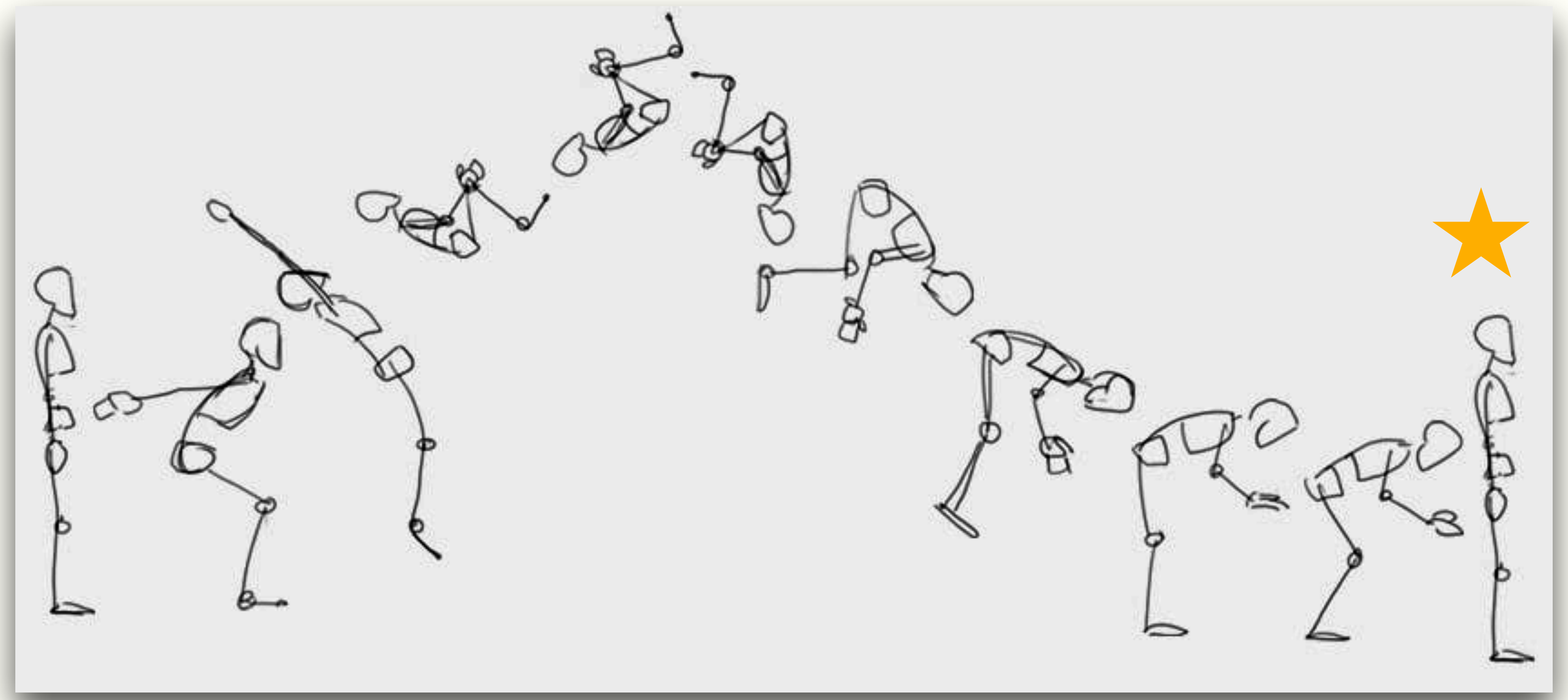
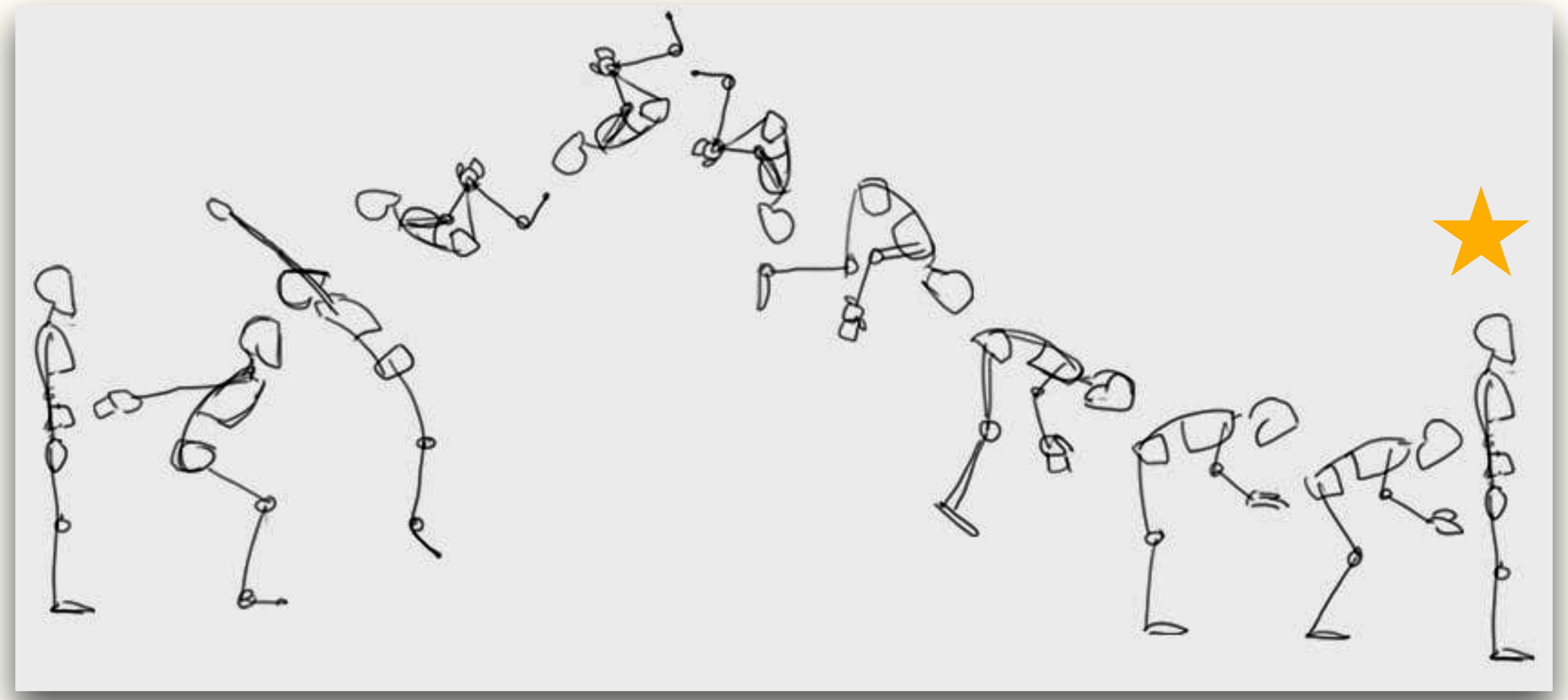
Settling the Reward Hypothesis

Michael Bowling, John D. Martin, David Abel, Will Dabney

Correspondence: jdmartin86@gmail.com





"All of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)."

— Rich Sutton and Michael Littman

On the Expressivity of Markov Reward

David Abel
DeepMind
dmabel@deepmind.com

Will Dabney
DeepMind
wdabney@deepmind.com

Anna Harutyunyan
DeepMind
harutyunyan@deepmind.com

Mark K. Ho
Department of Computer Science
Princeton University
mho@princeton.edu

Michael L. Littman
Department of Computer Science
Brown University
mlittman@cs.brown.edu

Doina Precup
DeepMind
doinap@deepmind.com

Satinder Singh
DeepMind
baveja@deepmind.com

We settle the reward hypothesis by specifying the implicit requirements on goals needed for the hypothesis to hold.

Summary of Contributions

Summary of Contributions

- Formalize the reward hypothesis as a set of formal assumptions.

Summary of Contributions

- Formalize the reward hypothesis as a set of formal assumptions.
- Specify the conditions under which the reward hypothesis holds.

Summary of Contributions

- Formalize the reward hypothesis as a set of formal assumptions.
- Specify the conditions under which the reward hypothesis holds.
- Translate results to the objective goals case.

Summary of Contributions

- Formalize the reward hypothesis as a set of formal assumptions.
- Specify the conditions under which the reward hypothesis holds.
- Translate results to the objective goals case.
- Describe an algorithm that can construct rewards for any "goal."

Summary of Contributions

- Formalize the reward hypothesis as a set of formal assumptions.
- Specify the conditions under which the reward hypothesis holds.
- Translate results to the objective goals case.
- Describe an algorithm that can construct rewards for any "goal."
- Frame results in context of common reactions to the hypothesis.

Summary of Contributions

- Formalize the reward hypothesis as a set of formal assumptions.
- Specify the conditions under which the reward hypothesis holds.
- Translate results to the objective goals case.
- Describe an algorithm that can construct rewards for any "goal."
- Frame results in context of common reactions to the hypothesis.

“All of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward).”

— Rich Sutton and Michael Littman

Assumption: Subjective Goals

"All of what we mean by goals and purposes" can be expressed as a binary preference relation* on distributions over finite histories, denoted by $\succsim$.

For $A, B \in \Delta(\mathcal{H})$

* Inspired by the work of

- Pitis (2019)
- Shakerinava and Ravanbakhsh (2022)

Assumption: Subjective Goals

"All of what we mean by goals and purposes" can be expressed as a binary preference relation* on distributions over finite histories, denoted by $\succsim$.

For $A, B \in \Delta(\mathcal{H})$ $\qquad A \succsim B$

* Inspired by the work of
- Pitis (2019)
- Shakerinava and Ravanbakhsh (2022)

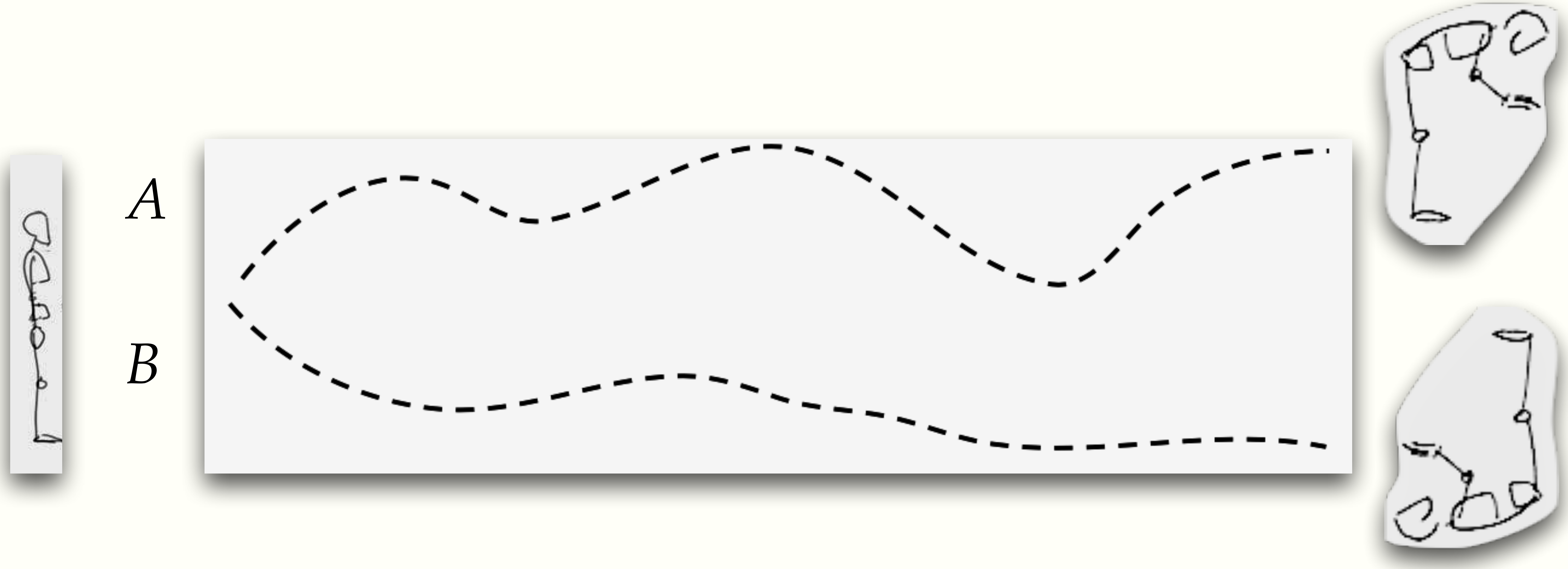
Assumption: Subjective Goals

"All of what we mean by goals and purposes" can be expressed as a binary preference relation* on distributions over finite histories, denoted by $\succsim$.

For $A, B \in \Delta(\mathcal{H})$ $\qquad A \succsim B$

* Inspired by the work of

- Pitis (2019)
- Shakerinava and Ravanbakhsh (2022)



"All of what we mean by goals and purposes can be well thought of as
maximization of the expected value of the cumulative sum of a received scalar signal (reward)."
— Rich Sutton and Michael Littman

Assumption: Cumulative Sum of Rewards

The "maximization of the expected value of the cumulative sum of a received scalar signal (reward)" means that there is a reward function and a transition-dependent discount function

$$r\colon \mathcal{O} \times \mathcal{A} \to [0,1], \qquad \gamma\colon \mathcal{O} \times \mathcal{A} \to [0,1],$$

such that we weakly prefer π_1 to π_2 under our reward if and only if there exists an N such that for all $V_n^{\pi_1} \geq V_n^{\pi_2}$ for all $n \geq N$, where

$$V_n^{\pi} \stackrel{\text{def}}{=} E\left[\sum_{i=1}^{n}\left(\prod_{j=1}^{i-1}\gamma(O_j, A_j)\right) r(O_i, A_i)\middle|\pi, e\right].$$

Generalized discounting from White, 2017.

"All of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)."

— Rich Sutton and Michael Littman

Assumption 4 (The Reward Hypothesis)

What the reward hypothesis means by "well thought of" is that for any preference relation on distributions of histories there exists r and γ such that

$$\pi_1 \succsim_g \pi_2 \iff \pi_1 \succsim_r \pi_2$$

von Neumann Morgenstern Utility Theory

A preference relation satisfies *rationality axioms* if and only if there exists a utility function consistent with the relation.

von Neumann Morgenstern Utility Theory

A preference relation satisfies *rationality axioms* if and only if there exists a utility function consistent with the relation.

Rationality Axioms

- Completeness
- Transitivity
- Independence
- Continuity

Axiom 1 (Completeness). *For all* $A, B \in \Delta(\mathcal{H})$, $A \gtrsim B$ *or* $B \gtrsim A$ *(or both, if* $A \sim B$*).*

Some judgement is needed for every pair of outcomes.

Axiom 2 (Transitivity). *For all* $A, B, C \in \Delta(\mathcal{H})$, *if* $A \gtrsim B \gtrsim C$, *then* $A \gtrsim C$.

No cyclical preferences.

Axiom 3 (Independence). *For all* $A, B, C \in \Delta(\mathcal{H})$ *and* $p \in (0, 1)$, $A \gtrsim B$ *if and only if*

$$pA + (1 - p)C \gtrsim pB + (1 - p)C$$

Mixing outcomes doesn't change anything.

Axiom 4 (Continuity). *For all* $A, B, C \in \Delta(\mathcal{H})$ *if* $A \gtrsim B \gtrsim C$, *then there exists* $p \in [0, 1]$ *such that,*

$$pA + (1 - p)C \sim B$$

Between any two outcomes is a continuum of preferences

Axiom 5: Temporal Gamma Indifference

For all $A, B \in \Delta(\mathcal{H})$ and transitions $t \in T$, *with* $\gamma=1$

$$\frac{1}{2}(t \cdot A) + \frac{1}{2}B \sim \frac{1}{2}(t \cdot B) + \frac{1}{2}A$$

Ask us questions and read the paper for more!