Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning

Abstract

Reinforcement learning (RL) commonly assumes access to well-specified reward functions, which many practical applications do not provide. Instead, recently, more work has explored learning what to do from interacting with humans. So far, most of these approaches model humans as being (nosily) rational and, in particular, giving unbiased feedback. We argue that these models are too simplistic and that RL researchers need to develop more realistic human models to design and evaluate their algorithms. In particular, we argue that human models have to be *personal*, *contextual*, and *dynamic*. This paper calls for research from different disciplines to address key questions about how humans provide feedback to AIs and how we can build more robust human-in-the-loop RL systems.

1 Introduction

Reinforcement learning (RL) has been successful in solving complex tasks, such as playing video games [20] or controlling robotic systems [9]. However, it remains challenging to apply RL to tasks without a well-specified reward function, such as autonomous driving [16]. RL from human feedback is a promising alternative that aims to interactively learn from human feedback instead of a fixed reward function [3].

Unfortunately, existing approaches to learning from human feedback rely on simple human models, such as Boltzmann distributions [15]. We argue that currently used models are too simplistic. We discuss challenges we expect such methods to encounter as they are being increasingly deployed in practical applications and present opportunities for further research towards improved modeling of human feedback.

In this paper, we, first, a brief overview of commonly used types of human-AI interaction and work in RL using human feedback (Section 2). Then, we present a range of open research questions about modeling human feedback (Section 3) and key dimensions for designing RL systems that can robustly learn from humans (Section 4). We raise many research questions about human modeling that are highly relevant to any human-AI interaction. We close by discussing the implications of our analysis and a call for interdisciplinary research addressing these questions (Section 5).

Position Statement. We argue that human-in-the-loop reinforcement learning needs more realistic human models.

Contributions. We, propose a framework for modeling human feedback as *personal*, *contextual*, and *dynamic*. For each category, we discuss open research questions both from a social science and a computer science perspective, and call for interdisciplinary research to address these questions.

2 Related Work

Research on human-centered machine learning has borrowed many methods from social science and the humanities to assess human feedback and interaction. However, most of the assumptions made about the reasoning processes, interaction and communication dynamics, as well as the integration of such feedback into learning models has remained unchallenged, lacking effective human-centered evaluations [26].

In this section, we provide a brief overview of types of human-AI interaction, with a specific focus on RL.

2.1 Collaborative Human-AI Interaction

To give an overview of the most common types of human-AI interaction, we provide a categorization in Figure 1. We distinguish three categories of interaction: *Instruction*, *Evaluation*, and *Cooperation*. For a more thorough review of possible types of human-AI interaction in the context of RL, we refer to recent reviews [15, 19].

Instruction. Human feedback is *instructive* if it tells the agent *what to do*. For example, demonstrations show the agent how to do a specific task. In some situations, the agent can also indirectly obtain information about what to do simply by observing the state of the world [25]. More directly, the human can tell the agent what to do by *correcting* it physically [1], or providing *improvements*, e.g., an alternative action in a specific situation [13].

Evaluation. Human feedback is *evaluative* if it tells the agent *how well it is doing*. Humans can provide this information directly, e.g., by comparing trajectories [3] or shutting off the agent [10]. The agent can also get implicit evaluations, e.g., by measuring user engagement [29], or by monitoring gestures or facial expressions [4].

Cooperation. More complicated forms of human-AI interaction need to be modeled in a cooperation framework [11]. While we could also model all instructive and evaluative feedback as a form of cooperation, the latter allows for more general forms of interaction beyond giving feedback. For example, a human and an RL agent might have to solve a problem together, requiring them to learn from each other [2].

2.2 Models of Human Feedback in RL

The most popular way to integrate human feedback in RL is learning from demonstrations, where the agent learns a task from observing a (human) expert's *demonstrations* of a task. This can happen either via imitation learning [12], or via inverse reinforcement learning [21]. The main limitation of this approach is that the demonstrations have to be (close to) optimal, which is often difficult for humans to achieve. Instead of demonstrations, a human can also provide a direct *reinforcement* signal, e.g., a binary rating "good" or "bad" [17]. In many applications, it can be easier to provide contrastive feedback, e.g., by comparing two possible actions or trajectories, rather than an absolute evaluation. This motivates *preference-based* RL [28].

All of these methods require a model of how humans give feedback. They commonly assume that humans are perfectly rational or at least unbiased. One of the most common ways to model noisy feedback is using a Boltzmann distribution [15]. Some work tries to evaluate the effect of this *misspecification* [8]. Other work tries to learn biases in human feedback from data, with mixed results [24]. Overall, current methods in RL from human feedback are heavily affected if their assumptions about the human model are not satisfied [18].

The B-Pref benchmark [18] models *irrationalities* that humans might exhibit when giving feedback, intending to move towards more realistic human models. The benchmark considers humans making systematic mistakes, skipping comparisons of very similar or very different trajectories, or giving myopic feedback, i.e., weighting recently seen things higher when making decisions. We consider this work a promising step in the right direction. However, none of the simulations in B-Pref is grounded in the existing literature on cognitive biases, and the benchmark is not based on human data.

Some literature on recommender systems has explored the use of more advanced human models, but these methods are often tailored to the specific applications and have not made it to more general applications of RL yet [14, 23].

3 Challenges in Human Feedback Modelling

Most existing approaches to incorporating human feedback into RL, assume humans act noisily but are unbiased. It is common to assume that humans are goal-driven and act rationally consistent. This model is, of course, wrong, and humans are far from rational and unbiased.

We argue that in practice, human-AI interaction is *personal*, *contextual*, and *dynamic*. This section discusses these three aspects, focusing on open research questions. The next section will discuss potential implications for designing systems that interactively learn from humans.

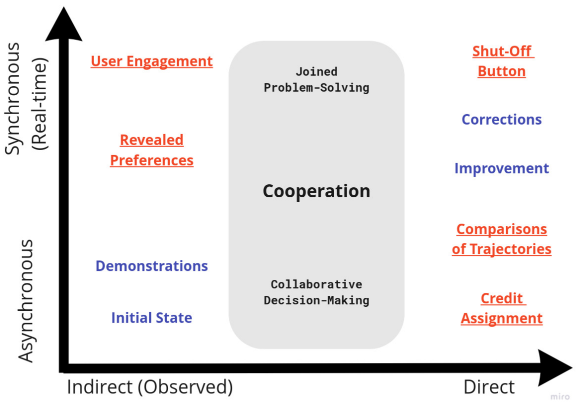

Figure 1: We classify a subset of common types of human feedback in RL along three dimensions. Humans can give *direct* feedback to an RL agent performing a task. The agent can also obtain *indirect* feedback by observing the human. Feedback can be provided *synchronously*, i.e., in real-time, or *asynchronously*, i.e., before or after the agent acts. Finally, feedback is **instructive** if it tells the agent what to do explicitly; feedback is **evaluative** if it only tells the agent how good it is doing.

Most of our discussion applies to any human-AI interaction. However, to be concrete, we focus on RL agents receiving evaluative or instructive feedback. This setting inherits most problems of modeling humans in general, but it allows for more concrete research on practical applications.

3.1 *Personalized* Feedback & Interaction

We cannot hope to find a universal model that describes how humans interact with RL systems that we can use to design these systems. Human-AI interaction is inherently personal, and we need to model it as such.

How a human interacts with an RL agent will depend on personality factors [22]. For example, it is likely that in many tasks, the user's conscientiousness will affect how they evaluate an agent. A crucial first step towards designing personalized interaction is to understand how personality affects human-AI interactions.

In addition to different personalities, each person also has prior knowledge they bring into an interaction. Whether the user is a domain expert, an ML expert, or a novice, will affect which situations they can judge and which types of feedback they can give reliably. For example, an expert might be able to provide near-optimal demonstrations, whereas a novice might not. Still, the novice might be able to evaluate the task competently when giving the right user interface [5].

As a first step towards building personalized human models, we propose to study the following research questions:

> RQ1: *How do **personality factors** influence how humans interact with AI systems?*
>
> RQ2: *Are there **measurable** personality factors with a clear impact on the human feedback and interaction dynamic to allow for personalization?*
>
> RQ3: *How can we **quantify** prior knowledge and semantic understanding to model interaction dynamics?*

3.2 *Contextualized* Feedback & Interaction

To model human-AI interaction accurately, we need to take into account the (sociotechnical) *context* of the interaction. In particular, the interaction dynamics depend on when and where it is happening and which social and cultural norms exist around the interaction [6, 7]. For example, a medical doctor might be more careful about evaluating the information from an AI system than the average user of a personal smartphone assistant. In the context of RL the context of the interaction depends on the environment in which the agent and the human act, and how the human relates to this environment.

To enable contextualized modeling of human-AI interaction, we need to understand:

> RQ4: *Which aspects of the* ***sociotechnical context*** *affect human-AI interaction?*

> RQ5: *How do* ***cultural*** *(and other)* ***differences*** *influence the interaction?*

> RQ6: *How do contextual factors influence individual personality factors, and can they be modeled* ***independently****?*

3.3 *Adaptive* Feedback & Interaction

A major limitation of most human models used in the RL literature is that they are static, i.e., they do not change throughout the interaction. This is unrealistic; in practice, both the user and the AI system will accumulate knowledge that changes how they interact with each other [27]. Other factors, such as the user's energy and motivation levels, might change over time and affect the interaction. Moreover, the context of the interaction might change due to external factors.

RL systems have to adapt to these *dynamic* factors and adapt how they interact with humans over time. To build robust adaptive systems, it is necessary to investigate:

> RQ7: *Which factors have temporal and interaction-dependent* ***variation****?*

> RQ8: *Can we* ***measure*** *and* ***predict*** *changes in these factors during interaction sequences?*

> RQ9: *How do personal and contextual factors* ***adapt*** *to changes in the interaction?*

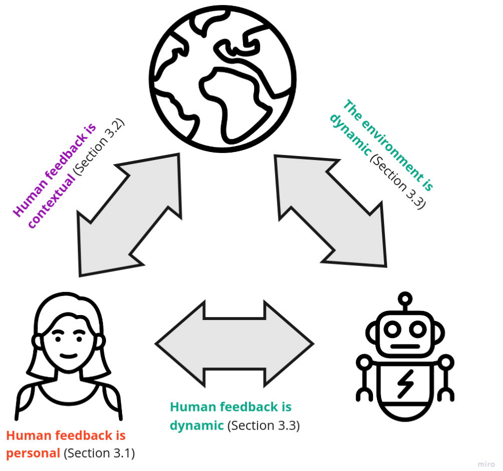
Human feedback is contextual (Section 3.2)
The environment is dynamic (Section 3.3)
Human feedback is personal (Section 3.1)
Human feedback is dynamic (Section 3.3)
miro

Figure 2: Schematic overview of RL with human feedback. In Section 3, we discuss the key components of the interaction. Section 4 discusses key dimensions for robust interactive learning: **Personalized Learning** (Section 4.1), **Contextualized Modeling** (Section 4.2), and **Adaptive Querying** (Section 4.3).

4 Implications for RL Applications

Now, let us turn our focus towards designing RL systems that learn from interacting with humans. Along our three dimensions of modelling human feedback and interaction, we now highlight how to design RL systems that are personalized, contextualized and adaptive.

4.1 *Personalized* Learning

To build RL systems that can learn from different people, we need to ensure that all parts of the interaction are personalized. Building on a better understanding of the aspects discussed in Section 3.1, we could decide which system designs are most appropriate for which users. This includes answering the following research questions:

> RQ10: *How should we choose the* ***type*** *of human feedback for an individual user?*

> RQ11: *How much can an RL system* ***trust*** *the responses of an individual user?*

> RQ12: *Which* ***explanations*** *can an RL system provide to an individual user to allow them to give better feedback?*

4.2 *Contextualized* Modeling

Similar to building personalized learning systems, we want to ensure to make them aware of the sociotechnical context of the interaction as well. Building on the questions discussed in Section 3.2, we want to answer similar questions from an RL perspective:

> RQ13: *In which context can an RL system* ***rely*** *on users providing high-quality feedback?*

> RQ14: *How can an RL system provide* ***context-appropriate*** *explanations of it's behavior?*

> RQ15: *How can we ensure RL systems can learn from human feedback reliably even in* ***novel*** *contexts?*

4.3 *Adaptive* Querying

We need to enable RL systems to adapt to changes in their interaction with humans. This is particularly important for designing human-in-the-loop systems that do not have a separate phase of learning from humans, but continuously interact with humans during their deployment. Such systems need to have uncertainty estimates that are well-calibrated with respect to how users and interaction patterns might change. We propose several concrete research questions in this direction:

> RQ16: *How can RL systems make* ***situation-aware*** *queries to users?*

RQ17: *How can RL systems maintain appropriate* ***uncertainty*** *to* ***detect*** *changes in the interaction?*

RQ18: *How much can systems learn about a user's interaction patterns* ***online****, and how many inductive biases do we need to* ***encode****?*

5 Discussion

After discussing the most important dimensions for designing more human-centered RL systems, let us highlight the key challenges and research opportunities.

5.1 Challenges

An Interdisciplinary Research Challenge. Modeling human feedback is not *only* a technical problem. The most crucial open research questions need to be answered from a human-centered perspective. However, no single research discipline is currently offering insights on the right level of granularity. On the one hand, research in Neuroscience aims to provide a detailed understanding on the basis of human behavior. However, its insights are not (yet) actionable for our purpose. On the other hand, behavioral and cognitive psychology research asks many similar questions. However, its focus is often too broad to yield actionable insights on how to model human feedback in RL systems.

More Difficulty for RL. The main reason for using simple human models in current work is that they make learning more tractable. If we can no longer get unbiased feedback about rewards, how can we hope to still learn a good RL policy. Developing more realistic human models will likely make RL more difficult.

5.2 Research Opportunities

An Interdisciplinary Approach. To make progress in modeling human feedback, we need to answer multi- and interdisciplinary research questions. These will require research from a wide range of disciplines to engage with these questions, including, but not limited to, researchers from Computer Science, (Cognitive) Psychology, Ethics, Philosophy, Behavioral Science, Communication Sciences, Sociology, and Neuroscience. To answer the research questions we posed, we need to first compare the current understanding from different disciplines. Then, we can aim to operationalize the open questions within specific disciplines and design experiments to answer them.

Towards More Robust RL. Ultimately, we hope to gain insights into measurable factors that affect human-AI interaction, which can be used to build better human models. Such models promise to make it easier to design new RL algorithms that learn from humans and benchmark existing methods in more realistic simulations without running user studies to evaluate small algorithmic changes.

An Opportunity for New Algorithms. A human-centered perspective on the RL learning problem creates challenges for applying existing methods, that assume access to a reward function. But, it provides an opportunity for developing novel, human-centered learning algorithms that are designed with the goal of learning from and with humans.

6 Conclusion

We reviewed research on human feedback in RL from a human-centered perspective. We argued that current human models used in RL are too simple and that we need better human models if we want to design systems that can learn from humans robustly in the real world. We argued that we need personal, contextual, and dynamic models to design robust RL systems that learn from humans. We hope to start an interdisciplinary discussion around these topics with the goal of building better human models and designing interaction protocols that can work outside of simulations.

References

[1] Andrea Bajcsy, Dylan P Losey, Marcia K O'Malley, and Anca D Dragan. Learning robot objectives from physical human interaction. In *Conference on Robot Learning*, pages 217–226. PMLR, 2017.

[2] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

[3] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in neural information processing systems*, 2017.

[4] Yuchen Cui, Qiping Zhang, Alessandro Allievi, Peter Stone, Scott Niekum, and W Bradley Knox. The empathic framework for task learning from implicit human feedback. *arXiv preprint arXiv:2009.13649*, 2020.

[5] John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37, 2018.

[6] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. The who in explainable ai: how ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*, 2021.

[7] Upol Ehsan and Mark O Riedl. Human-centered explainable ai: towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*, pages 449–466. Springer, 2020.

[8] Rachel Freedman, Rohin Shah, and Anca Dragan. Choice set misspecification in reward inference. In *IJCAI-PRICAI Workshop On Artificial Intelligence Safety*, 2020.

[9] Tuomas Haarnoja, Sehoon Ha, Aurick Zhou, Jie Tan, George Tucker, and Sergey Levine. Learning to walk via deep reinforcement learning. *arXiv preprint arXiv:1812.11103*, 2018.

[10] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[11] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[12] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

[13] Ashesh Jain, Shikhar Sharma, Thorsten Joachims, and Ashutosh Saxena. Learning preferences for manipulation tasks from online coactive feedback. *The International Journal of Robotics Research*, 34(10):1296–1313, 2015.

[14] Anthony Jameson, Martijn C Willemsen, Alexander Felfernig, Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Li Chen. Human decision making and recommender systems. In *Recommender systems handbook*, pages 611–648. Springer, 2015.

[15] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.

[16] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. Reward (mis) design for autonomous driving. *arXiv preprint arXiv:2104.13906*, 2021.

[17] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16, 2009.

[18] Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. B-Pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026*, 2021.

[19] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He, and Guangliang Li. A review on interactive reinforcement learning from human social feedback. *IEEE Access*, 8:120757–120765, 2020.

[20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[21] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, volume 1, page 2, 2000.

[22] Maria Augusta SN Nunes and Rong Hu. Personality-based recommender systems: an overview. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 5–6, 2012.

[23] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164, 2011.

[24] Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*, 2019.

[25] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. Preferences implicit in the state of the world. In *International Conference on Learning Representations*, 2019.

[26] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, D Horng Chau, Alex Endert, and Daniel Keim. A survey of human-centered evaluations in human-centered machine learning. In *Computer Graphics Forum*, volume 40, pages 543–568. Wiley Online Library, 2021.

[27] Fabian Sperrle, Astrik Jeitler, Jürgen Bernard, Daniel Keim, and Mennatallah El-Assady. Co-adaptive visual data analysis and guidance processes. *Computers & Graphics*, 100:93–105, 2021.

[28] Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

[29] Qian Zhao, F Maxwell Harper, Gediminas Adomavicius, and Joseph A Konstan. Explicit or implicit feedback? engagement or satisfaction? a field experiment on machine-learning-based recommender systems. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 1331–1340, 2018.