A Common Description Language for Human Reinforcement Learning Paradigms

Maria K. Eckstein[1], Kevin J, Miller[1], Angela Radulescu[2]

[1]Google DeepMind, [2]Icahn School of Medicine at Mount Sinai

Introduction

- **Reinforcement learning** (RL) is the dominant framework to model reward-based learning in humans and other animals
- A **variety of task paradigms** has been created; and a **variety of model variants** is used to analyze these tasks
- The literature has revealed **major differences in modeling results**, suggesting a **lack of generalizability** [Eckstein et al., 2021]

➢ ***Is there a <u>general algorithm</u> that underlies reward-based learning? How do we find it?***

➢ ***Do task differences affect reward learning? How? Do people use different strategies? Which?***

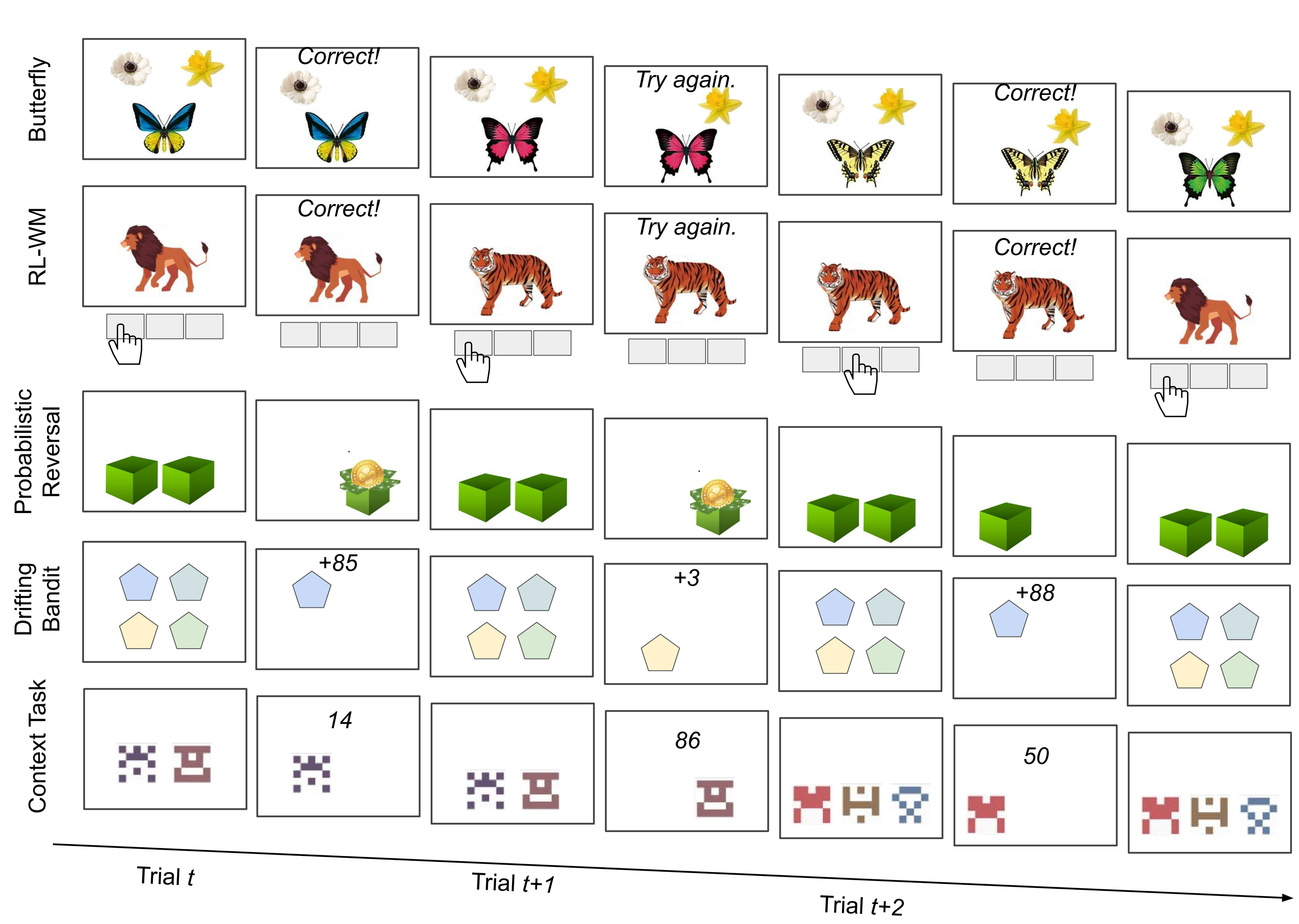
Method

- Survey the literature of existing task designs
- Identify commonalities and differences
- Create an **abstract description language** of reward-based learning tasks (identify the "axes" of the "task space")

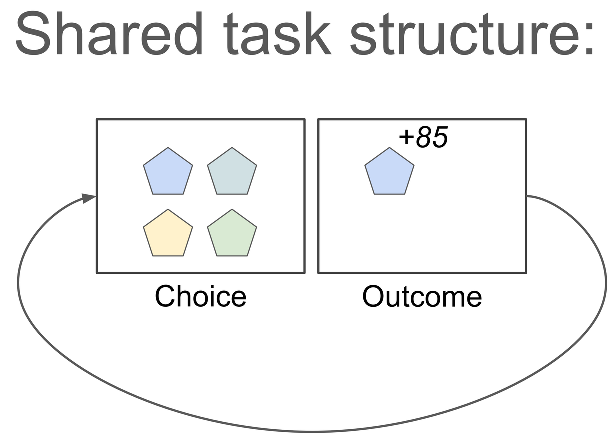
➢ This will allow us to 1) **quantify task differences** and 2) **automatically design new tasks**

Results

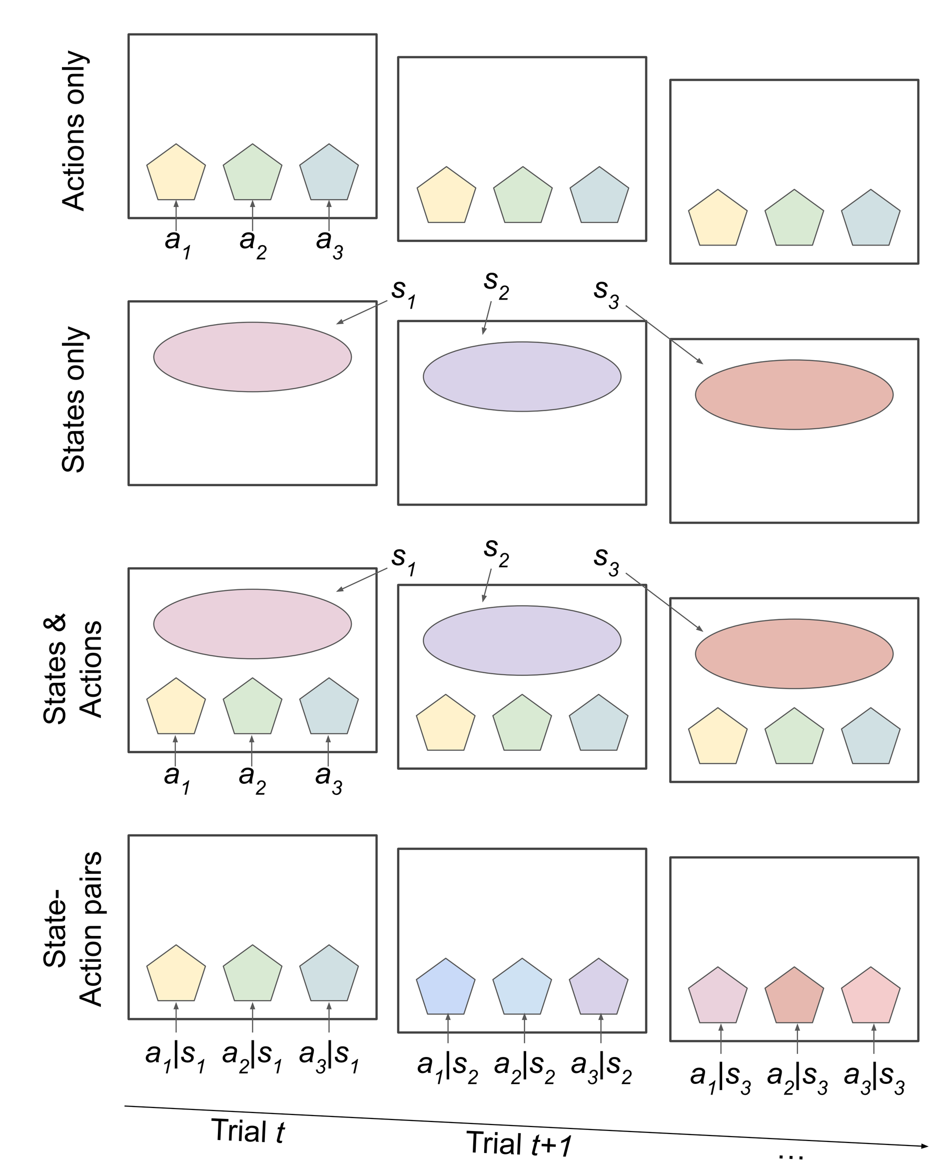
- Large literature, mostly consisting of bandit task variants:



- Shared task structure:



- Abstract description as **Markov Decision Process (MDP)**:
  - *State space* $\mathcal{S}$
  - *Action space* $\mathcal{A}$
  - *Transition function* $\mathcal{P}$: $P_a(s, s') = Pr(s_{t+1}=s'|s_t=s, a_t=a)$
  - *Reward function* $\mathcal{R}$: $R_a(s, s')$ is the immediate reward received after transitioning from state s to state s' due to action a
- **Problem 1**: Mapping is ambiguous (What are the states What are the actions? What is shown on the screen?)



- **Logic**: Each visual token represents one action. Underlying task state is not shown.
- **Examples**: Probabilistic Switching tasks [Hampton et al., 2006; Tai et al., 2012; Ouden et al., 2013; Eckstein et al., 2022], 4-armed drifting bandit task [Daw et al., 2006], volatility task [Behrens et al., 2007]

- **Logic**: Each visual token represents one task state. Which actions are available is not shown.
- **Examples**: Reinforcement Learning and Working Memory (RL-WM) Task [Collins & Frank, 2012], reward-based learning [Frank et al., 2015], task set learning [Collins & Koechlin, 2012; Donoso et al., 2014]

- **Logic**: Separate visual tokens represent the current state (top) and available actions (bottom).
- **Examples**: Butterfly task [Davidow et al., 2016]

- **Logic**: Each visual token represents a specific state-action pair.
- **Examples**: Context-based learning task [Frank et al., 2004; Lefebvre at al., 2017; Klein et al., 2017; Bavard & Palminteri, 2023]

- **Problem 2**: Many features are not included (common features: switching / drifting; binary / continuous; number of bandit arms; bandit features; prediction task)

➢ ***Many common task features are not captured by MDP framework.***

- **Alternative Proposal**: 7 features
  - *Visibility type* $\mathcal{V}$: states, actions, states+actions, states*actions
  - *Number of states* $|\mathcal{S}|$: 1, 2, …, n_s
  - *Number of actions* $|\mathcal{A}|$: 1, 2, …, n_a
  - *Outcome type* $\sim\mathcal{R}$: binary / continuous; stable / drifting
  - *Probability type* $\sim\mathcal{P}$: binary / continuous; stable / drifting
  - *Relation type* $\mathcal{C}$: identical, antisymmetric, independent
  - *Block change type* $\mathcal{B}$: "high" / "low"; feature

Conclusion

- Concise description language
- Captures majority of existing tasks
  - Open question: What about others? Exponential explosion of task space with every new feature
- Reveals that current literature exploits minuscule regions of task space, while vast regions are unexplored
- Reveals how similar existing tasks are to each other
- Allows interpolation between existing tasks and creation of entirely new ones
- Allows uniform sampling from underlying task space