Robotics and AI Group
School of ECE
University of Tehran

Social Learning Group

Erfan Mirzaei
Amirhossein Mesbah
Banafshe Karimian

Dec 2021

Social Reinforcement Learning

PLOS BIOLOGY

RESEARCH ARTICLE

The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning

Anis Najar[1,2,3]*, **Emmanuelle Bonnet**[1,2,3], **Bahador Bahrami**[4,5,6], **Stefano Palminteri**[1,2,3]*

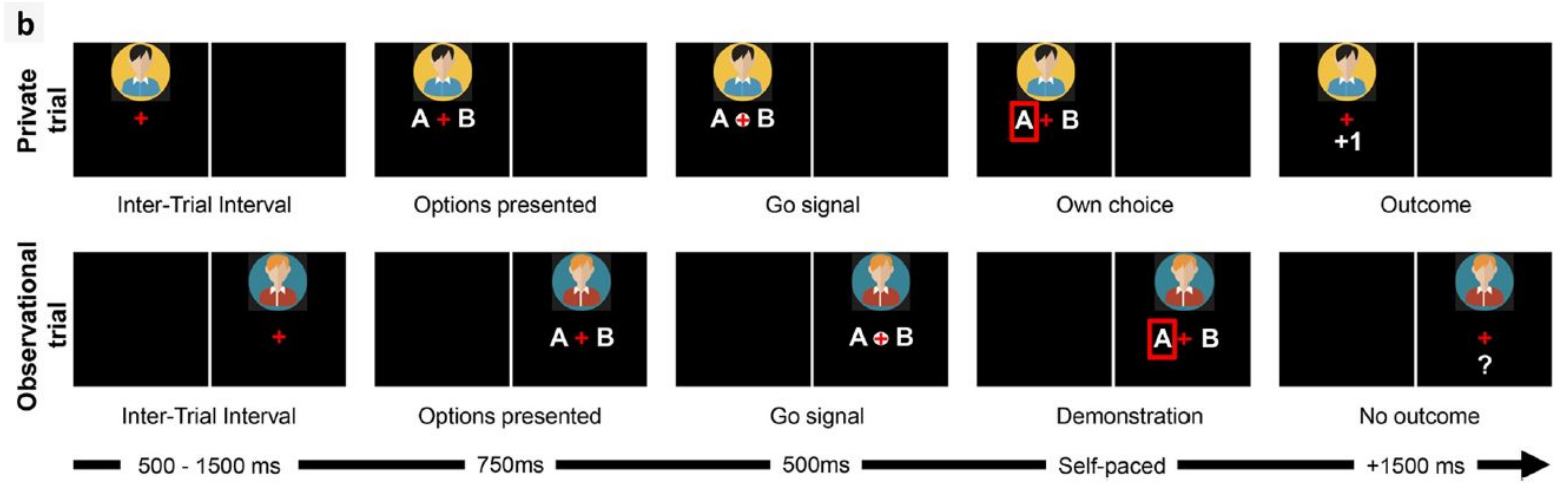
1 Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et de la Recherche Médicale, Paris, France, **2** Département d'Études Cognitives, École Normale Supérieure, Paris, France, **3** Human Reinforcement Learning team, Université de Paris Sciences et Lettres, Paris, France, **4** Ludwig-Maximilians Universität München, Faculty of Psychology and Educational Sciences, General and Experimental Psychology, Munich, Germany, **5** Department of Psychology, Royal Holloway University of London, London United Kingdom, **6** Max Planck Institute for Human Development, Center for Adaptive Rationality, Berlin, Germany

* anis.najar@ens.fr (AN); stefano.palminteri@ens.fr (SP)

Published: December 8, 2020

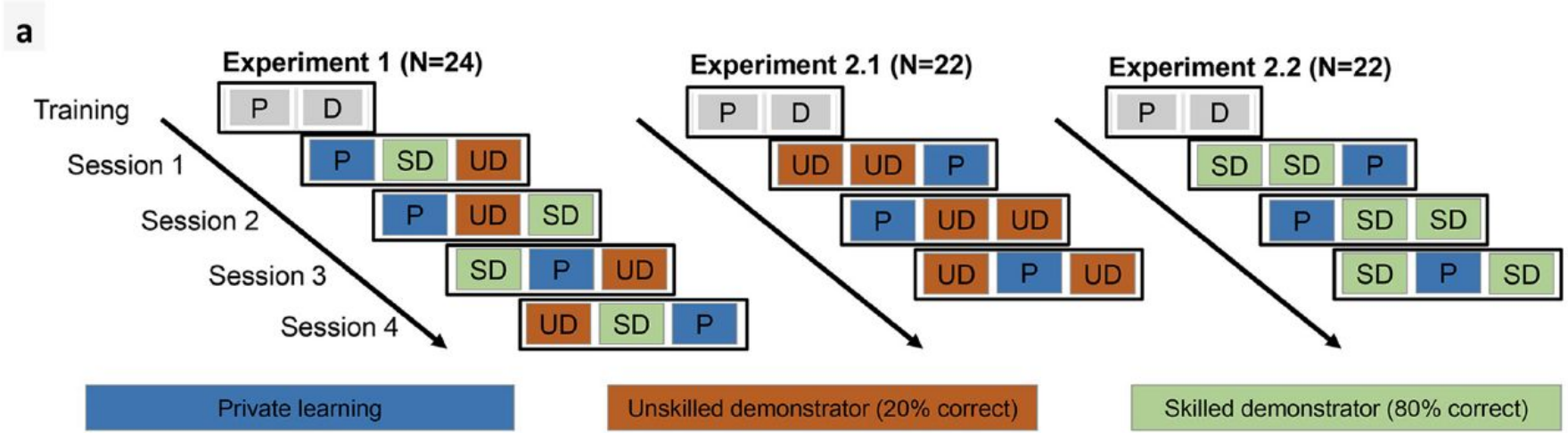
Motivation and Main Problem

- There is no doubt that social signals affect human reinforcement learning, BUT still no consensus about how implement it computationally



Social Reinforcement Learning task
(probabilistic instrumental learning task)

Motivation and Main Problem



Notes:

- 3 Conditions : 1) PL 2) UD 3)SD ----> Randomized at each session(20 SD, 20 UD, 20 PL)
- Demonstrations consist two trials : Private and Observational, But PL just has private trial
- Rewards: opposite and reciprocal winning and losing probabilities
- Goal : learn by trial-and error which of the 2 stimuli had the highest expected value.

Motivation and Main Problem

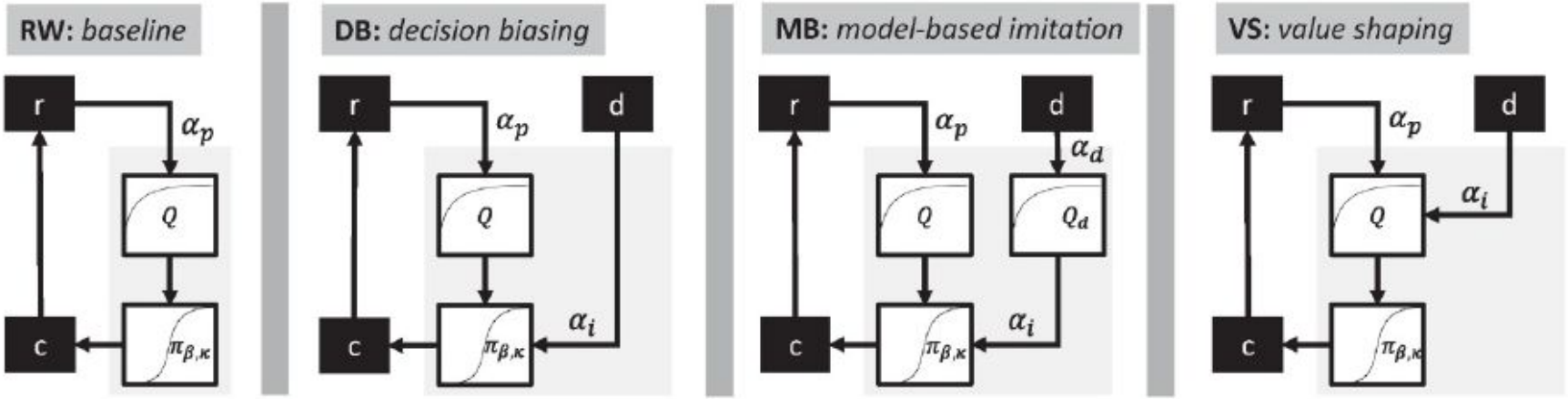
- Individual learning (classic RL) may be slow, costly, risky and also information-inefficient, so with social rl, we can facilitate learning of artificial agents and mitigate the costs and risks.

- The main question of this research is about how observing other agent's actions, can affect on our learning

- Imitation as one of main mechanism of social learning has been widely investigated in psychology and neuroscience, but they are still limited in their scope as imitation is treated in isolation from other learning processes, such as autonomous reinforcement learning.

Contributions

3 psychologically plausible hypotheses for modeling social RL(Imitation)



Imitation transiently biasing the learner's action selection without affecting their value functions

the learner infers the demonstrator's value function through inverse RL and uses it to bias action selection.

the demonstrator's actions directly affect the learner's value function

Related works

Decision Biasing:

- *[1] Burke CJ, Tobler PN, Baddeley M,* ***Schultz*** *W. Neural mechanisms of observational learning*
- *[2] Selbing I,* ***Lindstrom*** *B,* ***Olsson*** *A. Demonstrator skill modulates observational aversive learning*

One limitation : Doesn't allow to an extended effect of imitation over time **(Imitation as biasing exploration strategy)** but studies show that imitation has long lasting effect (*Bandura A. Social cognitive theory.*)

Model-based imitation:

- *[3] Baker CL, Saxe R,* ***Tenenbaum*** *JB. Action understanding as inverse planning.*
- *[4] Collette S, Pauli WM, Bossaerts P,* ***O'Doherty*** *J. Neural computations underlying inverse reinforcement learning in the human brain.*

Note: In previous paradigms, the model of the Demonstrator is generally used to predict her behavior, but these representations could easily be recycled to influence the behavior of the Learner.
Limitation: This approach is computationally demanding.Also there is unanswered questions of how the model of the Demonstrator is integrated into the behavior of the Learner

Related works

Value Shaping:

- *Advice taking*
  [5] Biele G, Rieskamp J, Krugel LK, Heekeren HR. The neural basis of following advice.

- *Learning from evaluative feedback*
  *[6] Ho MK, MacGlashan J, **Littman ML**, Cushman F. Social is special: A normative framework for teaching with and learning from evaluative feedback.*

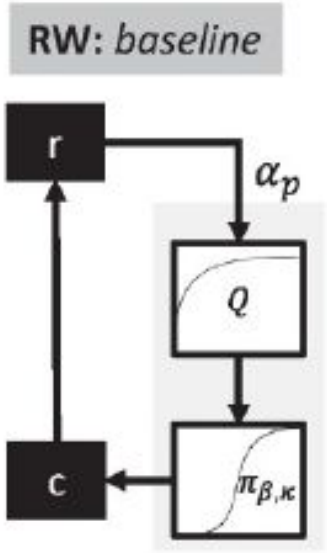
- *Human RL*
  [7] Najar A, Chetouani M. Reinforcement learning with human advice: a survey

One Note : This scheme allowing for a long-lasting influence of social signals while being computationally simple.

Models

Baseline Model: [RW1]

Initial Values Q = 0



$$Q(c) \leftarrow Q(c) + \alpha_p \times [r - Q(c)],$$

Choice

Private learning rate

Exploration temperature

Autocorrelation parameter

$$\pi(c) = \frac{1}{1 + e^{\beta \times [Q(\bar{c}) - Q(c) - \lambda(c).\kappa]}}$$

$$\lambda(c) = \begin{cases} 1 & \text{if c is the last performed action} \\ -1 & \text{otherwise} \end{cases}$$

Symmetric update value: [RW2]

$$Q(\bar{c}) \leftarrow Q(\bar{c}) + \alpha_p \times [-r - Q(\bar{c})].$$

Models

Decision Biasing Model: [DB1]

First the policy is derived from the value function

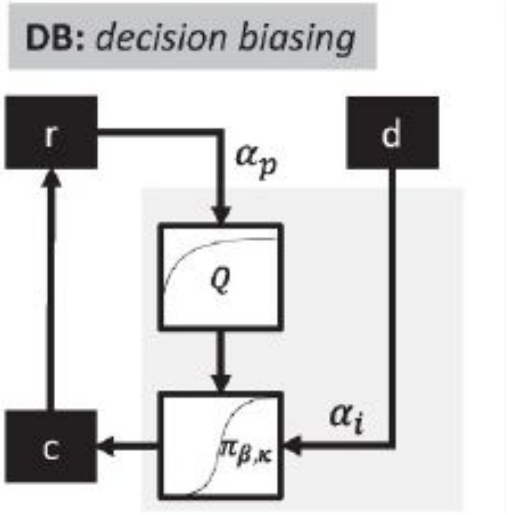
The Demonstrator's choice is perceived as a greedy action

$$\pi(d) \leftarrow \pi(d) + \alpha_i \times [1 - \pi(d)],$$

Demonstrated action

Imitation decision bias rate

$$\pi(\bar{d}) \leftarrow 1 - \pi(d),$$



Models

Decision Biasing Models:

DB6, in which the action prediction error is computed based on the Q-value (i.e., a value update).

$$Q'(d) \leftarrow Q'(d) + \alpha_i \times [1 - Q'(d)].$$

$$Q'(d) \leftarrow Q'(d) + \alpha_i \times [1 - Q'(d)],$$

$$Q'(\bar{d}) \leftarrow Q'(\bar{d}) + \alpha_i \times [-1 - Q'(\bar{d})].$$

Symmetric update value

The policy is derived from Q' instead of Q.

Note: Because the policy π (resp. Q0) is derived from Q each time a demonstration is provided,they do not allow the accumulation of successive demonstrations.

Models

Model-Based Imitation Model: [MB9] **Symmetric update value**

Observational trial

$$Q_D(d) \leftarrow Q_D(d) + \alpha_d[1 - Q_D(d)]$$

$$Q_D(\overline{d}) \leftarrow Q_D(\overline{d}) + \alpha_d\left[-1 - Q_D(\overline{d})\right]$$

$$d^* \leftarrow argmax(Q_D)$$

Private trial

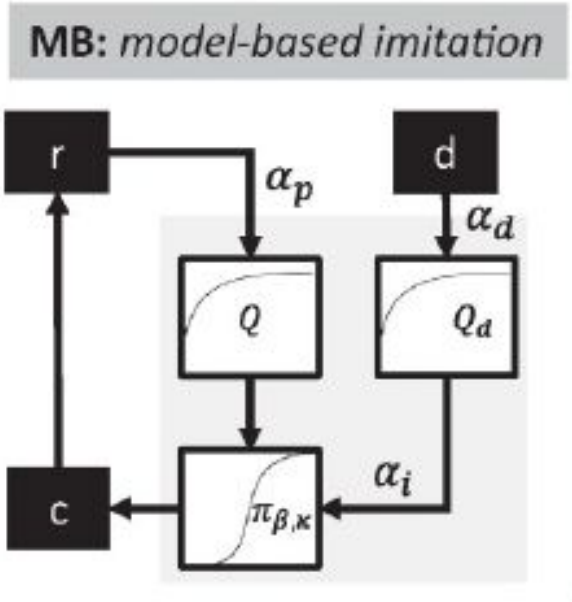
$$Q' \leftarrow Q$$

$$Q'(d^*) \leftarrow Q'(d^*) + \alpha_i[1 - Q'(d^*)]$$

$$Q'(\overline{d^*}) \leftarrow Q'(\overline{d^*}) + \alpha_i\left[-1 - Q'(\overline{d^*})\right]$$

$$\pi \leftarrow softmax_{\beta,\kappa}(Q')$$

The Demonstrator's most valuable action perceived as a greedy action



Note: The modeling can be done by policy shaping just like the previous model.

Models

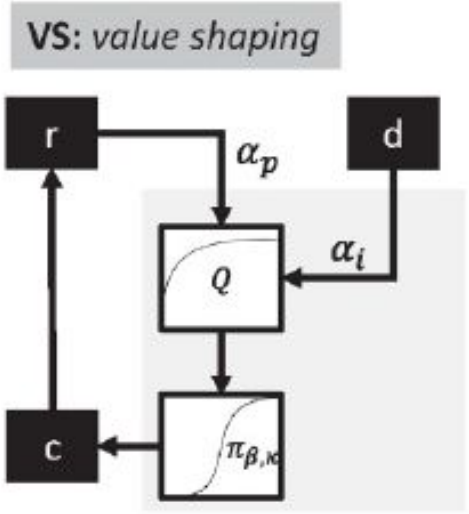
The Demonstrator's choice is perceived as a positive outcome (or **surrogate** reward).

Value Shaping Model: [VS1]

$$Q(d) \leftarrow Q(d) + \alpha_i \times [1 - Q(d)],$$

Demonstrated action

Imitation learning rate



Symmetric update value: [VS2]

$$Q(\bar{d}) \leftarrow Q(\bar{d}) + \alpha_i \times [-1 - Q(\bar{d})],$$

Models

Meta learning

Auxiliary learning rate

$$\alpha_i(s) \leftarrow \alpha_i(s) + \alpha_m * (\tau - \alpha_i(s)),$$

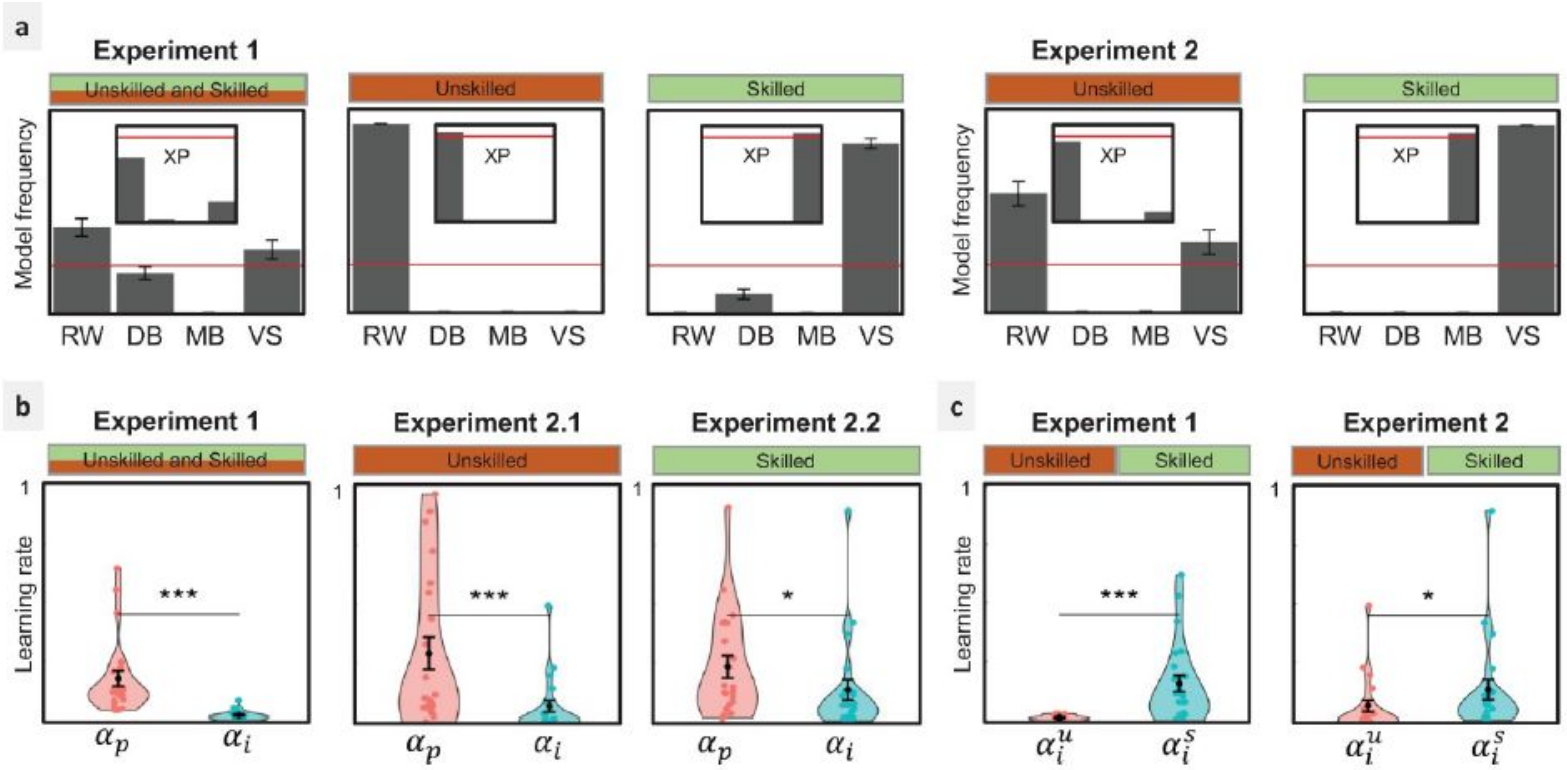
Initial Values alpha_i = 0

Imitation learning rate

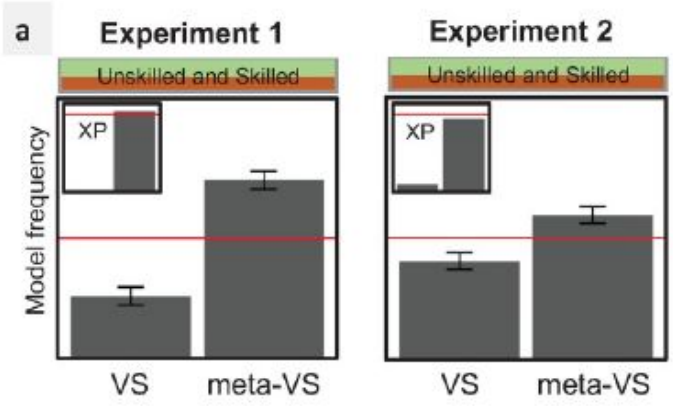
$$\tau = \begin{cases} 1 & \text{if } Q(d) = max(Q(d), Q(\bar{d})) \\ 0 & \text{otherwise} \end{cases}$$

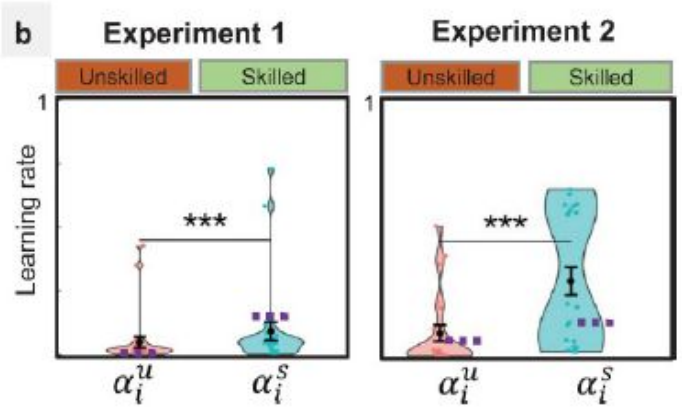
Note: Since the Demonstrator's choice outcomes are not directly accessible to the Learner, we postulated that the participant uses their own value function to assess the Demonstrator's skill on the task.

Experimental Results



Experimental Results





Results

- Imitation takes form of Value Shaping, which implies that the choices of the Demonstrator affect the value function of the Learner.
- On top of that, we found that imitation is modulated by a meta-learning process, such that it occurs when it is adaptive (i.e., SD).
- In other terms, imitation is instantiated as a model-free learning process.
- A comparison between reward and imitation learning rates suggests that privately generated outcomes are considered overweighted compared to social information.
- We found that humans can correctly infer the skill of a Demonstrator and modulate their imitation accordingly
- For imitation to be adaptive, it should be modulated by several environmental and social signals. Here, we focused on the Demonstrator's skills and found that when the Demonstrator was not skilled, the imitation learning rate was down-regulated.

Discussion

- **Value Shaping vs. Decision Biasing**
  - A notable advantage of VS is that it can easily account for observational learning in Pavlovian settings where no decisions are involved, while DB cannot.
  - Their VS method is equivalent to the outcome bonus method that has been proposed for integrating advice into reinforcement learning *[5].* Just like reward shaping in RL,the advised options are perceived more positively and thus acquire an extra reward bonus. which corresponds to augmenting the reward function with extra rewards in order to speed up the learning process. However, it can lead to suboptimal solutions that fail to account for human behavior[6].
  - 2 approach to address this problem:
    - Value Shaping: affects the preference for “advised” actions without modifying the Learner’s reward specifications
    - Policy Shaping: policy shaping which affects the Learner’s behavior without modifying its value function with long-lasting effect
  - **Nevertheless** in this single-state task there is no distinction between reward shaping, policy shaping and value shaping.(adding an extra reward bonus to an action = augmenting its expected value = augmenting its probability of being selected) [Further research need to assess which one is better]

Discussion

- **Value Shaping vs. Model-based imitation**
  - Distinction between VS and MB echoes the classical distinction between model-free and model-based methods in the RL literature.
  - The results may seem in contrast with previously reported results showing that people infer a model of a Demonstrator, but it can be explained by the fact that in these works,participants were explicitly instructed to predict other participant's behavior.
  - Moreover, in these previous works, demonstrations were the only available source of information for the Learner, while in our task, demonstrations were in competition with self-generated outcomes.
  - When only demonstrations are provided inferring the Demonstrator's goal is the only way for learning the task demonstrations only play a secondary role.
  - MB implies that people build distinct representations of the same task, one of which can be a representation of a Demonstrator's goal. Each representation can then influence the others, be switched on and off, and be more or less considered by the Learner.VS, on the other hand, implies a deeper effect of imitation. People would integrate others' behavior and adopt their preferences as their own. This "subversive" effect of imitation can be found in other works [28], where over-imitation is explained not as a mimicry mechanism, but by the modification of the inner representation of the causal model of the environment.

Limitations

- Used models just for binary action selection in bandits.
- There is just one demonstrator in the environment.
- The rewards was assumed to be just 1 or -1.
- They postulated the participants have the same utility functions.
- In first 3 hypotheses, they postulated that the demonstrator is an expert
- In meta VS, they assume participants uses their own value function to assess the demonstrator's skill on the task [Ali Naghdi's work], that would be dangerous in complex tasks.

Our Inspiration

I think the main differences of these models with our work is that they update their policy, value, etc just by observing an action, but in our methods we update just by do actions, which this has some positive and negative consequences. In fact in this work there are 2 types of trial: private and Observational, but in our work all trials are mixed.

- (meta)VS is compatible with pavlovian settings,where no decisions(actions) are involved, while DB and BK's cannot. We can compare our results with this
- Implement policy shaping too.
- Decrease preference of other actions of the most prefered agent without doing them