CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Models of Human Preference

Learning reward functions from preferences

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

Deep Reinforcement Learning from Human Preferences

Paul F Christiano, OpenAI, paul@openai.com
Jan Leike, DeepMind, leike@google.com
Tom B Brown, nottombrown@gmail.com
Miljan Martic, DeepMind, miljanm@google.com
Shane Legg, DeepMind, legg@google.com
Dario Amodei, OpenAI, damodei@openai.com

B-Pref: Benchmarking Preference-Based Reinforcement Learning

Kimin Lee, Laura Smith, Anca Dragan, Pieter Abbeel
UC Berkeley

Active Preference-Based Learning of Reward Functions

Dorsa Sadigh, Anca D. Dragan, Shankar Sastry, and Sanjit A. Seshia
University of California, Berkeley, {dsadigh, anca, sastry, sseshia}@eecs.berkeley.edu

Value Alignment Verification

Daniel S. Brown [*1] Jordan Schneider [*2] Anca Dragan [1] Scott Niekum [2]

Safe Imitation Learning via Fast Bayesian Reward Inference from Preferences

Daniel S. Brown [1] Russell Coleman [1 2] Ravi Srinivasan [2] Scott Niekum [1]

Article

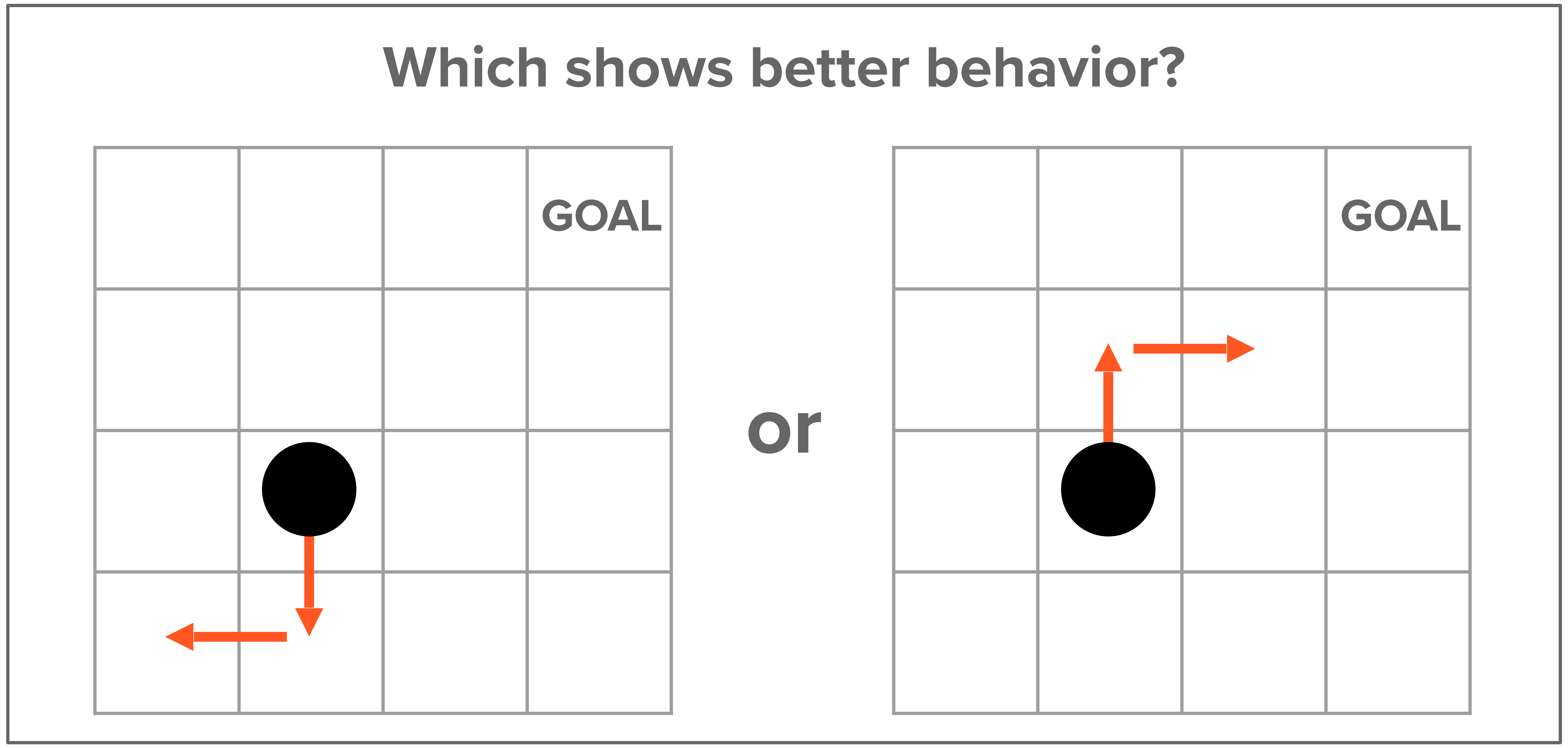
The International Journal of Robotics Research
2022, Vol. 41(1) 45–67

DOI: 10.1177/02783649211041652
journals.sagepub.com/home/ijr
SAGE

Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences

Erdem Bıyık[1], Dylan P. Losey[2], Malayandi Palan[2], Nicholas C. Landolfi[2], Gleb Shevchuk[2] and Dorsa Sadigh[1,2]

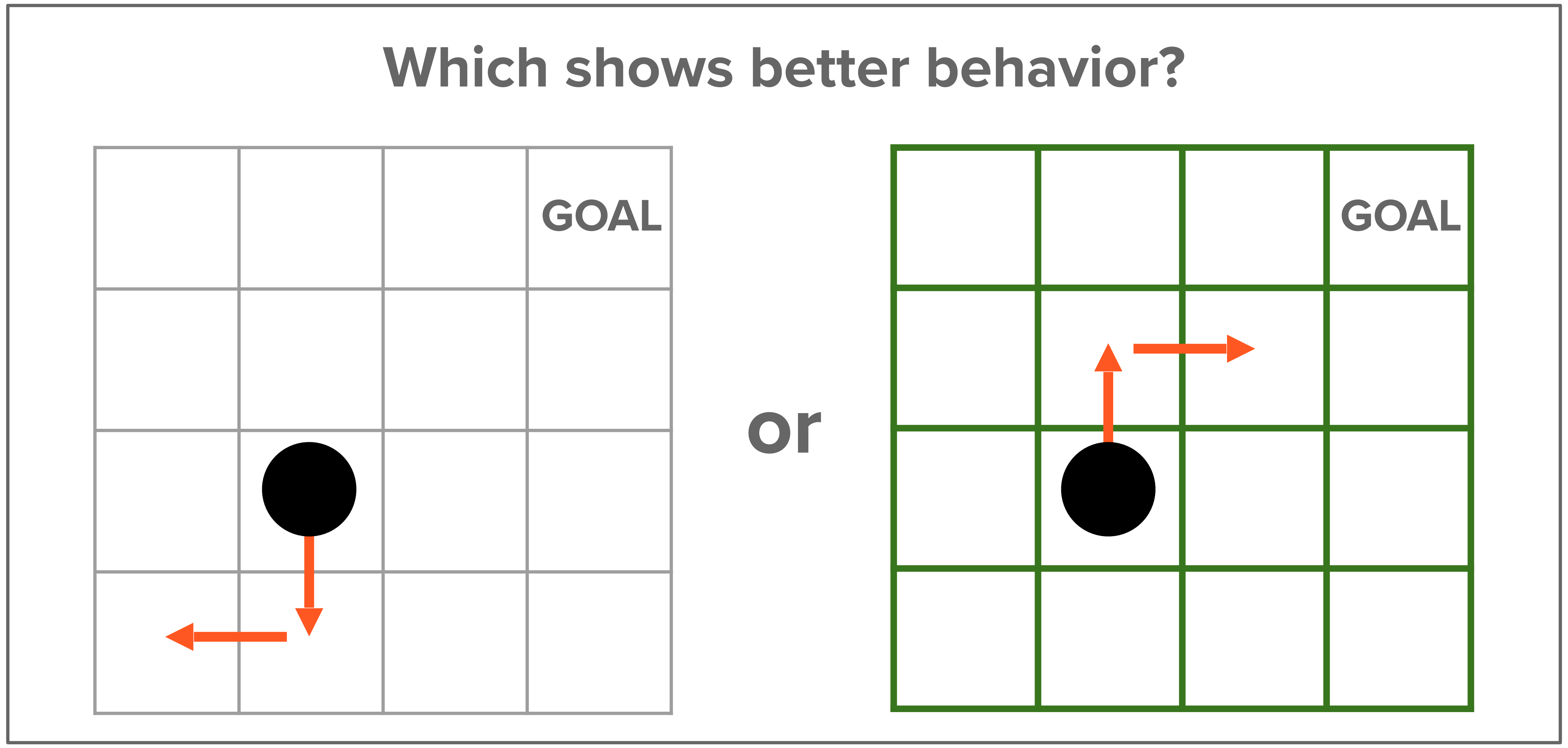
Preferences over segment pairs



W. Knox, S. Hatgis-Kessell, S. Booth, S. Niekum, P. Stone, A. Allievi.
Models of Human Preference for Learning Reward Functions.
Transactions on Machine Learning Research (TMLR), January 2024.

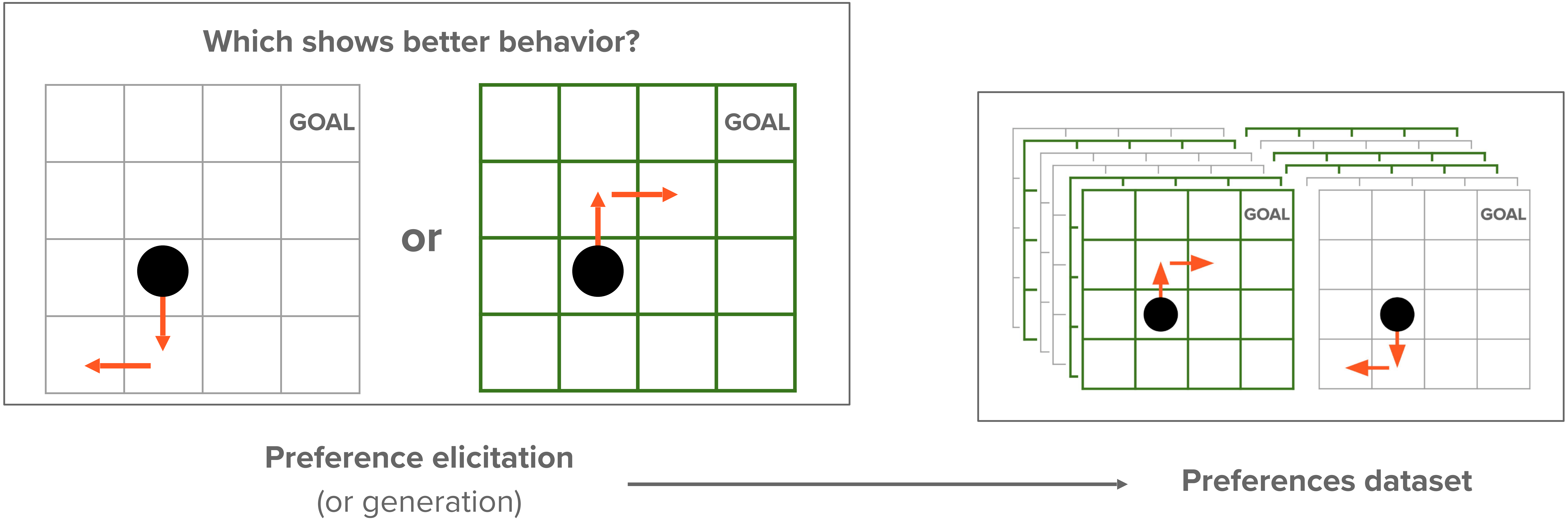
Slide credit: W. Bradley Knox

Preferences over segment pairs



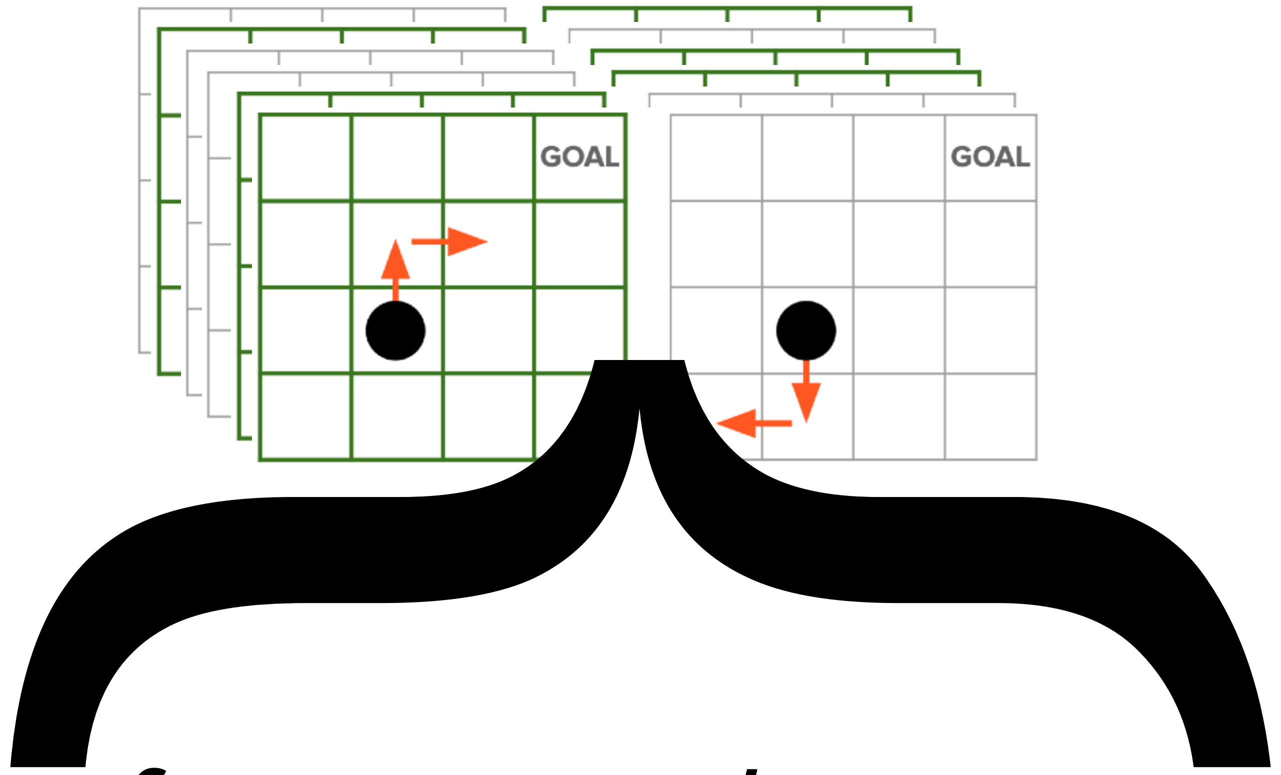
Slide credit: W. Bradley Knox

Preferences over segment pairs



Slide credit: W. Bradley Knox

Learning a reward function from preferences



Given a preference model $P(\sigma_1 \succ \sigma_2 | \hat{r})$

optimize $\hat{r}$ to maximize the likelihood of the *preferences dataset.*

Slide credit: W. Bradley Knox

Why preferences?

- Established technique in reward learning
- Intuitive for humans
- Judgment may be easier than control
- Connects to expected utility theory
- *In ideal settings, the reward function underlying the preferences can be recovered*

The missing piece: the model of preference

$$P(\sigma_1 \succ \sigma_2) = \frac{\exp\left[f(\sigma_1)\right]}{\exp\left[f(\sigma_1)\right] + \exp\left[f(\sigma_2)\right]}$$

$$= logistic(f(\sigma_1) - f(\sigma_2))$$

The missing piece: the model of preference

$$P(\sigma_1 \succ \sigma_2) = logistic\Big(f(\sigma_1) - f(\sigma_2)\Big)$$

Slide credit: W. Bradley Knox

The missing piece: the model of preference

$$P(\sigma_1 \succ \sigma_2) = logistic\Big(f(\sigma_1) - f(\sigma_2)\Big)$$

Current dominant model:

Partial return

$$f(\sigma) = \text{sum of reward in } \sigma, \quad \boxed{\sum_{t=0}^{|\sigma|-1} \gamma^t \tilde{r}_t}$$

$$P(\sigma_1 \succ \sigma_2) = logistic(\Sigma_{\sigma_1} r - \Sigma_{\sigma_2} r)$$

Slide credit: W. Bradley Knox

The missing piece: the model of preference

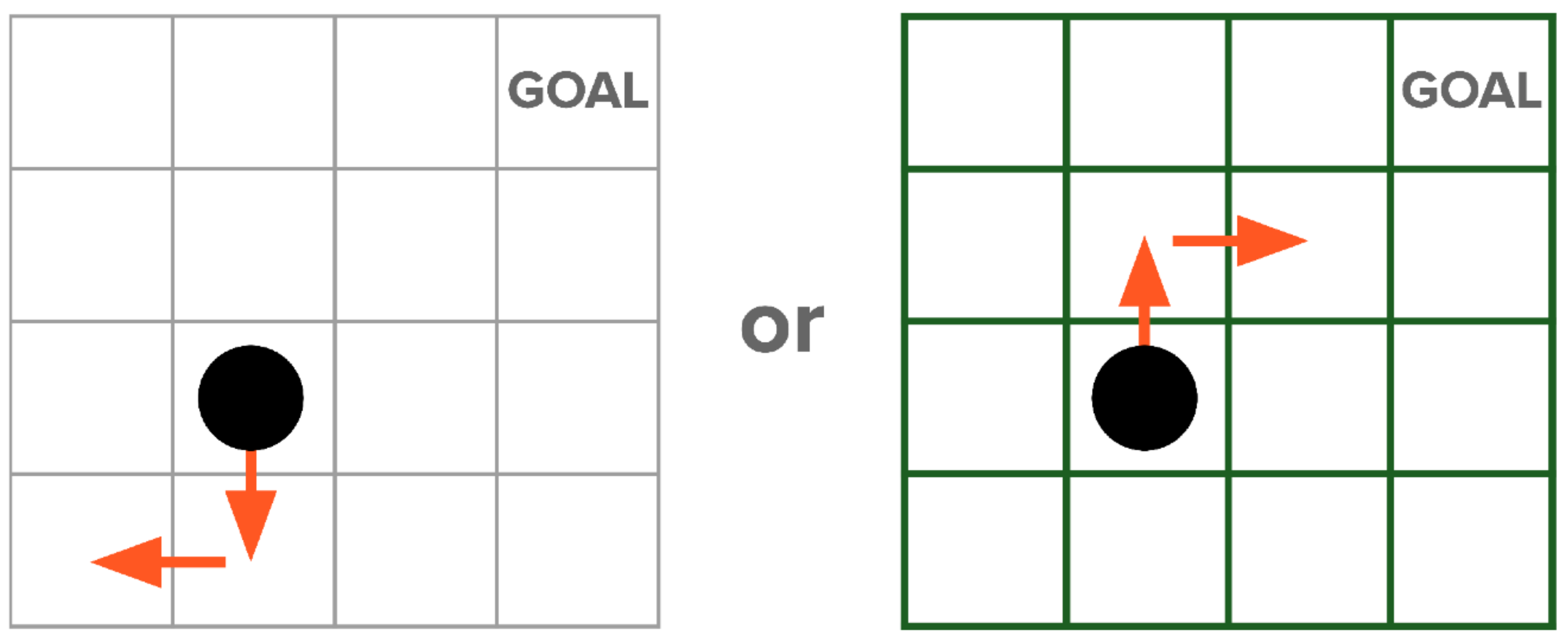
$$P(\sigma_1 \succ \sigma_2) = logistic\Big(f(\sigma_1) - f(\sigma_2)\Big)$$

Partial return: $f(\sigma) = \text{sum of reward in } \sigma$

Assume -1 reward per step.

Partial return is indifferent!

Which shows better behavior?



σ_1 σ_2

Slide credit: W. Bradley Knox

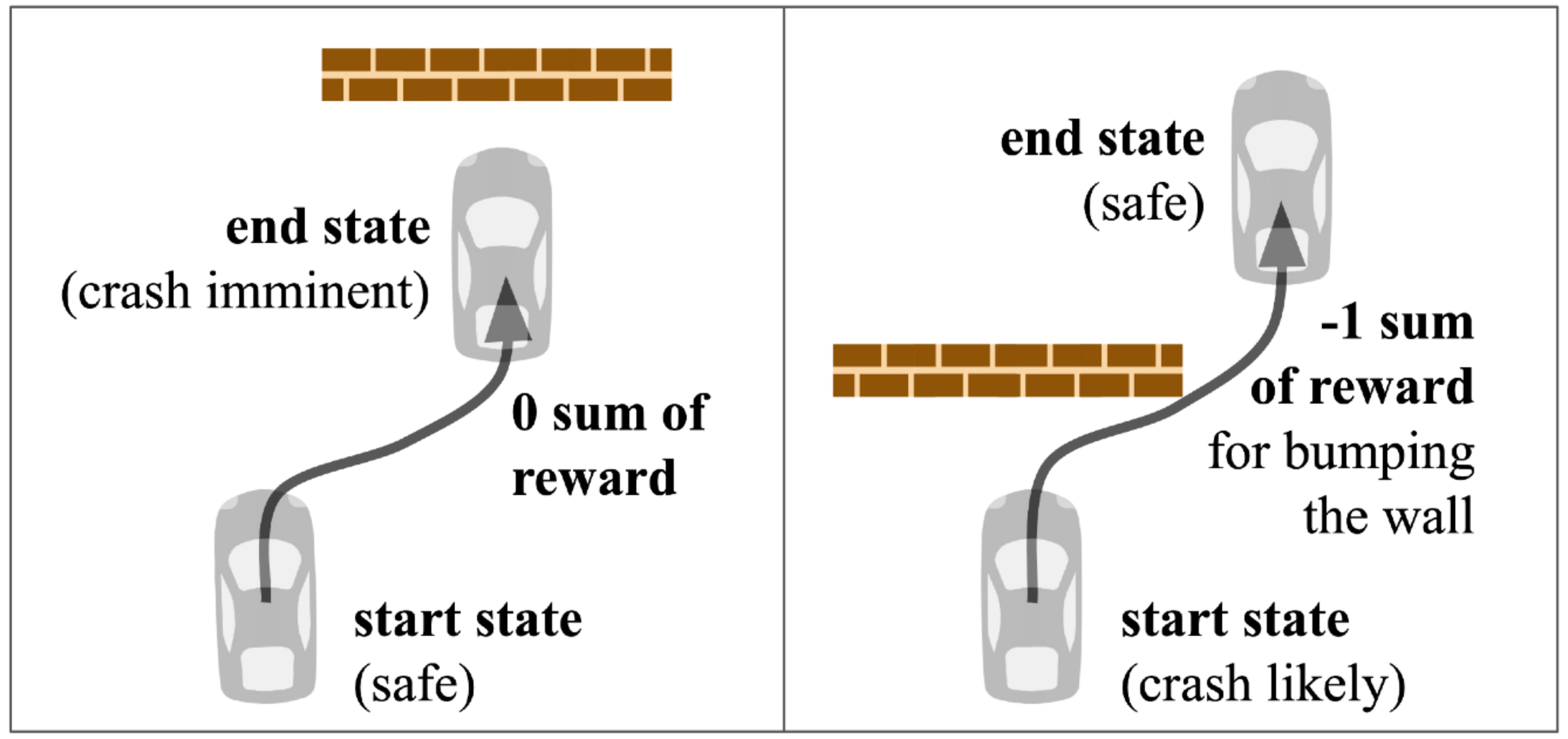
The missing piece: the model of preference

$$P(\sigma_1 \succ \sigma_2) = logistic\Big(f(\sigma_1) - f(\sigma_2)\Big)$$

Partial return: $f(\sigma) = \text{sum of reward in } \sigma$

Partial return prefers the left segment!

Which shows better behavior?



σ_1 σ_2

Slide credit: W. Bradley Knox

The missing piece: the model of preference

$$P(\sigma_1 \succ \sigma_2) = logistic\Big(f(\sigma_1) - f(\sigma_2)\Big)$$

Proposed preference model: Regret

$$f(\sigma) = -regret(\sigma)$$

$$= \text{sum of } A^*(s,a) \text{ for each } (s,a) \text{ in } \sigma$$

$$regret(\sigma|\tilde{r}) = \sum_{t=0}^{|\sigma|-1} regret(\sigma_t|\tilde{r}) = \sum_{t=0}^{|\sigma|-1} \Big[V_{\tilde{r}}^*(s_{\sigma,t}) - Q_{\tilde{r}}^*(s_{\sigma,t}, a_{\sigma,t})\Big] = \sum_{t=0}^{|\sigma|-1} -A_{\tilde{r}}^*(s_{\sigma,t}, a_{\sigma,t})$$

when all transitions are deterministic

$$regret_d(\sigma|\tilde{r}) \triangleq \sum_{t=0}^{|\sigma|-1} regret_d(\sigma_t|\tilde{r}) = V_{\tilde{r}}^*(s_{\sigma,0}) - (\Sigma_\sigma \tilde{r} + V_{\tilde{r}}^*(s_{\sigma,|\sigma|}))$$

Slide credit: W. Bradley Knox

The missing piece: the model of preference

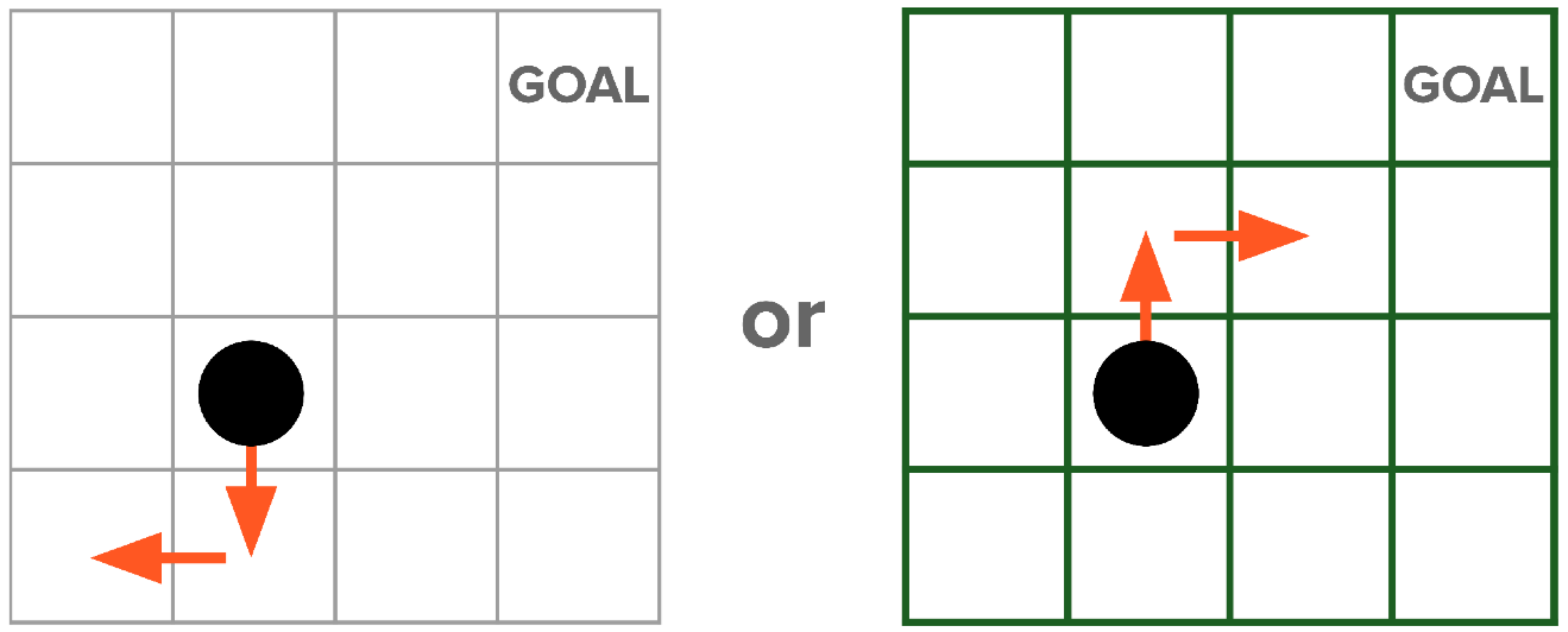
$$P(\sigma_1 \succ \sigma_2) = logistic\Big(f(\sigma_1) - f(\sigma_2)\Big)$$

Regret: $f(\sigma) = \text{sum of } A^*(s, a) \text{ for each } (s, a) \text{ in } \sigma$

Which shows better behavior?

Assume -1 reward per step.

Regret prefers σ_2.



σ_1 σ_2

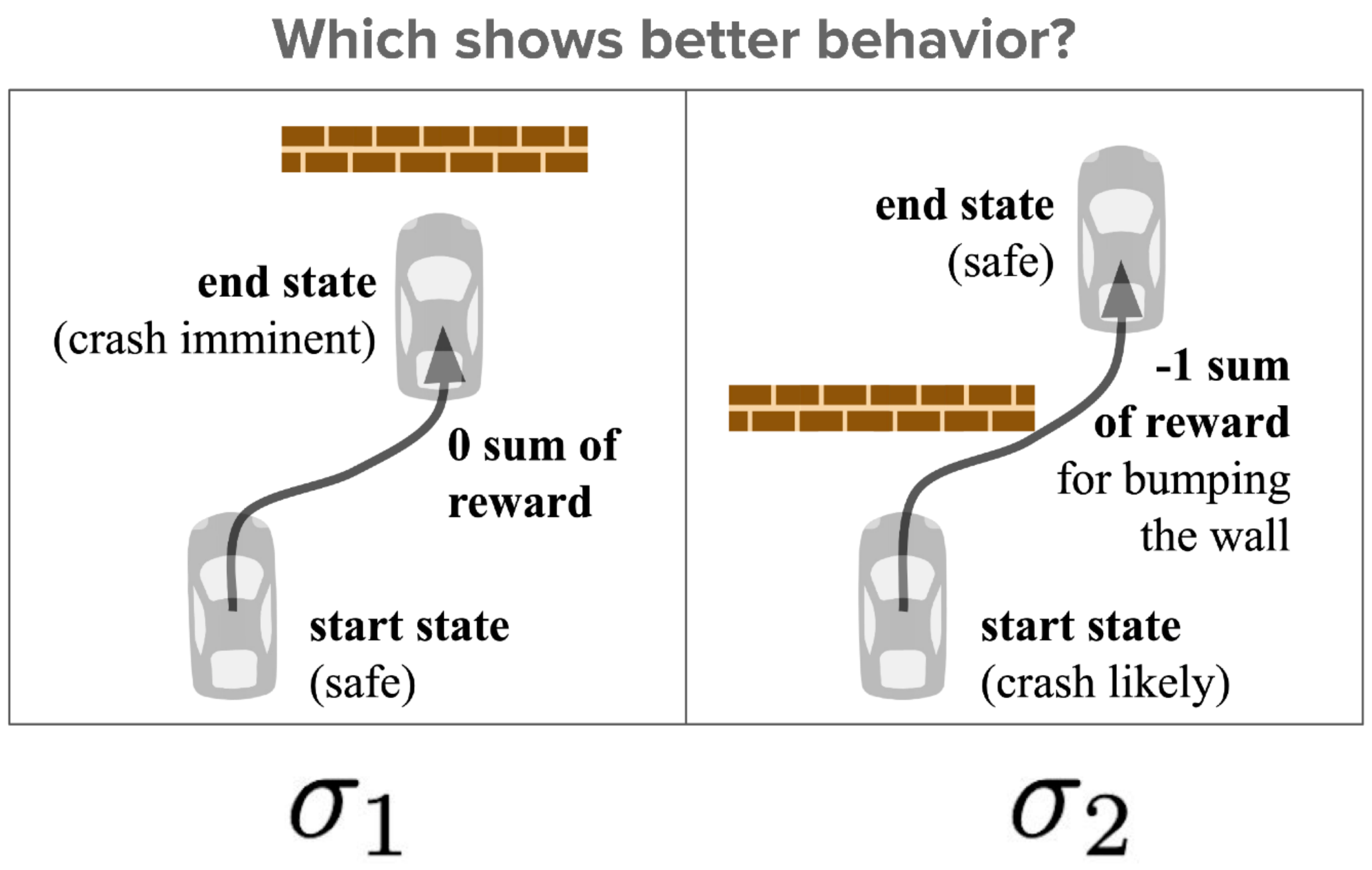
Slide credit: W. Bradley Knox

The missing piece: the model of preference

$$P(\sigma_1 \succ \sigma_2) = logistic\Big(f(\sigma_1) - f(\sigma_2)\Big)$$

Regret: $f(\sigma) = \text{sum of } A^*(s,a) \text{ for each } (s,a) \text{ in } \sigma$

Regret prefers σ_2.



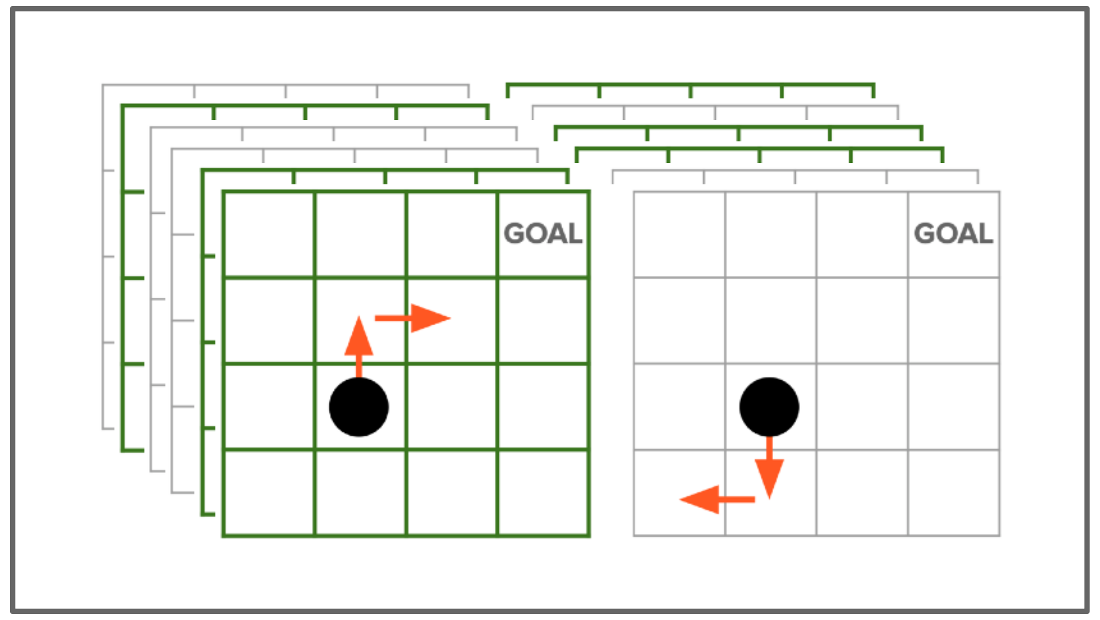
Slide credit: W. Bradley Knox

Reward Identifiability

Reward identifiability

definition:

Ground truth r $\longrightarrow$ $P(\sigma_1 \succ \sigma_2 | r)$ $\longrightarrow$



Infinitely exhaustive preferences dataset

Equivalent
(same set of optimal policies)

Learned $\hat{r}$

∃ an algorithm that guarantees …

Given preferences generated by a preference model and a reward function, where the preferences infinitely cover every segment pair,
does the preference dataset contain sufficient information to recover an equivalent reward function?

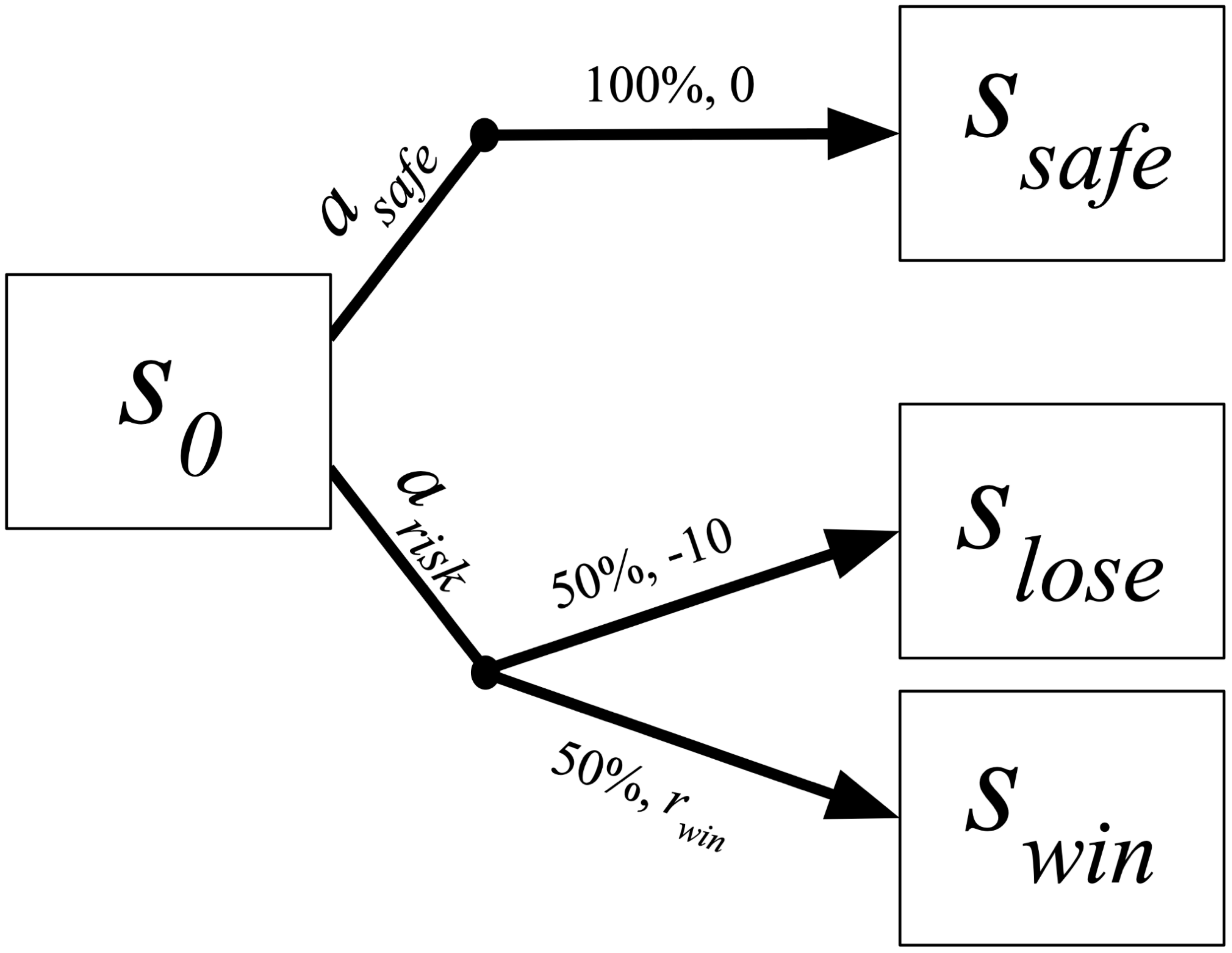
Slide credit: W. Bradley Knox

Reward identifiability

Reward is identifiable with **regret**-based preferences for any MDP.

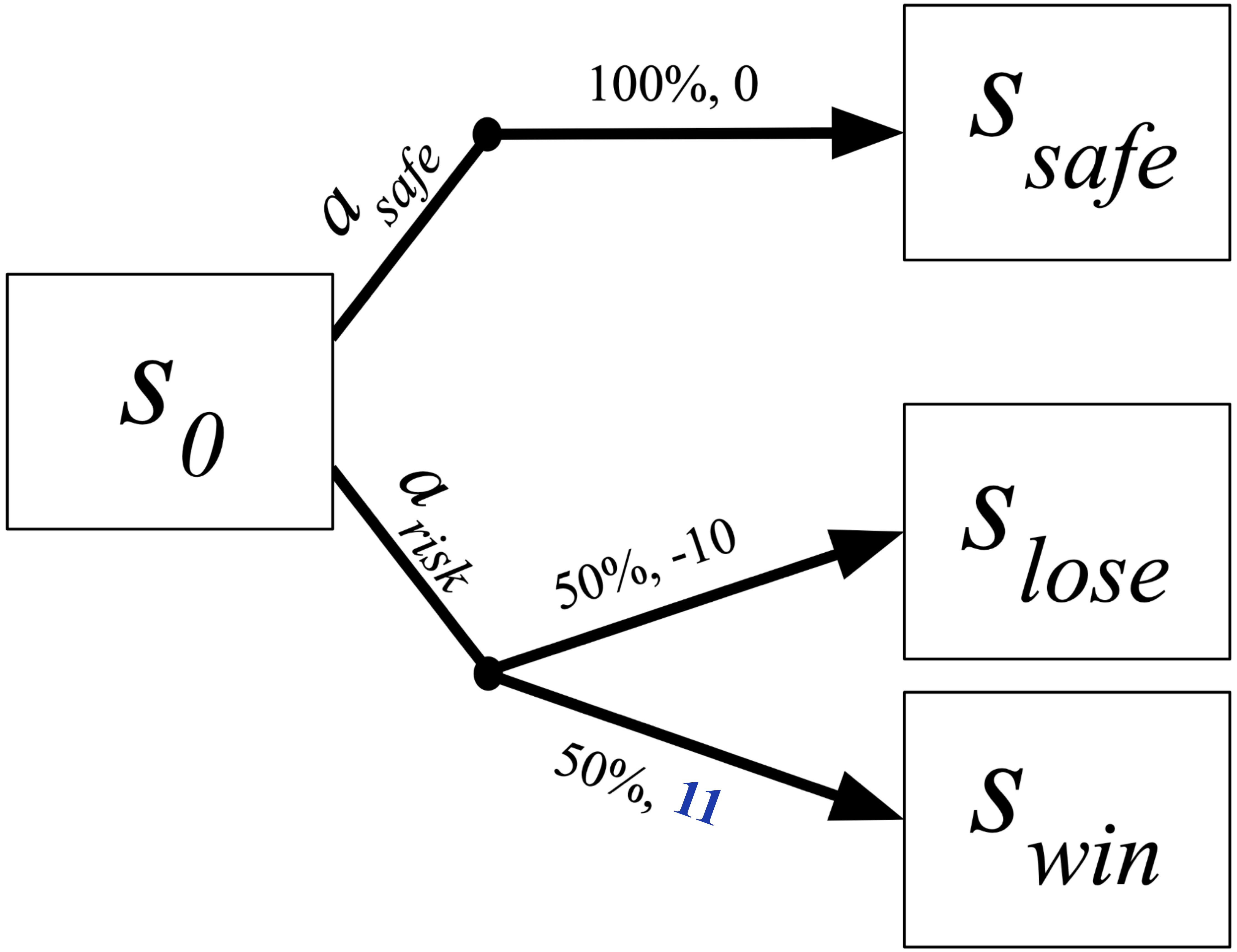
Reward identifiability

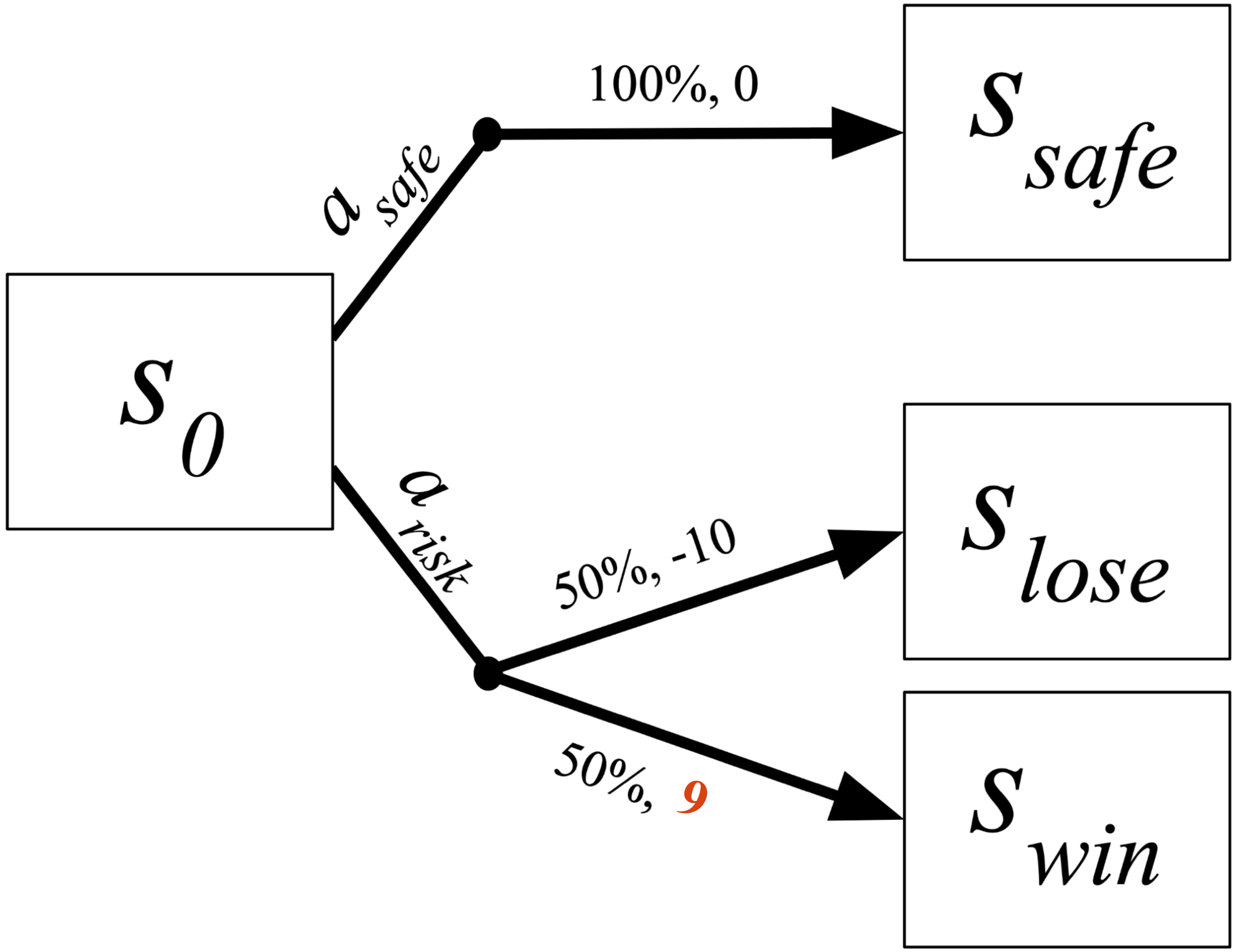
With partial return, reward is not generally identifiable without preference noise that reveals rewards' relative proportions.



Slide credit: W. Bradley Knox

Reward identifiability





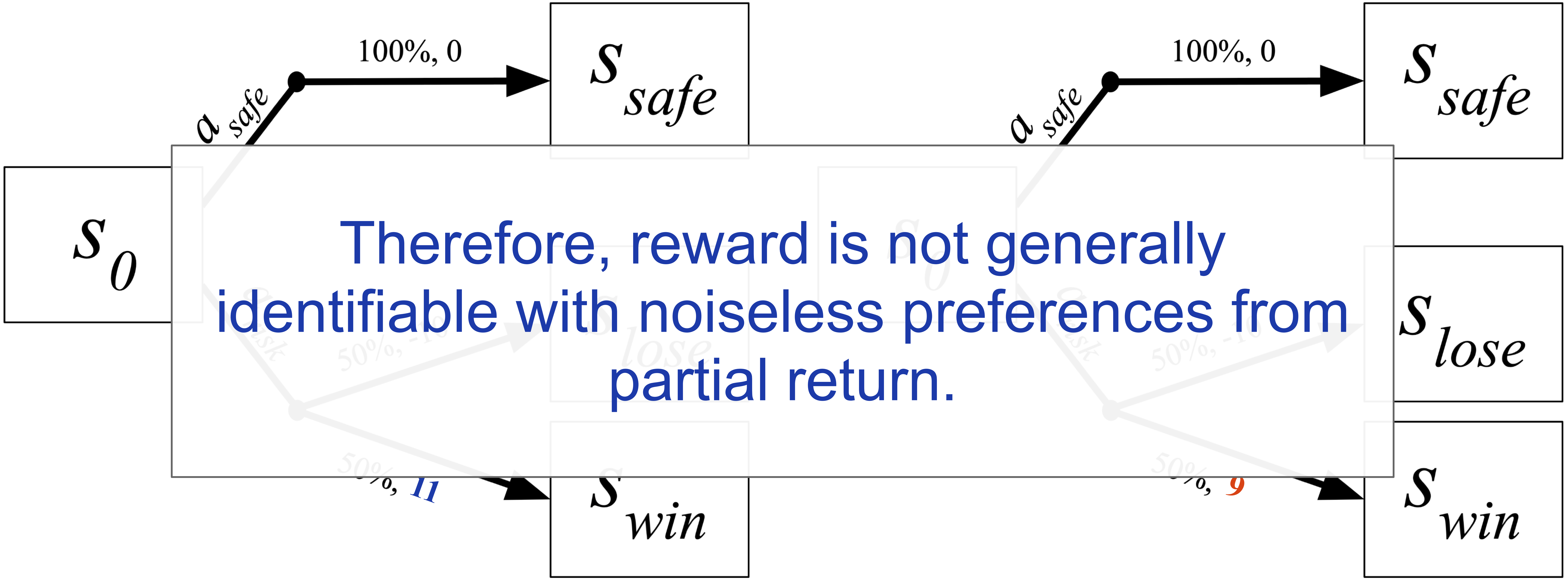
If $r_{win} = 11$, a_{risk} is optimal.

If $r_{win} = 9$, a_{safe} is optimal.

Yet both create the same (noiseless) preferences!!

Slide credit: W. Bradley Knox

Reward identifiability



If $r_{win} = 11$, a_{risk} is optimal.

If $r_{win} = 9$, a_{safe} is optimal.

Yet both create the same (noiseless) preferences!!

Slide credit: W. Bradley Knox

How to learn under the regret-based model?

The regret preference model

$$P_{regret}(\sigma_1 \succ \sigma_2 | \tilde{r}) \triangleq logistic\Big(regret(\sigma_2|\tilde{r}) - regret(\sigma_1|\tilde{r})\Big)$$

Efficiently estimating value functions

$$P(\sigma_1 \succ \sigma_2) = logistic\Big(f(\sigma_1) - f(\sigma_2)\Big)$$

Regret preference model

$$f(\sigma) = -regret(\sigma)$$
$$= \text{sum of } A^*(s,a) \text{ for each } (s,a) \text{ in } \sigma$$

$$regret(\sigma|\tilde{r}) = \sum_{t=0}^{|\sigma|-1} regret(\sigma_t|\tilde{r}) = \sum_{t=0}^{|\sigma|-1} \left[V^*_{\tilde{r}}(s_{\sigma,t}) - Q^*_{\tilde{r}}(s_{\sigma,t}, a_{\sigma,t})\right] = \sum_{t=0}^{|\sigma|-1} -A^*_{\tilde{r}}(s_{\sigma,t}, a_{\sigma,t})$$

$$regret_d(\sigma|\tilde{r}) \triangleq \sum_{t=0}^{|\sigma|-1} regret_d(\sigma_t|\tilde{r}) = V^*_{\tilde{r}}(s_{\sigma,0}) - (\Sigma_\sigma \tilde{r} + V^*_{\tilde{r}}(s_{\sigma,|\sigma|}))$$

We assume linear reward functions and use successor features to quickly estimate Q* and V* for new reward parameters.

Slide credit: W. Bradley Knox

Learning a reward function with synthetic preferences

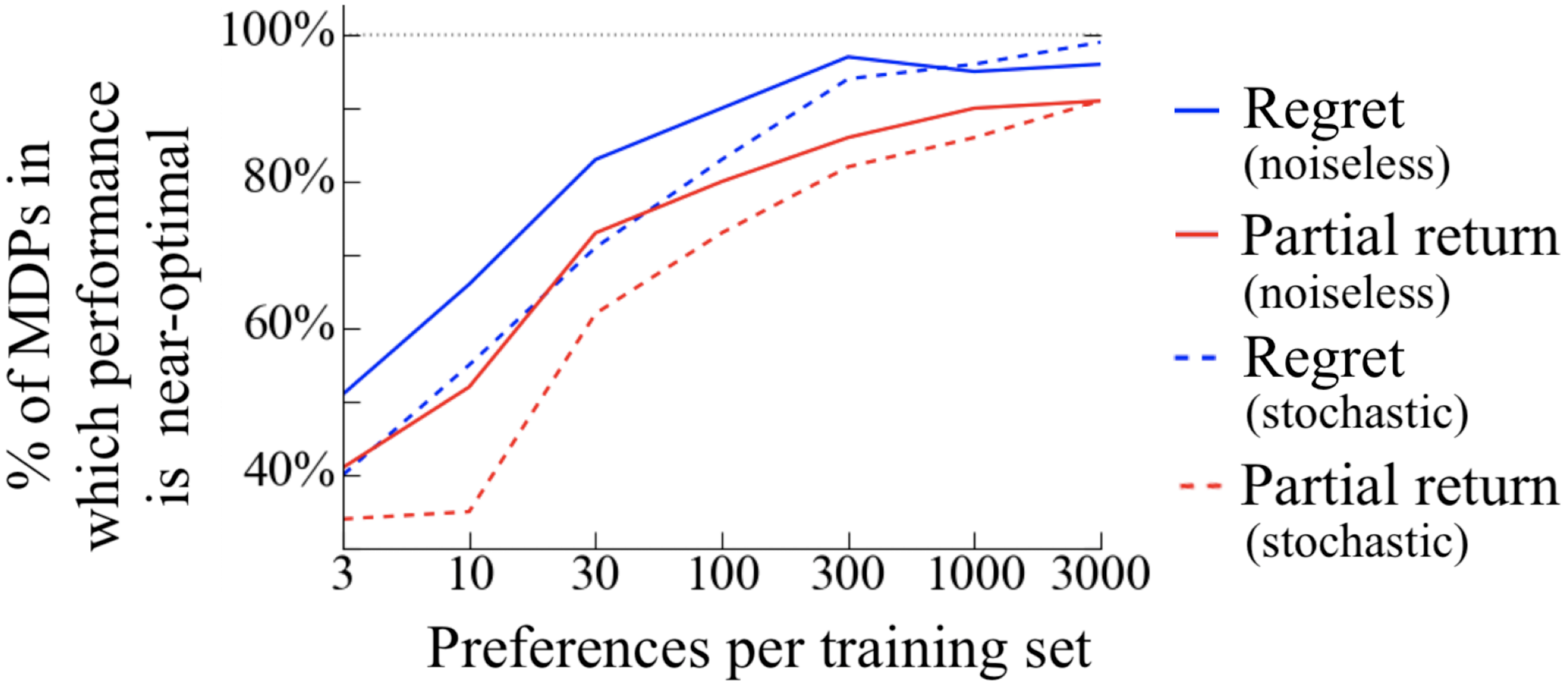
The delivery domain

Slide credit: W. Bradley Knox

Slide credit: W. Bradley Knox

100 randomly generated MDPs

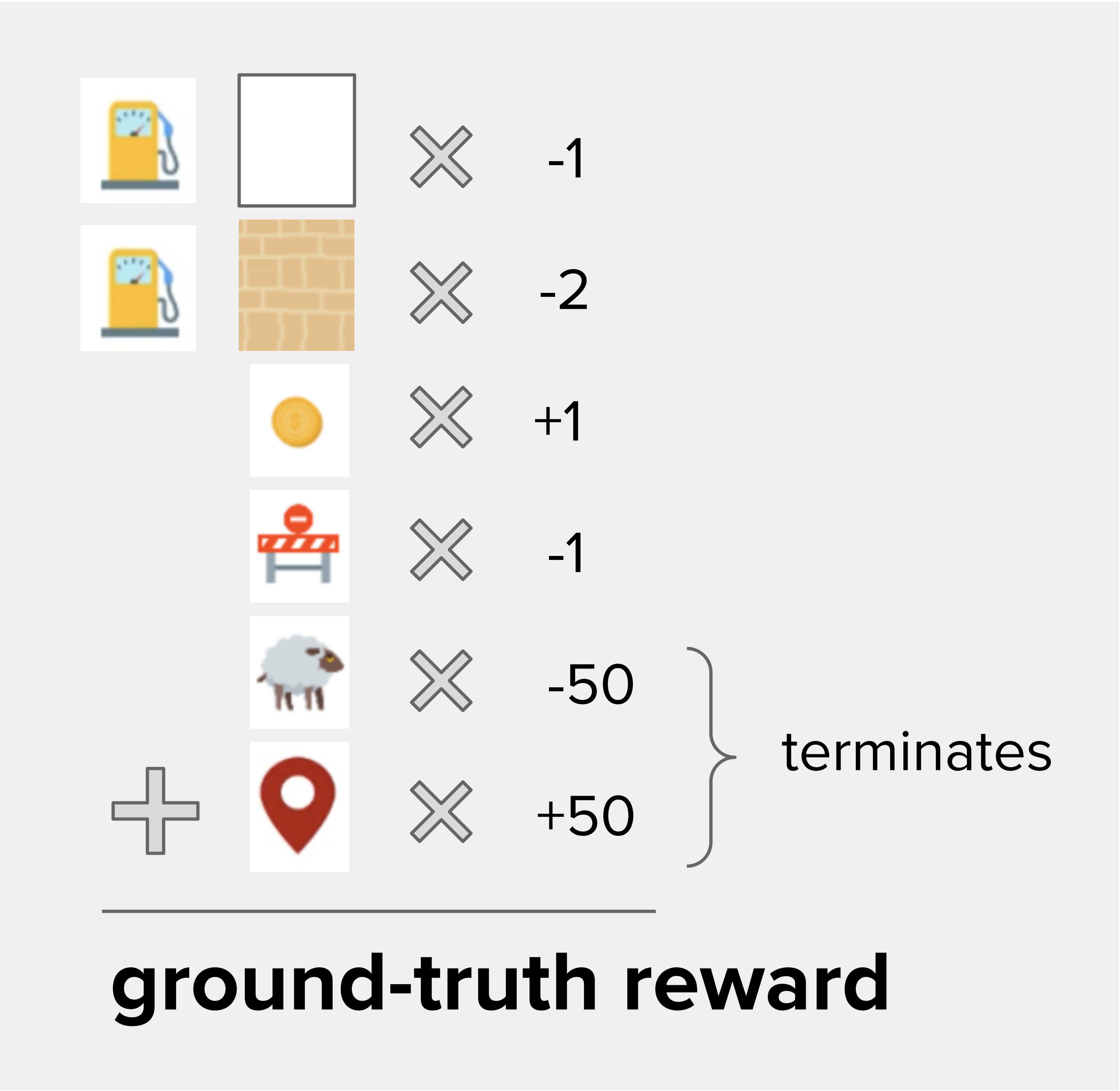
When each model is perfect, because it creates its own preference dataset



Slide credit: W. Bradley Knox

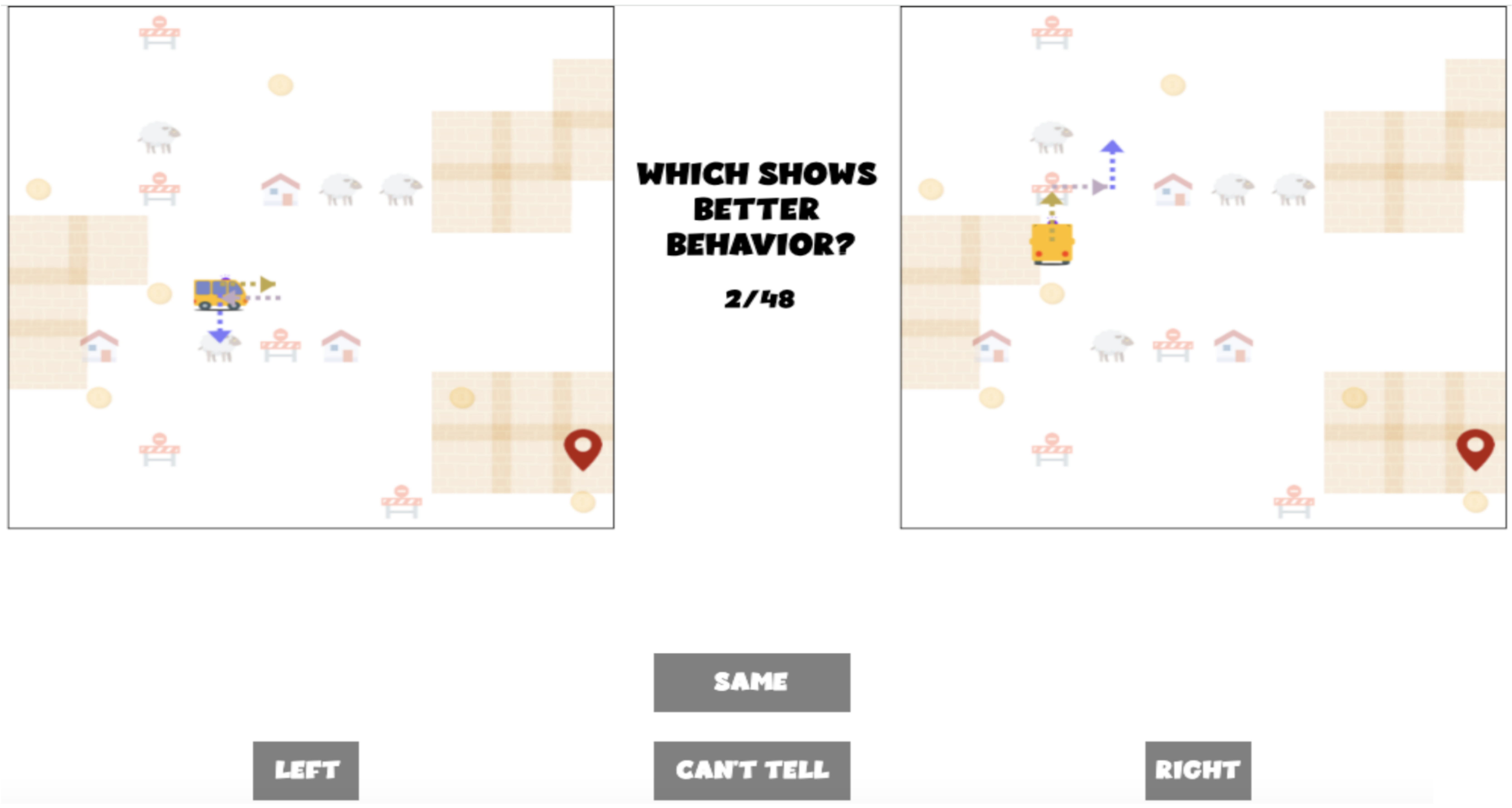
A human preference dataset

The delivery task



Slide credit: W. Bradley Knox

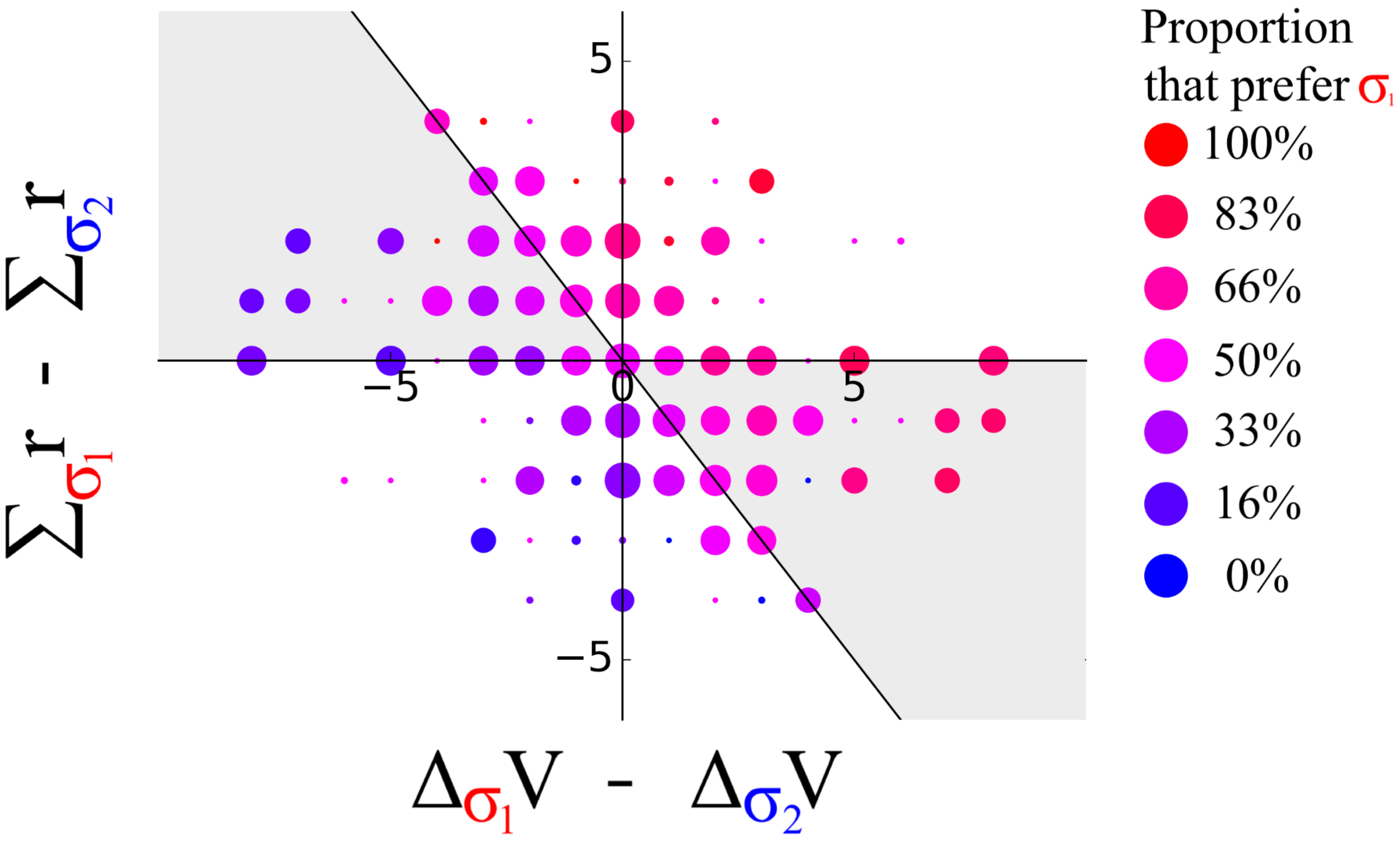
Preference elicitation



Slide credit: W. Bradley Knox

Human preferences visualized

Recall $regret_d(\sigma|\tilde{r}) \triangleq \sum_{t=0}^{|\sigma|-1} regret_d(\sigma_t|\tilde{r}) = V^*_{\tilde{r}}(s_{\sigma,0}) - (\Sigma_\sigma \tilde{r} + V^*_{\tilde{r}}(s_{\sigma,|\sigma|}))$

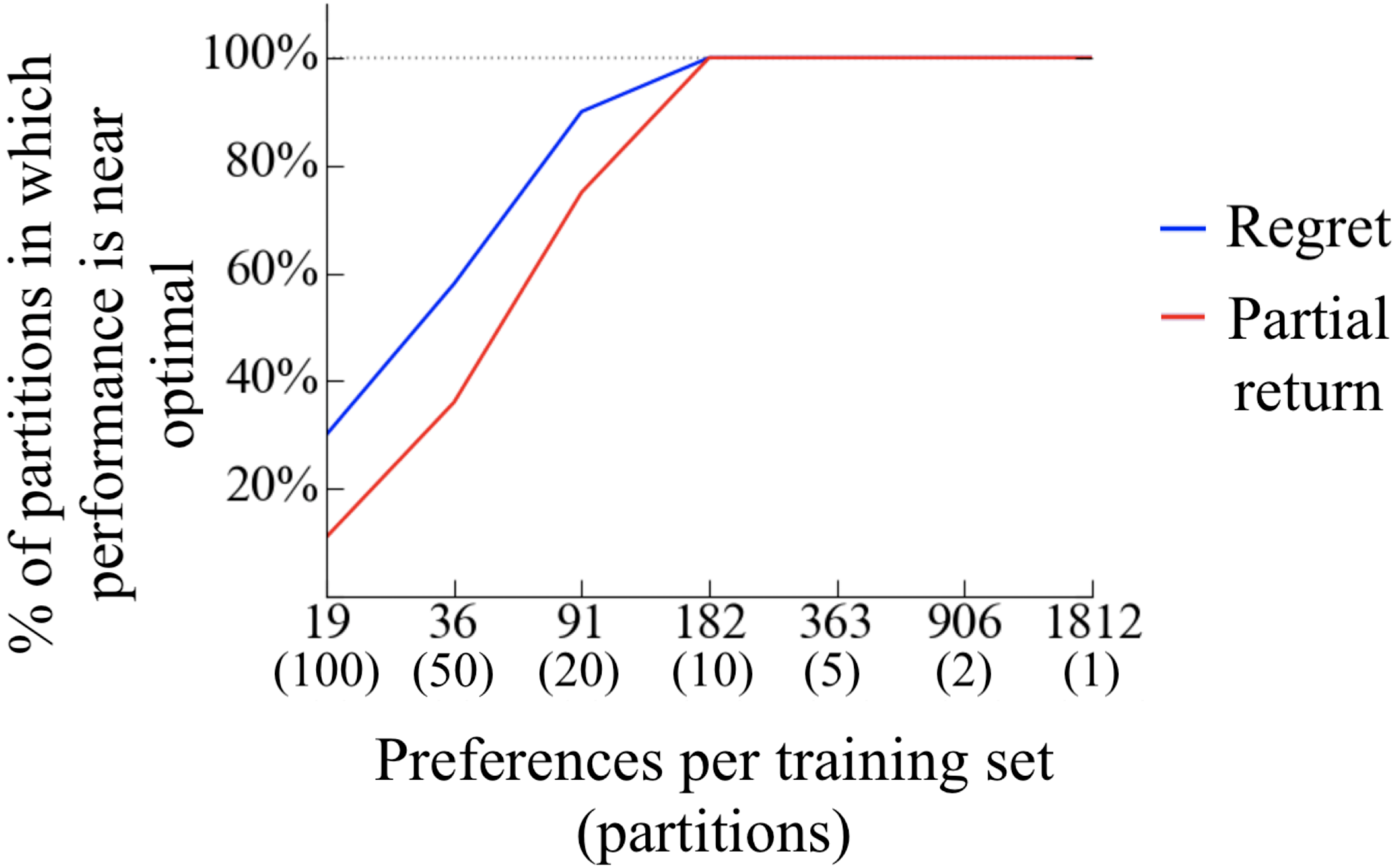


Explaining human preferences with different preference models

Preference model	Loss
$P(\cdot) = 0.5$ (uninformed)	0.69
P_{Σ_r} (partial return)	0.62
P_{regret}	**0.57**

Mean cross-entropy test loss over 10-fold cross validation (n=1812) from predicting human preferences. Lower is better.

Performance with random partitions of human preferences dataset



Slide credit: W. Bradley Knox

Benefits of the regret preference model (over the partial return model)

1. Humans intuitively appear to consider state value. The regret preference model also considers state value (in expectation).
2. Always prefers optimal segments over suboptimal segments, making it reward identifiable with noiseless preferences or stochastic preferences.
3. More sample efficient
    - when learning from its own preferences.
    - when learning from human preferences.
4. When $|\sigma| = 1$, the discount factor is considered, which is critical because the discount factor and the reward function *interact* to determine the set of optimal policies.

Slide credit: W. Bradley Knox

Choose set sensitivity

- Human shown (A,B) — prefers A to B
- Human shown (A,B,C) — prefers B to A
- Why?

Decoy effect

Laptop A

Cost: $100

Storage: 1 TB

Laptop B

Cost: $200

Storage: 2 TB

Decoy effect

Laptop A

Cost: $100

Storage: 1 TB

Laptop B

Cost: $200

Storage: 2 TB

Laptop C

Cost: $400

Storage: 3 TB

Decoy effect

Option A

Helpfulness: 5

Toxicity: 1

Option B

Helpfulness: 8

Toxicity: 4

Decoy effect

Option A

Helpfulness: 5

Toxicity: 1

Option B

Helpfulness: 8

Toxicity: 4

Option C

Helpfulness: 9

Toxicity: 8

What do preferences really mean?

…and how should we model them?