Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Bonn-Aachen
International Center for
Information Technology

Interactive Reinforcement Learning

Guest Lecture

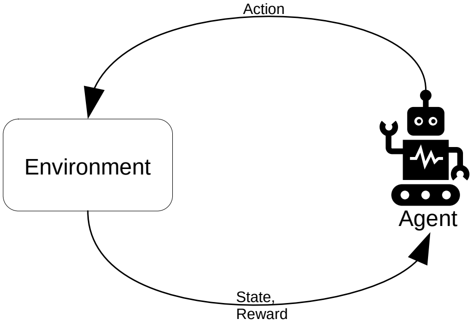
June 13th, 2024

Michał Stolarz
Prof. Dr. Teena Chakkalayil Hassan

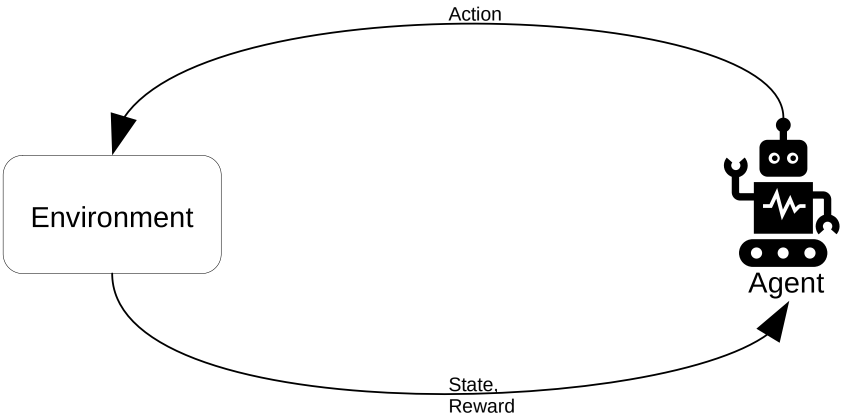
1. What is Interactive Reinforcement Learning?
2. Teaching signals used in IRL
3. Human evaluative feedback
4. Learning from evaluative feedback
5. TAMER
6. SABL
7. COACH
8. Combining evaluative feedback and environmental reward
9. Real-life application
10. Summary

Reinforcement Learning (RL)

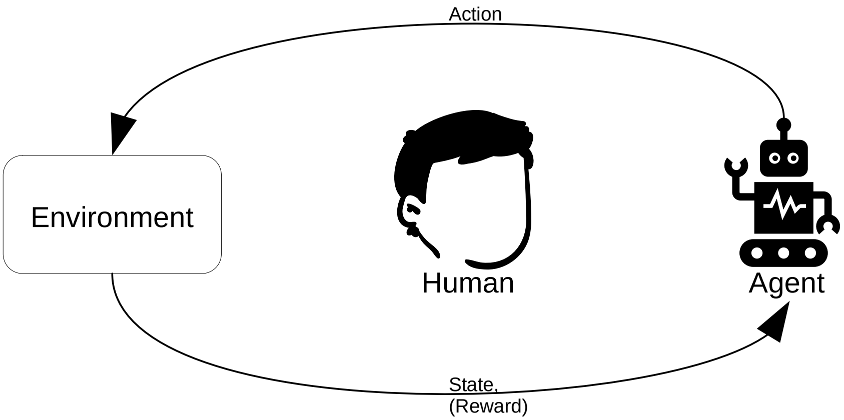
- RL allows an agent to learn an optimal policy that maximises the **cumulative reward** by interacting with an environment.
- In real-world tasks, the agent is faced with the **sample efficiency problem**, making the learning slow.



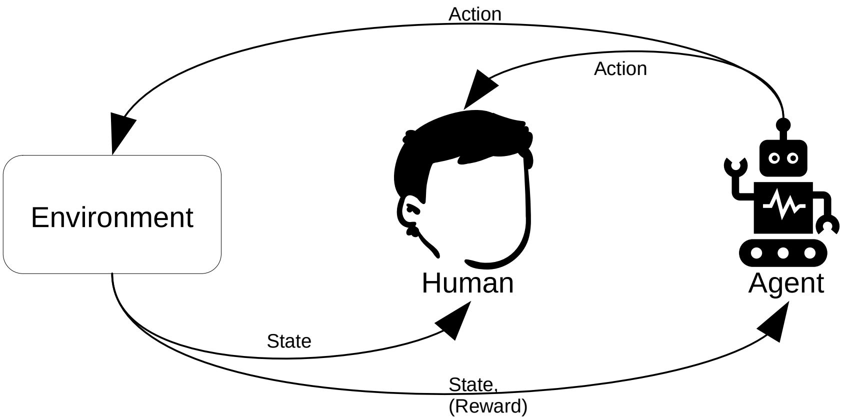
Reinforcement Learning



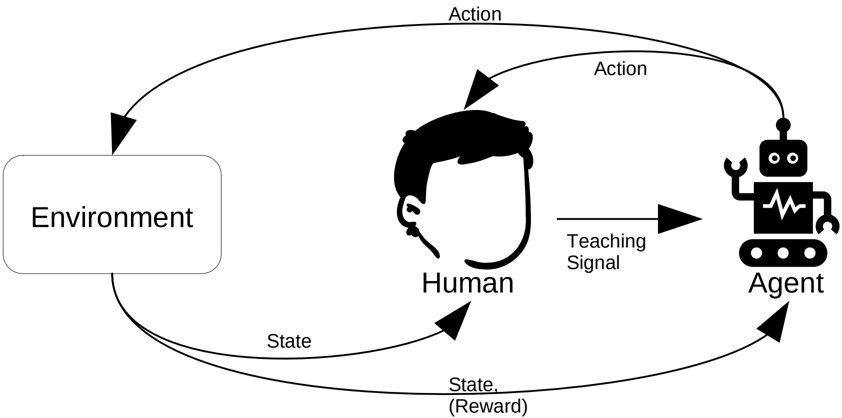
(Interactive) Reinforcement Learning



(Interactive) Reinforcement Learning



Interactive Reinforcement Learning (IRL)



Advantages of IRL

- Humans possess **knowledge** about the environment and have **experience** in acting in that environment.

- Human input could be used to guide and **accelerate the robot's learning**, or to change its optimal behaviour (personalisation).

Teaching Signals

Feedback

✓ Human feedback is provided with the intention of **evaluating** the robot's action.

✓ The value of the **evaluative feedback** depends on the last action performed by the robot.

Demonstration

✓ Demonstration is produced with the intention of **showing** a **state-action sequence** to robot.

✗ Teaching with a demonstration strategy imposes a **significant burden** on the human teacher.

Instruction

✓ An instruction is produced with the intention of **communicating** the **action** to be performed in a given task state.

✓ Learning from instructions - mapping instructions (e.g. natural language) to a sequence of executable actions.

[1] M. Chetouani, "Interactive Robot Learning: An Overview," ECCAI Advanced Course on Artificial Intelligence, pp. 140–172, 2021

Teaching Signals

Teaching signals		Feedback	Demonstration	Instruction
Nature	Notation	$H(s, a)$	$D = \{(s_t, a_t^*), (s_{t+1}, a_{t+1}^*)\ldots\}$	$I_\pi(s) = a_t^*$
	Value	Binary/Scalar	State-Action pairs	Probability of an action
Time-step	t − 1		✓	✓
	t		✓	
	$t+1$	✓		
Human	Intention	Evaluating/Correcting	Showing	Telling
	Teaching cost	Low	High	Medium
Robot	Interpretation	State-Action evaluation Reward-/Value-like	Optimal actions Policy-like	Optimal action Policy-like
	Learning cost	High	Low	High

[1] M. Chetouani, "Interactive Robot Learning: An Overview," **ECCAI Advanced Course on Artificial Intelligence**, pp. 140–172, 2021

Teaching Signals

Teaching signals		Feedback	Demonstration	Instruction
Nature	Notation	$H(s,a)$	$D = \{(s_t, a_t^*), (s_{t+1}, a_{t+1}^*)\ldots\}$	$I_\pi(s) = a_t^*$
	Value	Binary/Scalar	State-Action pairs	Probability of an action
Time-step	t − 1		✓	✓
	t		✓	
	$t+1$	✓		
Human	Intention	Evaluating/Correcting	Showing	Telling
	Teaching cost	Low	High	Medium
Robot	Interpretation	State-Action evaluation Reward-/Value-like	Optimal actions Policy-like	Optimal action Policy-like
	Learning cost	High	Low	High

[1] M. Chetouani, "Interactive Robot Learning: An Overview," **ECCAI Advanced Course on Artificial Intelligence**, pp. 140–172, 2021

Teaching Signals

Teaching signals		Feedback	Demonstration	Instruction
Nature	Notation	$H(s, a)$	$D = \{(s_t, a_t^*), (s_{t+1}, a_{t+1}^*) \ldots\}$	$I_\pi(s) = a_t^*$
	Value	Binary/Scalar	State-Action pairs	Probability of an action
Time-step	t − 1		✓	✓
	t		✓	
	$t + 1$	✓		
Human	Intention	Evaluating/Correcting	Showing	Telling
	Teaching cost	Low	High	Medium
Robot	Interpretation	State-Action evaluation Reward-/Value-like	Optimal actions Policy-like	Optimal action Policy-like
	Learning cost	High	Low	High

[1] M. Chetouani, "Interactive Robot Learning: An Overview," **ECCAI Advanced Course on Artificial Intelligence**, pp. 140–172, 2021

Teaching Signals

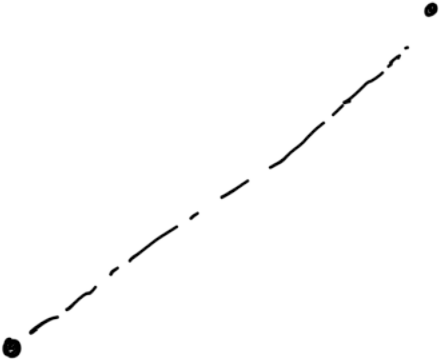
Teaching signals		Feedback	Demonstration	Instruction
Nature	Notation	$H(s,a)$	$D = \{(s_t, a_t^*), (s_{t+1}, a_{t+1}^*) \ldots\}$	$I_\pi(s) = a_t^*$
	Value	Binary/Scalar	State-Action pairs	Probability of an action
Time-step	t − 1		✓	✓
	t		✓	
	$t+1$	✓		
Human	Intention	Evaluating/Correcting	Showing	Telling
	Teaching cost	Low	High	Medium
Robot	Interpretation	State-Action evaluation Reward-/Value-like	Optimal actions Policy-like	Optimal action Policy-like
	Learning cost	High	Low	High

[1] M. Chetouani, "Interactive Robot Learning: An Overview," ECCAI Advanced Course on Artificial Intelligence, pp. 140–172, 2021

How Can Humans Deliver Evaluative Feedback?

Two **ways** of delivering evaluative **feedback**:

- Hardware
- Natural interaction modalities



How Can Humans Deliver Evaluative Feedback?

Two **ways** of delivering evaluative **feedback**:

- Hardware
- Natural interaction modalities

Two **types** of evaluative **feedback**:

How Can Humans Deliver Evaluative Feedback?

Two **ways** of delivering evaluative **feedback**:

- Hardware
- Natural interaction modalities

Two **types** of evaluative **feedback**:

- **Explicit** - user gives **direct feedback** to the robot e.g. through the graphical interface.

How Can Humans Deliver Evaluative Feedback?

Two **ways** of delivering evaluative **feedback**:

- Hardware
- Natural interaction modalities

Two **types** of evaluative **feedback**:

- **Explicit** - user gives **direct feedback** to the robot e.g. through the graphical interface.
- **Implicit** - **spontaneous behaviour** of the human user is analysed and used as feedback. Feedback is estimated based on social signals such as valence, engagement, facial expressions.

Arity of Human Evaluative Feedback

- Unimodal → Only one feedback modality is used.
- Multimodal:
  - Multiple modalities are used either **disjunctively** (OR) or **conjunctively** (AND).
  - If **one** modality is unavailable, then the others can be used → **robustness**.
  - If **all** are available, then the feedback is more **reliable**.
  - Examples:
    - Speech and gesture
    - Laugh (audio) and smile (visual)
    - Facial expressions of emotions + task-related features
- Human evaluative feedback can also be combined with a pre-defined environmental reward function.

Challenges with Human Feedback

- Delayed feedback
  - E.g. Due to reaction times involved in evaluating the action and providing the feedback.
  - To which action should the feedback be mapped?
- Interpersonal variability
  - E.g. The reaction times differ from person to person.
  - The social signals used for feedback vary in modality, in expression, and in intensity.
  - The same teacher might change the feedback strategy or type over time (e.g. change from binary to categorical feedback).

Challenges with Human Feedback

- Decay in feedback frequency
  - Intense feedback at the beginning, sparse feedback later.
- Multiple feedback
  - Should the feedback be aggregated or should only one of those multiple feedback be used?
- Noise in feedback channel
  - Discrepancies between what the teacher intends to convey and what the agent actually observes.

1. What is Interactive Reinforcement Learning?
2. Teaching signals used in IRL
3. Human evaluative feedback
4. Learning from evaluative feedback
5. TAMER
6. SABL
7. COACH
8. Combining evaluative feedback and environmental reward
9. Real-life application
10. Summary

Human-centered RL

- TAMER [2]
  - Human feedback is mapped to numeric value.
- SABL [3]
  - Human feedback is mapped to categorical strategies.
- COACH [4]
  - Human feedback is mapped to agent's policy (selecting actions).

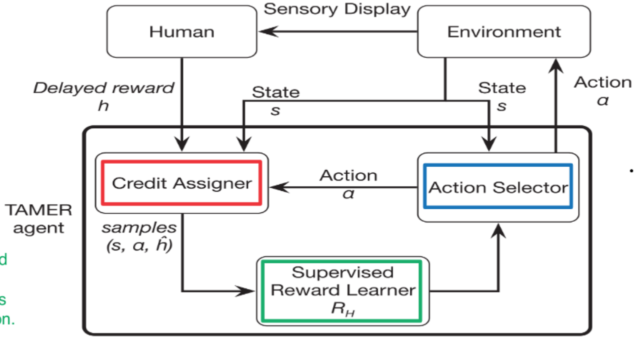
[2] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The TAMER framework," in **Proc. of the Fifth Int. Conf. on Knowledge Capture**, 2009, pp. 9–16

[3] R. Loftin **et al.**, "A Strategy-Aware Technique for Learning Behaviors from Discrete Human Feedback," in **Proc. of the AAAI Conf. on Artificial Intelligence**, vol. 28, no. 1, 2014

[4] J. MacGlashan **et al.**, "Interactive Learning from Policy-Dependent Human Feedback," in **Proc. of the 34th Int. Conf. on Machine Learning**, ser. Proc. of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2285–2294

TAMER

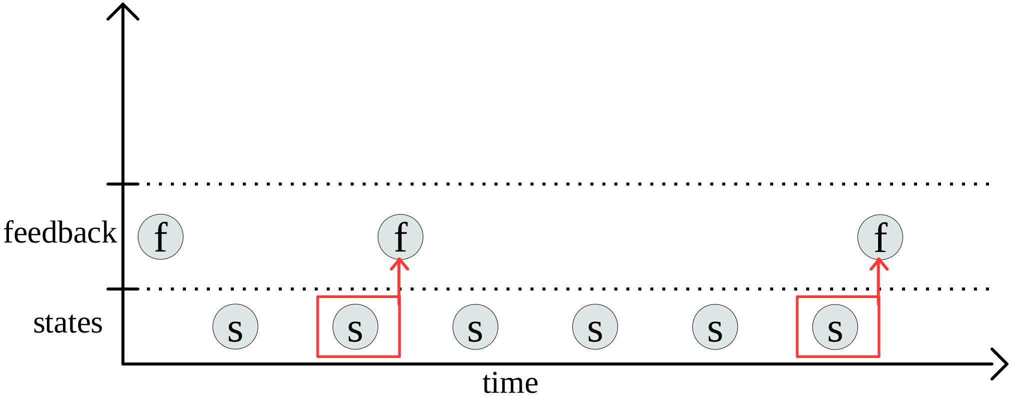
- Probabiliy density function to model the feedback delay.
- Learns a parameterised model for expected human reward, which is used for action selection.



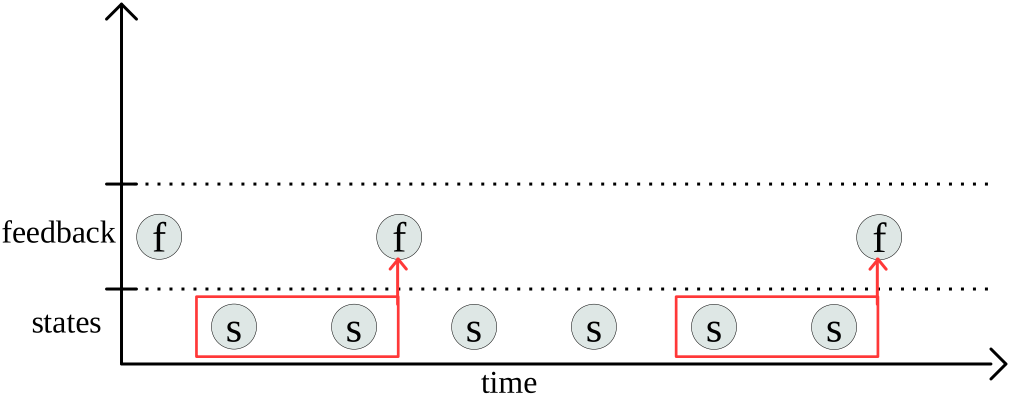
- Myopic rewards (γ=0) under the assumption that the human takes the long-term consequences into account in their evaluative feedback.

[5] W. B. Knox, "Learning from human-generated reward," Ph.D. dissertation, University of Texas at Austin, 2012

Without Credit Assignment



With Credit Assignment



Practical Example

- Agent is deployed in a simple **Mountain Car** environment, with
  - **continuous state space** - position of the car along the x-axis and velocity of the car
  - **discrete action space** - accelerate to the left, don't accelerate, accelerate to the right
- The **user** can give positive/negative **feedback** using the keyboard or no feedback (by not giving any input)

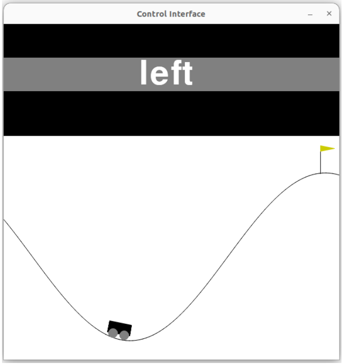
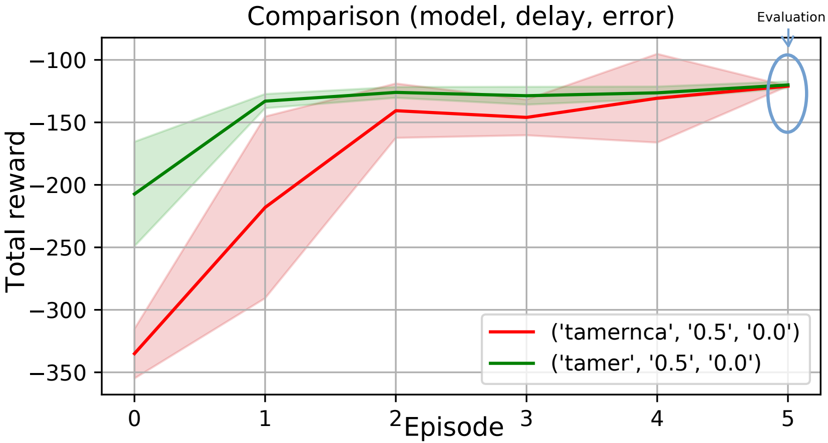


Figure 1: View of the deployment environment

Advantage of Credit Assignment



Going "deeper" ...

- One needs to find features (e.g. x car position, velocity) that can sufficiently define a state for TAMER.
- Would be easier if TAMER finds these features itself based on the image (like humans do).
- This problem is addressed by Deep TAMER [6].

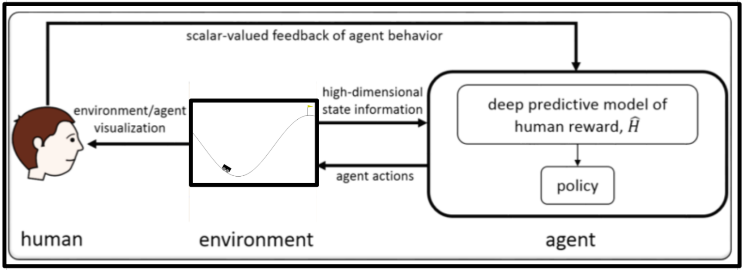


Figure 2: Based on [6]

[6] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone, “Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces,” in **Proc. of the AAAI Conf. on Artificial Intelligence**, vol. 32, no. 1, 2018

Is Deep TAMER a Solution?

- Needs pretraining of the encoder part (a lot of data necessary).
- Needs two input images for some environments (e.g. to estimate velocity).
- Hard to assess the quality of the extracted features.

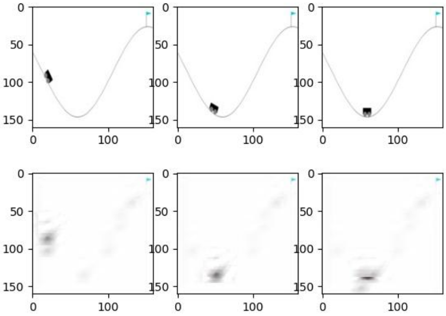


Figure 3: Reconstruction of the images from the features extracted by the encoder part of Deep TAMER

Categorical Feedback Strategies

A human teacher's feedback can be categorized into four types (inspired by behaviourism and animal training):

- Positive reward (R+)
  - Explicit feedback for correct behaviour.
- Negative reward (R-)
  - No feedback for correct behaviour.
- Positive punishment (P+)
  - Explicit feedback for wrong behaviour.
- Negative punishment (P-)
  - No feedback for wrong behaviour.

Categorical Feedback Strategies

- Different combinations of these feedback types are possible and it forms the teacher's feedback strategy.
  - Reward-focused $\rightarrow$ R+/P-
  - Punishment-focused $\rightarrow$ P+/R-
  - Balanced $\rightarrow$ R+/P+ (explicit reward and explicit punishment)
  - Inactive $\rightarrow$ R-/P- (rarely gives explicit feedback)

- The teacher can change the strategy during the course of training.
  - Teacher's feedback modelled probabilistically [7] and used with SABL algorithm.
  - Parameters:
    - » $\mu+$ $\rightarrow$ Probability that teacher will not give explicit feedback for correct behaviour
    - » $\mu-$ $\rightarrow$ Probability that teacher will not give explicit feedback for wrong behaviour
    - » ϵ $\rightarrow$ Probability that teacher misjudges the correctness of an action

[7] R. Loftin **et al.**, "Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning," **Autonomous Agents and Multi-Agent Systems**, vol. 30, pp. 30–59, 2016

Strategy-Aware Bayesian Learning (SABL)

- Bayessian inference.
- Assumes that teacher's strategy is **known**, i.e. $\mu+$ and $\mu-$ are known.
- Policy is updated based on the categorical probability of the given human feedback.
- Can be used only for low-dimensional discrete state space.
- Variant: **Inferring-SABL** or **I-SABL**
  - In reality, the teacher's strategy (i.e. $\mu+$ and $\mu-$) is **unknown**.
  - I-SABL **infers** the **teacher's strategy** by analyzing the feedback history.

Motivation for COACH

The following **characteristics** of training strategies used by humans:

- Correct actions are given **less positive feedback** progressively, as the **agent learns** to use that action succesfully.
- **Strength** of feedback varies depending on how much **improvement** or **deterioration** is observed in the agent's behaviour.
- Suboptimal actions may receive **positive** feedback if it **improves** the agent's behaviour; after the behaviour improves, the same suboptimal actions are given **negative feedback**.

[4] J. MacGlashan **et al.**, "Interactive Learning from Policy-Dependent Human Feedback," in **Proc. of the 34th Int. Conf. on Machine Learning**, ser. Proc. of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2285–2294

COnvergent Actor-Critic by Humans (COACH)

- COACH [4] is actor-critic-based reinforcement learning algorithm where human feedback is used as an **advantage function**.
  - advantage function - function that describes an **advantage** of selecting a certain action **over** the agent's policy

- The **sparse feedback** (also delayed feedback) problem is faced with **eligibility traces** which can smooth observed human feedback over past transitions.

- **Deep COACH** [8], namely COACH for **high-dimensional** input.

[4] J. MacGlashan **et al.**, "Interactive Learning from Policy-Dependent Human Feedback," in **Proc. of the 34th Int. Conf. on Machine Learning**, ser. Proc. of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2285–2294

[8] D. Arumugam, J. K. Lee, S. Saskin, and M. L. Littman, "Deep Reinforcement Learning from Policy-Dependent Human Feedback," **arXiv preprint arXiv:1902.04257**, 2019

Improving Learning from Evaluative Feedback

- **Human feedback** is usually:
  - **dense** (at the beginning of the interaction)
  - **flawed** (people generally make mistakes evaluating the agent's behaviour)
- On the other hand, **environmental reward** is usually:
  - **sparse**
  - **flawless** (determines optimal behaviour)
- Why not combine **human feedback** (HF) and **environmental reward** (ER) for agent learning?

Combining HF and ER

- **Reward shaping**
  - Modelled human feedback $\hat{H}$ is interpreted as a **reward**.
  - $r'(s,a) = r(s,a) + \beta * \hat{H}(s,a)$
- **Value shaping**
  - Modelled human feedback $\hat{H}$ is interpreted as **action-value function** (expected cumulative reward given that the agent starts with action a from s following policy π).
  - $Q'(s,a) = Q(s,a) + \beta * \hat{H}(s,a)$
- **Policy shaping**
  - Modelled human feedback $\hat{H}$ employed to directly influence the agent's **policy**.
  - e.g. $P(a = argmax(\hat{H}(s,a))) = min(\beta, 1)$

[1] M. Chetouani, "Interactive Robot Learning: An Overview," **ECCAI Advanced Course on Artificial Intelligence**, pp. 140–172, 2021

[9] W. B. Knox and P. Stone, "Combining manual feedback with subsequent MDP reward signals for reinforcement learning," in **Proc. of the 9th Int. Conf. on Autonomous Agents and Multiagent Systems: volume 1-Volume 1**. Citeseer, 2010, pp. 5–12

Example: Social Robotics

https://www.youtube.com/watch?v=VN1-bToWlac

[10] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal, "A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education," in **Proceedings of the AAAI Conference on Artificial Intelligence**, vol. 33, no. 01, 2019, pp. 687–694

Summary

- Human input can be used to **speed-up robot learning** in real-world tasks.
- Human input can take the form of **demonstrations**, **instruction**, or **evaluative feedback**.
- Learning from human evaluative feedback is called **human-centered reinforcement learning**.
- There are several **challenges** associated with obtaining, interpreting and using human input.
- Frameworks and methods that use human evaluative feedback include **TAMER**, **SABL** and **COACH**, to name a few.
- There are methods to combine human evaluative feedback and environmental reward including **reward shaping**, **value shaping** and **policy shaping**.

Thank you for your attention!