Neural Computing and Applications (2023) 35:16689–16691
https://doi.org/10.1007/s00521-023-08748-x

EDITORIAL

Human-aligned reinforcement learning for autonomous agents and robots

Francisco Cruz[1,6] · Thommen George Karimpanal[2] · Miguel A. Solis[3] · Pablo Barros[4] · Richard Dazeley[5]

Received: 31 May 2023 / Accepted: 3 June 2023 / Published online: 24 June 2023

Reinforcement learning (RL), powered by artificial neural networks, has made considerable advances in agent autonomy, both in simulation-based and real-world environments. As these agents become more prevalent in various aspects of our daily lives, it is imperative to ensure their behaviors are aligned with human interests and social values. Deploying agents without such constraints can be particularly damaging in embodied applications and in applications where the agent progressively attains higher degrees of independence. In addition to ensuring aligned behaviors, there exists a potential for agents to learn more efficiently and in a contextually aware manner if they can leverage the human inputs and feedback in a scalable manner. Such context awareness and efficiency are particularly important in robotics and other interactive applications, where interactions with the real world incur costs in terms of time and energy and where certain subsets of the action space could lead to undesirable consequences. Ideally, such agents should be able to explain their actions in a human interpretable manner, perhaps even communicating the predicted outcomes of counterfactual actions or policies.

This special issue was designed to collate high-quality works addressing the above and other related issues in reinforcement learning, with the intention of presenting and discussing them via a diverse set of original research approaches, ranging from theoretically rooted frameworks to more practically relevant techniques. In all, we have accepted 17 papers to be included in this special issue. The guest editors greatly appreciate the patience, diligence, and dedicated efforts of all the reviewers, many of whom made insightful recommendations that helped push the quality of some of the submissions to well beyond the recommended threshold for acceptance into the special issue. We would also like to thank the editor-in-chief, Professor John MacIntyre of the Neural Computing and Applications journal for his support and encouragement, as well as the editorial staff, Rashmi Jenna and Annette Hinze for their assistance during the collation of articles and processing of this special issue. In the remainder of this article, we summarize and describe each of the accepted contributions in separate paragraphs, highlighting their novelty and significance in addressing some of the modern challenges in human-aligned RL.

The first paper by Rietz et al. proposed providing hierarchical goals as context for interpreting the behavior of RL agents. The key idea is that in the context of a hierarchical goal, the future behavior of the agent becomes more predictable. Through qualitative analysis of one-step reward decomposition explanations, this work first showed that interpretability was limited in scenarios with multiple distinct optimal policies and that by providing hierarchical

✉ Francisco Cruz
f.cruz@unsw.edu.au

Thommen George Karimpanal
thommen.karimpanalgeorge@deakin.edu.au

Miguel A. Solis
miguel.solis@unab.cl

Pablo Barros
pablo.barros@sony.com

Richard Dazeley
richard.dazeley@deakin.edu.au

[1] School of Computer Science and Engineering, University of New South Wales, High St, Kensington UNSW, Sydney, NSW 2052, Australia

[2] Applied Artificial Intelligence Institute, Deakin University, 75 Pigdons Road, Waurn, Ponds, Geelong, VIC 3216, Australia

[3] Faculty of Engineering, Universidad Andres Bello, Santiago, Chile

[4] Brussels Laboratory, Sony Group Corporation, Leonardo da Vincilaan 7, 1930 Zaventem, Belgium

[5] School of Information Technology, Deakin University, 75 Pigdons Road, Waurn Ponds, Geelong, VIC 3216, Australia

[6] Escuela de Ingeniería, Universidad Central de Chile, Santa Isabel 1186, 8330601 Santiago, Chile

goals as context, a high degree of interpretability was retained.

Mishra et al. described the theoretical foundations of the cart-pole balancing problem and proposed the use of a Huber loss function to learn a policy. The authors experimentally justified the choice of the Huber loss, showing that it leads to faster learning and convergence.

Millan-Arias et al. studied the behavior of RL agents in a proxemic-based environment, where the human comfort levels are dependent on the agent's proximity. In a simulated environment, where the non-conformity information is provided by an issuer, this work demonstrated that RL can be used to identify proxemic regions, as well as to learn the best path to approach the issuer.

Barros et al. proposed a novel reinforcement learning mechanism that utilizes the social impact of rivalry behavior, combining objective and social perception mechanisms to derive a rivalry score for modulating agent learning. The authors designed an interactive game scenario to examine how rivalry behaviors affect an agent's play, as well as the experience of human players in the game. Their study demonstrated that humans could discern specific social characteristics when playing against rival agents compared to common agents, leading to a direct impact on the performance of human players in subsequent games.

Andres et al. examined different ways in which heterogeneous RL agents can share information with each other, and proposed a collaborative learning framework, via a centralized critic module, with which agents can learn faster than they would individually. This framework was extensively evaluated in terms of their configurations on what information is shared and when. The work also exposed the need for regulating extrinsic and intrinsic rewards in order to avoid undesirable agent behaviors.

One of the important research areas in human-aligned RL deals with the question of an agent learning from experts. Love et al. proposed an approach, Cautiously Learning with Unreliable Experts (CLUE), which deals with the scenario where experts may not be reliable. By modeling and updating the reliability of each expert and employing a Bayesian approach for pooling advice, the agent is shown to be able to robustly imbibe the appropriate advice even in scenarios where the experts are unreliable.

Persiani and Hellstrom proposed a method to augment an RL agent with legibility, allowing the policy to be regularized after training, without having to modify the underlying RL algorithm. The proposed legibility is essentially a measure of how discernible the policy is, and is loosely associated with an agent's intention. The key idea behind their approach is based on evaluating how the optimal policy may produce state–action pairs that could cause an observer to infer the incorrect policy. As the underlying algorithm remains unchanged, this method is flexible and thus applicable to arbitrary agents and environments.

Emuna et al. proposed a generic approach through which an artificial learning agent can learn to autonomously drive with human-like driving skills. The stochastic behaviors of human experts were mimicked by learning distributions of task variables from examples. In a simulated highway driving environment, this work demonstrated the ability to drive in a human-like manner while being able to handle new road and obstacle distributions that were unseen during training.

Mathewson et al. introduced the idea of communicative capital as a resource that could be developed even passively through interactions between humans and machines. It was shown that the development of this resource enables the human–machine partnership to achieve superior task performance compared to either one of the individual entities (humans or machines). Using the example of tightly coupled human–machine interfaces such as prosthetics, the authors suggest how the idea of communicative capital could extend current viewpoints on how to best support the use of complex prosthetic devices.

In the next work, Harnack et al. consider the interactive RL viewpoint and examine the effect of feedback frequency, with the aim of isolating the best or most beneficial feedback frequencies for interactive RL. To this end, this work isolates and quantifies the effect of feedback frequencies in robotics tasks of continuous state and action spaces. The primary finding of this study reports that there is no ideal feedback frequency and that this is a quantity that depends on the task complexity and performance threshold, and the ideal value of it varies according to the agent's task proficiency.

Miras et al. examined a human–robot collaborative scenario where the human helps a mobile robot that could move about in its environment. The aim was to collect a set of items within a walled area, with the human being able to directly control the robot if undesired robot behavior is observed. In both simulated and real-world experiments, the robot behaviors were shown to be improved considerably through human assistance.

In the work by Krening, it was shown that Q-learning could be used to model utilitarian decision-making in a human–machine team. Through this model, the symbiotic team was shown to more closely follow utilitarianism than either individual (human or machine) could. The structure of Q-learning was suggested to naturally output the utilitarian decision, along with the ranked order of suboptimal actions. In addition, the ability to account for not just the magnitude of rewards but also the stochastic nature of occurrences made the proposed framework a more realistic utilitarian model.

Celemin and Kober proposed a method to improve human–robot interaction for non-experts and imperfect teachers. The agents were equipped with uncertainty estimation, particularly, with a lack of knowledge awareness, and with demonstration ambiguity, such that human inputs could be requested when deemed necessary. The method also enabled teachers to train with corrective demonstrations, evaluative reinforcements, and implicit positive feedback. Through a first set of experiments, the proposed method was shown to improve learning convergence in the presence of imperfect teachers. Further, components of this method, via a user study, were also shown to improve the teaching experience and the data efficiency of learning. Hence, the proposed method constituted a bi-directional approach that improved both agent's learning and the teaching experience in realistic settings where the teacher is imperfect.

Kokel et al. proposed a hybrid deep RL architecture, RePReL, to efficiently provide state abstractions. The architecture involved a high-level planner communicating with a low-level RL agent to learn useful state abstractions. The paper empirically demonstrated improved generalization and transfer capabilities which were afforded through efficient state abstractions. Being a plug-and-play framework, the architecture allows different planners to be used in conjunction with different RL agents. Moreover, as the architecture accommodates continuous state–action spaces, it could be useful in real-world applications.

Dazeley et al. proposed RL as a potential backbone for broad explainable artificial intelligence (Broad-XAI), to provide coherent explanations across multiple levels of explainability. They suggest that providing explanations in an RL context would allow for increased focus on the types of explanations possible in RL and how they can be combined to provide levels of explanation to better understanding as required by users. To this effect, the paper proposes the conceptual framework, Causal XRL, based on a previously proposed causal framework. In addition to this, the paper contributed a comprehensive overview of explainable RL, while identifying sub-topics of explainable RL for further research.

Harland et al. proposed an apologetic framework to simultaneously address the practical as well as social needs of a user. The framework is based on an Act–Assess–Apologize cycle, where the apology acknowledges incorrect behaviors, an explanation, and an intention for improved future behaviors. To handle auxiliary objectives that are contradictory, the paper used impact minimization techniques. The framework was shown to achieve alignment goals in several cases, and post-apologetic behaviors were shown to have statistically significant improvements in user-sensitive objectives.

Pierson et al. examined explanations for RL behaviors through user studies. The study first collected RL behaviors, the explanations for which were first generated used a variant of an existing algorithm, and then compared to previous works' video-based summaries in a user study. The user studies, based on both trust and performance, were then used to evaluate the explanations. The explanations from the proposed method were shown to be more favorable in terms of the agents and their strategies, and further, the users' reported trust in an explanation did not directly correlate with performance.