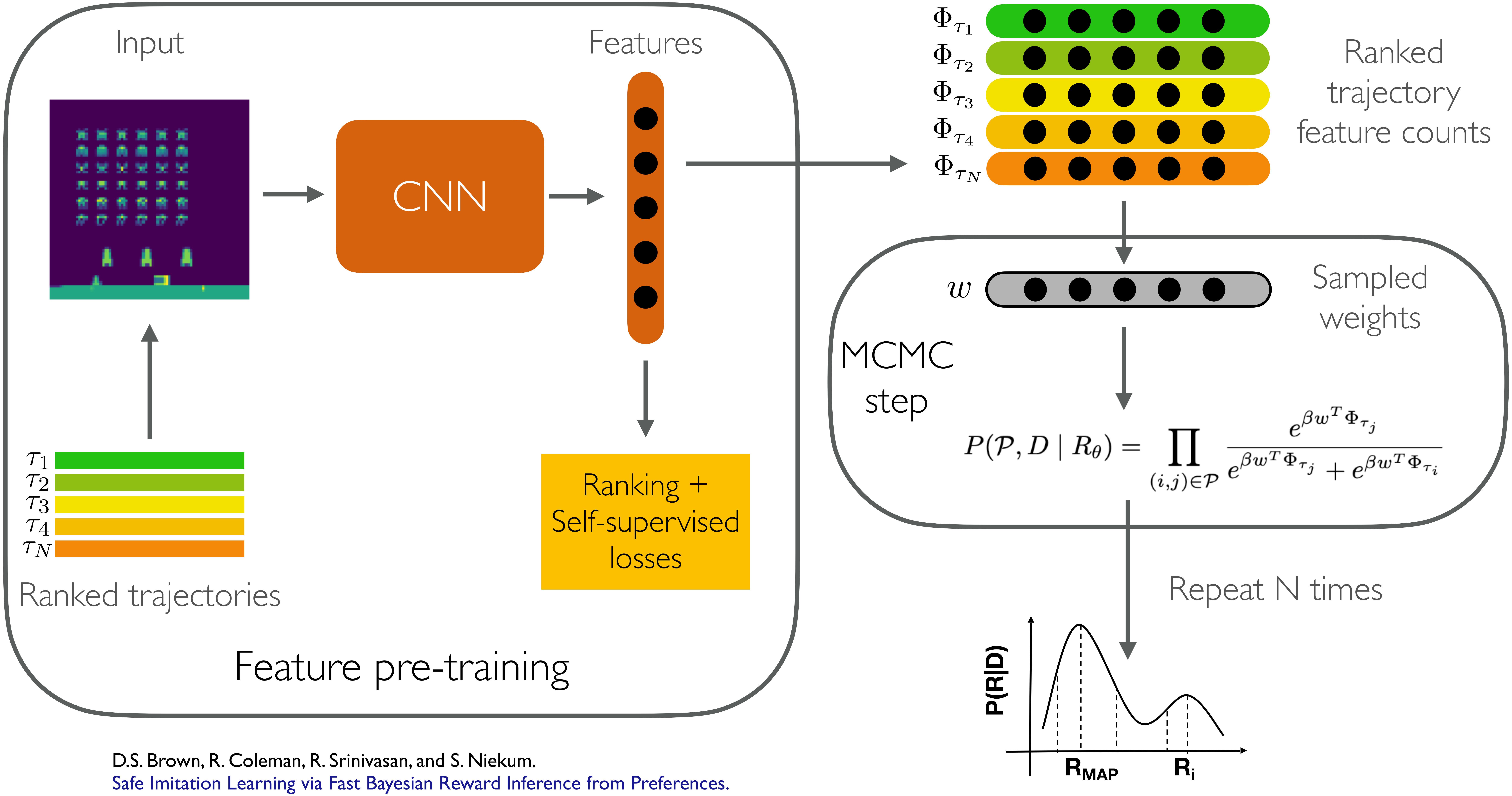
Upper bound the **policy loss** of the robot vs. human demonstrator with **high confidence**, *without knowing the ground-truth reward function.*



With probability $(1 - \delta)$:

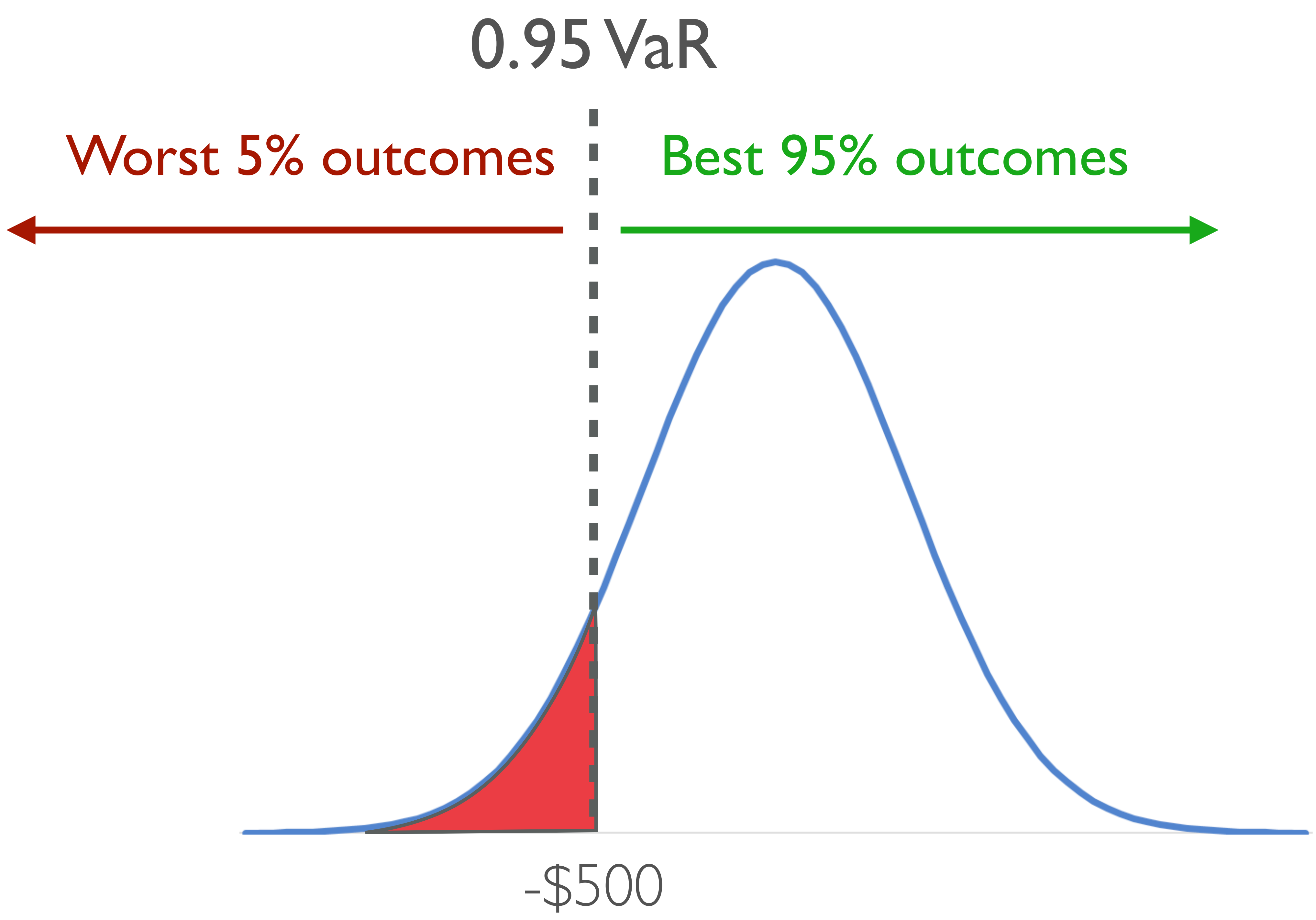
$$V_R^{\pi^*} - V_R^{\pi_{\text{robot}}} \leq \epsilon$$

Quantifying reward function alignment from preferences: Bayesian REX



D.S. Brown, R. Coleman, R. Srinivasan, and S. Niekum.
Safe Imitation Learning via Fast Bayesian Reward Inference from Preferences.
International Conference on Machine Learning (ICML), July 2020.

Reminder: α-value at risk

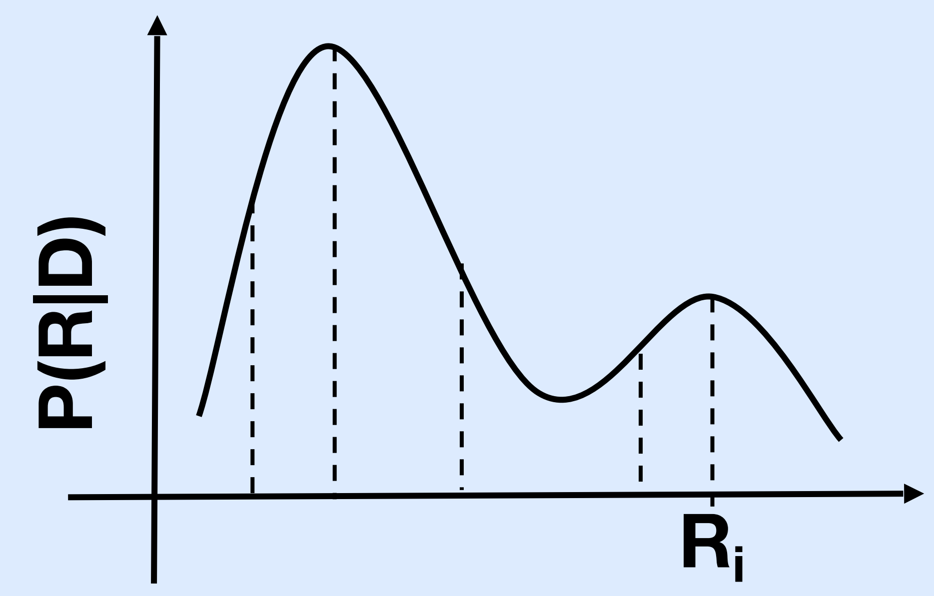


\+ Single-sided confidence bound

"With high confidence, you won't lose more than $500 more than 95% of the time when using this investing strategy"

Producing alignment guarantees

Bayesian REX

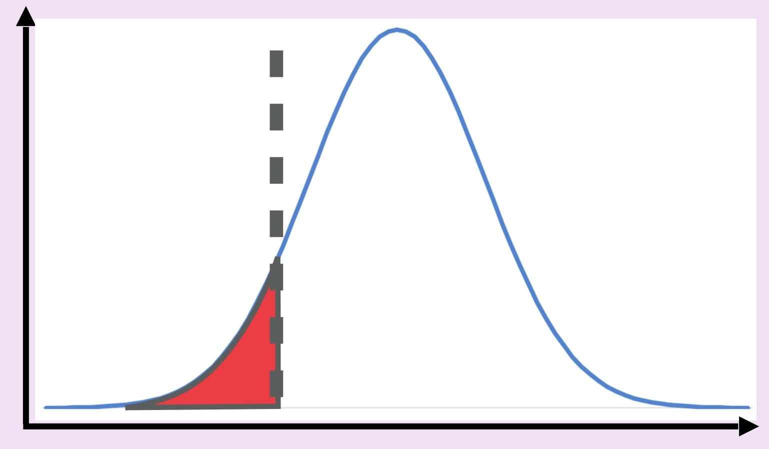


Calculate policy losses

$$V_{R_i}^{\pi_{expert}} - V_{R_i}^{\pi_{eval}}$$

π_{eval} can be any policy, learned or otherwise

Calculate Value at Risk



Plus a single-sided confidence bound

With probability $(1-\delta)$:

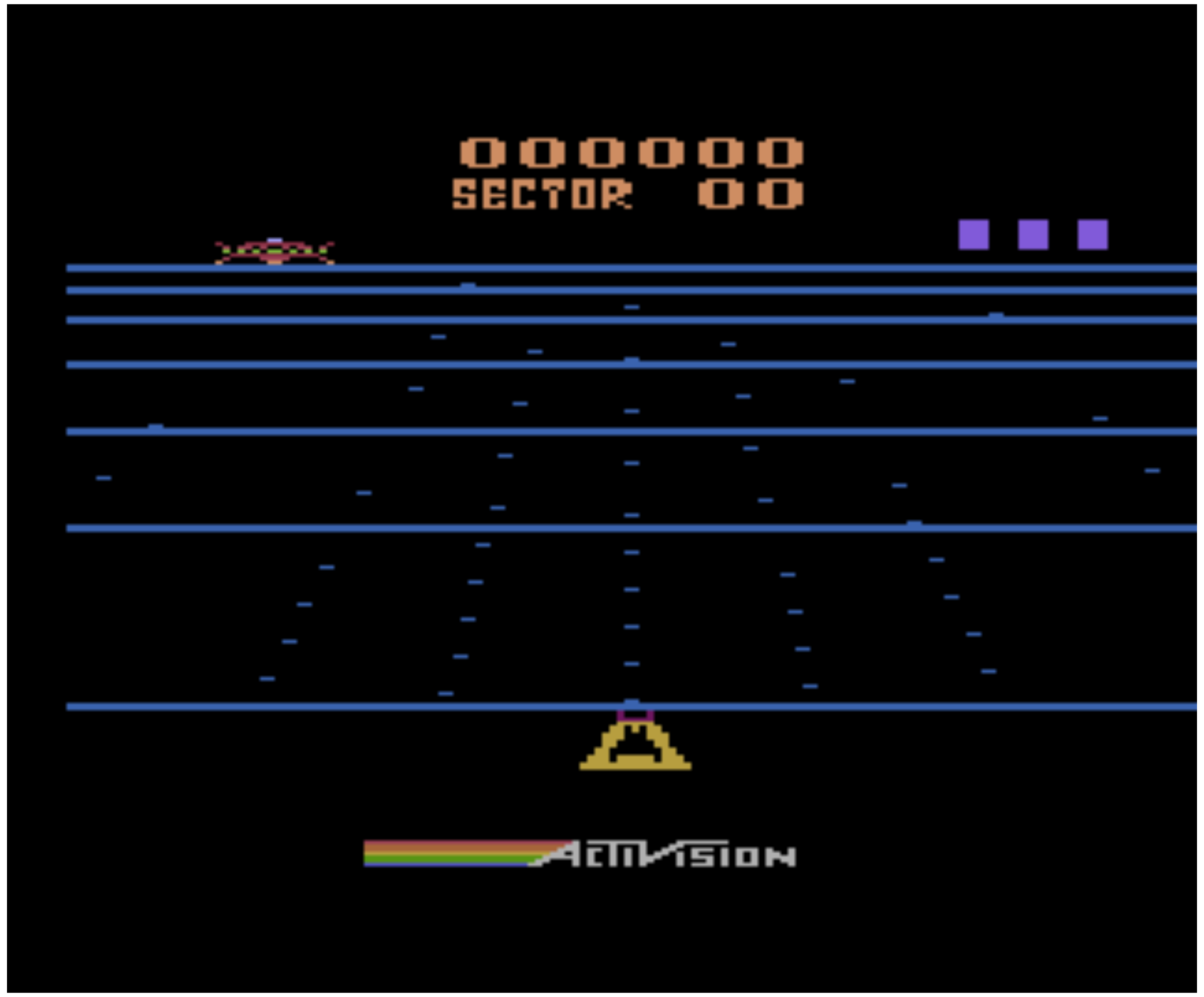
$$\mathrm{VaR}_\alpha[V_R^{\pi_{expert}} - V_R^{\pi_{eval}}] < \epsilon$$

Optimize to minimize risk of π_{eval}

BROIL

D.S. Brown, S. Niekum, and M. Petrik.
Bayesian Robust Optimization for Imitation Learning.
Neural Information Processing Systems, December 2020.

Bayesian REX: Results



Beamrider

Policy	Predicted Mean	Predicted 0.05-VaR	Ground Truth Avg. Score	Ground Truth Avg. Length
A	17.1	7.9	480.6	1372.6
B	22.7	11.9	703.4	1,412.8
C	45.5	24.9	1828.5	2,389.9
D	57.6	31.5	2586.7	2,965.0
No-Op	102.5	-1557.1	0.0	99,994.0

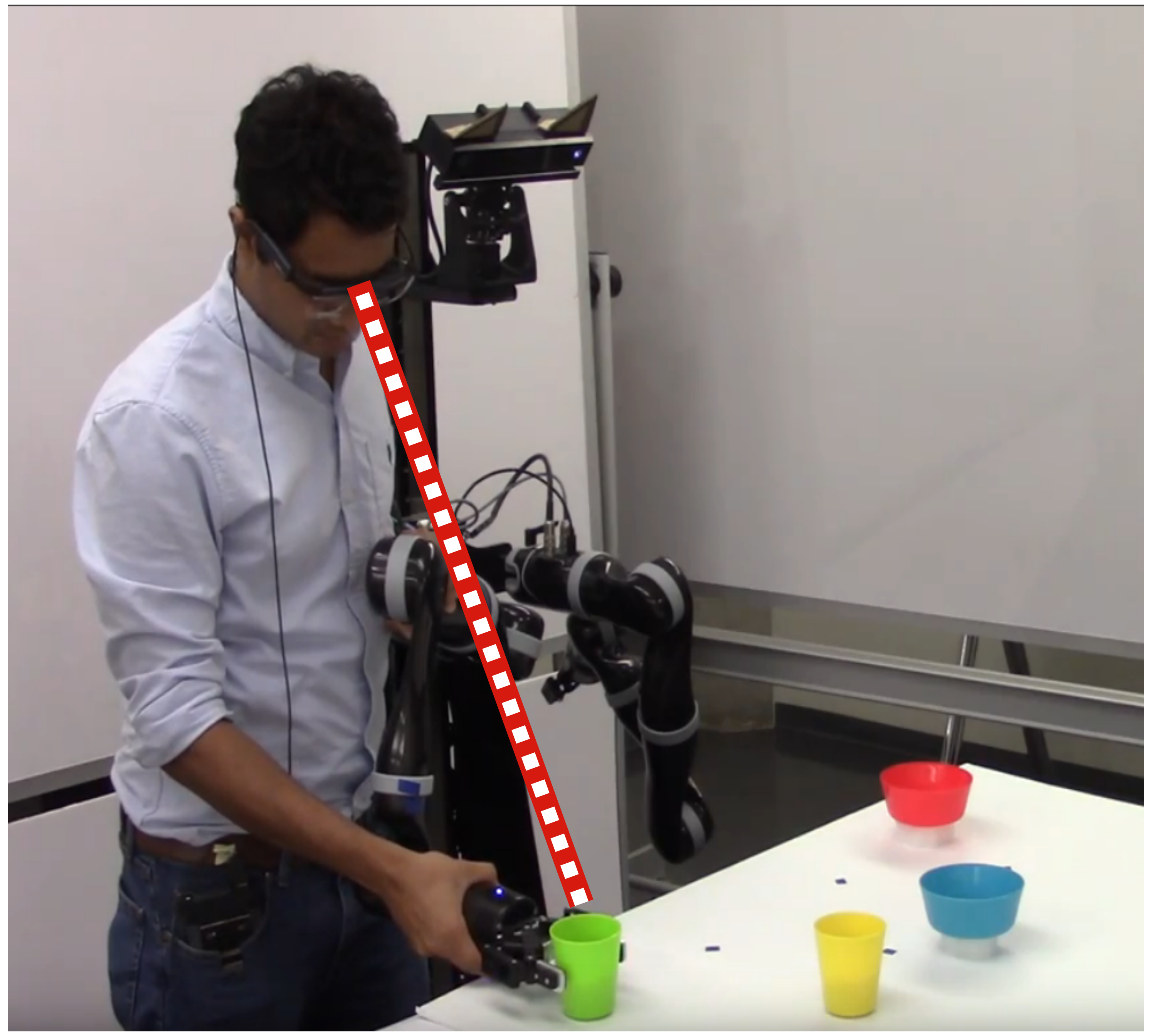
Not restricted to policy evaluation!

Can also learn policy to balance expected return and CVaR:

D.S. Brown, S. Niekum, and M. Petrik.
Bayesian Robust Optimization for Imitation Learning.
Neural Information Processing Systems, December 2020.

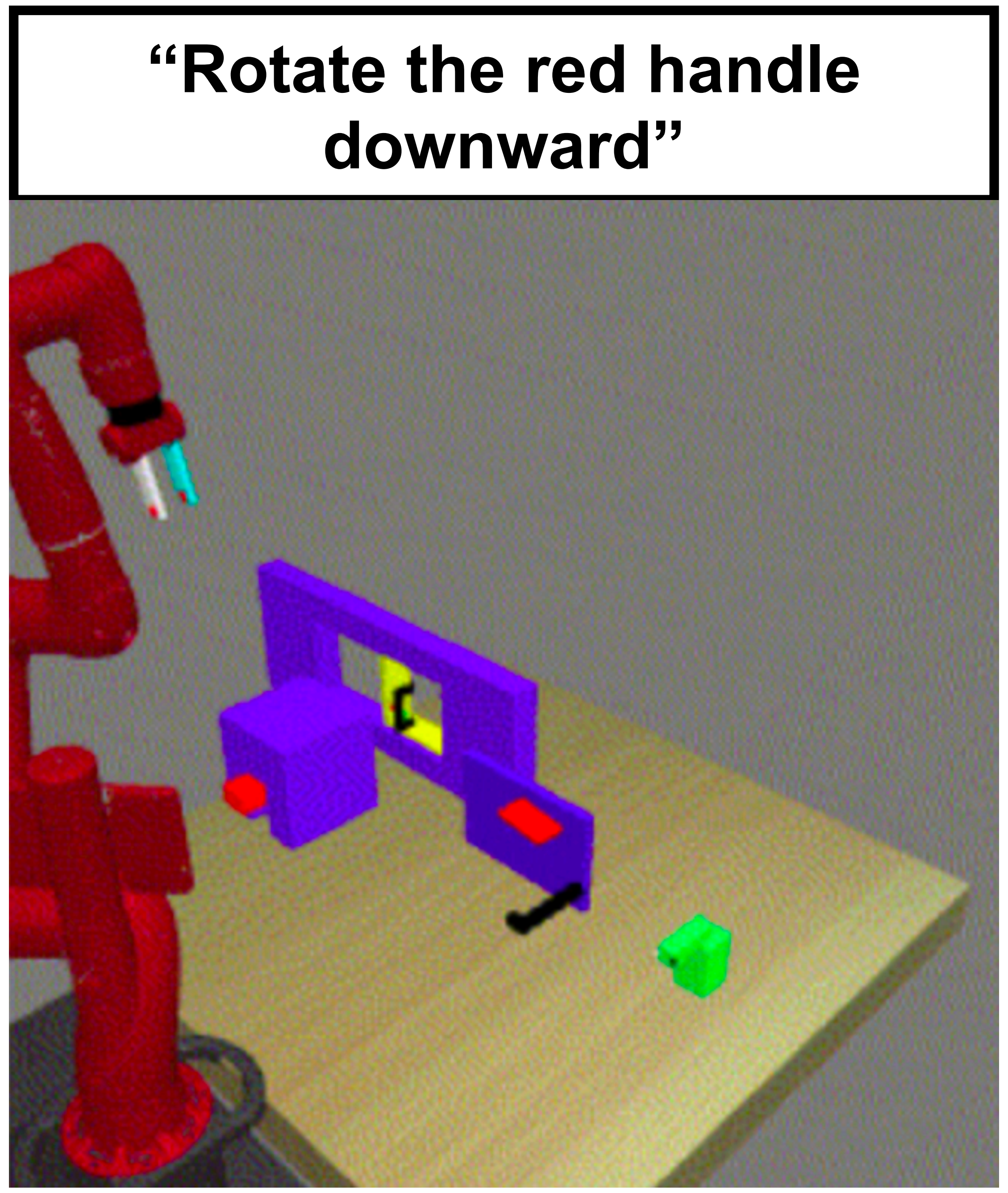
Alignment guarantee frontiers

Multimodal signals (with guarantees?)



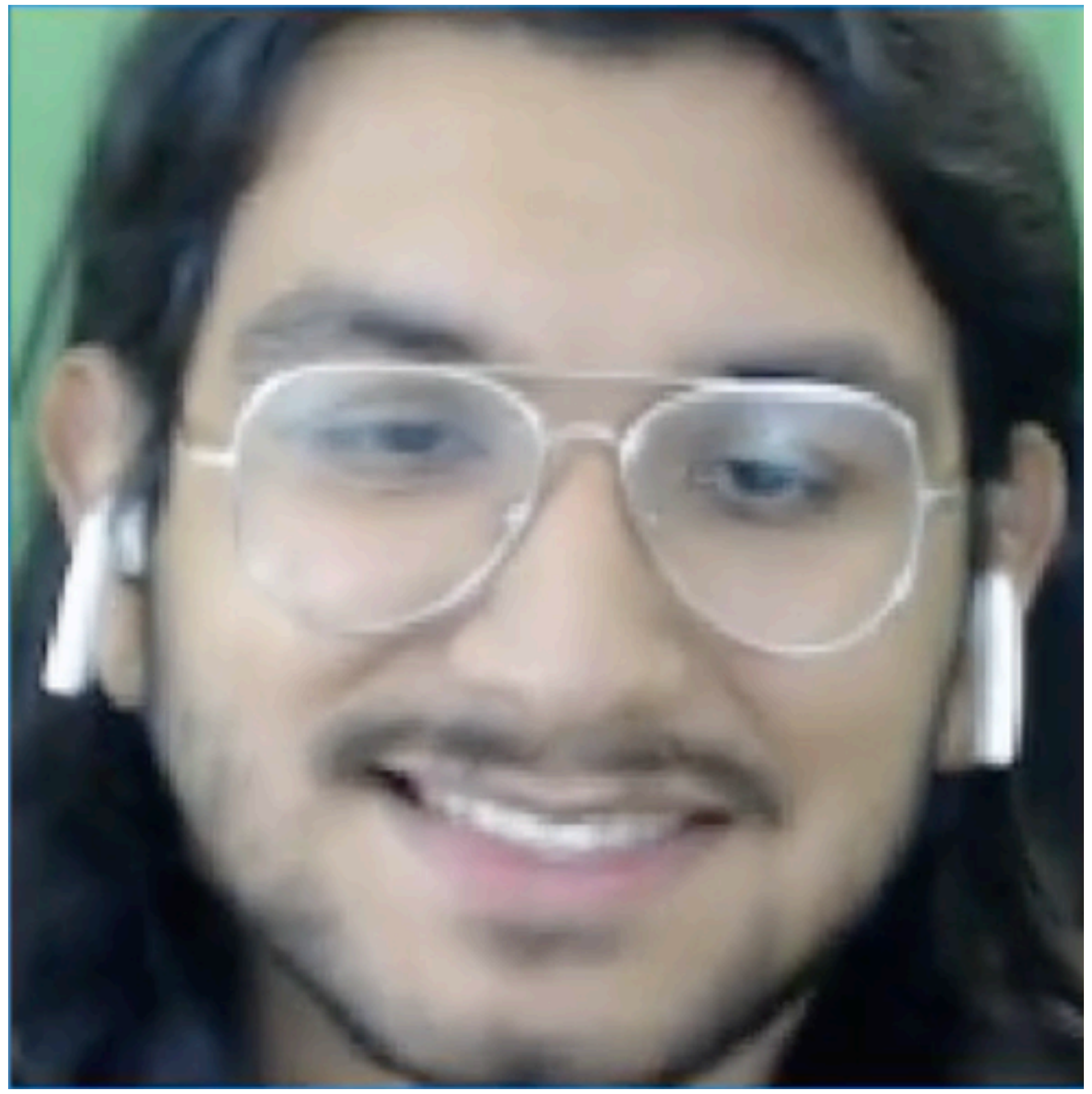
Human Gaze

A. Saran, E.S. Short, A.L. Thomaz, and S. Niekum.
Understanding Teacher Gaze Patterns for Robot Learning.
Conference on Robot Learning (CoRL), October 2019.



Natural language

P. Goyal, S. Niekum, and R. Mooney.
PixL2R: Guiding Reinforcement Learning Using Natural Language by Mapping Pixels to Rewards.
Conference on Robot Learning (CoRL), November 2020.



Facial reactions

Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. Knox.
The EMPATHIC Framework for Task Learning from Implicit Human Feedback.
Conference on Robot Learning (CoRL), November 2020.