frontiers | Frontiers in Neurorobotics

TYPE Perspective
PUBLISHED 09 November 2023
DOI 10.3389/fnbot.2023.1280341

OPEN ACCESS

EDITED BY
Alessandro Mengarelli,
Marche Polytechnic University, Italy

REVIEWED BY
Jiahao Chen,
Chinese Academy of Sciences (CAS), China
Wen Qi,
Polytechnic University of Milan, Italy

*CORRESPONDENCE
Shangding Gu
shangding.gu@tum.de

†These authors have contributed equally to this work

RECEIVED 20 August 2023
ACCEPTED 18 October 2023
PUBLISHED 09 November 2023

CITATION
Gu S, Kshirsagar A, Du Y, Chen G, Peters J and Knoll A (2023) A human-centered safe robot reinforcement learning framework with interactive behaviors.
Front. Neurorobot. 17:1280341.
doi: 10.3389/fnbot.2023.1280341

A human-centered safe robot reinforcement learning framework with interactive behaviors

Shangding Gu[1*†], Alap Kshirsagar[2†], Yali Du[3†], Guang Chen[4], Jan Peters[2] and Alois Knoll[1]

[1]Department of Computer Science, Technical University of Munich, Munich, Germany, [2]Department of Computer Science, Technical University of Darmstadt, Darmstadt, Germany, [3]Department of Informatics, King's College London, London, United Kingdom, [4]College of Electronic and Information Engineering, Tongji University, Shanghai, China

Deployment of Reinforcement Learning (RL) algorithms for robotics applications in the real world requires ensuring the safety of the robot and its environment. Safe Robot RL (SRRL) is a crucial step toward achieving human-robot coexistence. In this paper, we envision a human-centered SRRL framework consisting of three stages: safe exploration, safety value alignment, and safe collaboration. We examine the research gaps in these areas and propose to leverage interactive behaviors for SRRL. Interactive behaviors enable bi-directional information transfer between humans and robots, such as conversational robot ChatGPT. We argue that interactive behaviors need further attention from the SRRL community. We discuss four open challenges related to the robustness, efficiency, transparency, and adaptability of SRRL with interactive behaviors.

KEYWORDS

interactive behaviors, safe exploration, value alignment, safe collaboration, bi-direction information

1. Introduction

Deep learning has shown impressive performance in recent years (LeCun et al., 2015; Wang et al., 2022; Liu et al., 2023; Zhao and Lv, 2023). By leveraging deep learning, Reinforcement Learning (RL) has achieved remarkable successes in many scenarios and superhuman performance in some challenging tasks (Gu et al., 2022b, 2023), e.g., autonomous driving (Gu et al., 2022a), recommender system (Zhao et al., 2021), robotics (Brunke et al., 2021), games (Silver et al., 2018; Du et al., 2019; Han et al., 2019), and finance (Tamar et al., 2012). Most RL methods aim to maximize reward performance without considering safety constraints. However, safety is critical when deploying RL in real-world applications, especially in robotics. In Robot RL (RRL), a robot interacts with static or dynamic environments to learn the probability of better actions. When humans are also part of the robot's environment, ensuring their safety is crucial. This paper proposes a framework to achieve Safe Robot RL (SRRL) by leveraging interactive behaviors.

Interactive behaviors are behaviors that can mutually influence the interacting elements. Interaction is everywhere in human life (Kong et al., 2018), and agent-environment interaction is the basis of RL (Sutton and Barto, 2018). When humans and robots act in a shared environment, their actions can be influenced by each other through interactive behaviors (Thomaz and Breazeal, 2006; Knox and Stone, 2009; MacGlashan et al., 2017; Kazantzidis et al., 2022; Lou et al., 2023). For example, a robot navigating in a public space can plan its path to avoid collisions with pedestrians

and provide signals to pedestrians to move aside. The pedestrians can also plan their path to avoid collisions with the robot and provide signals to the robot to move aside. Figure 1 shows a schematic of interactive behaviors consisting of three elements: robots, environment, and humans. The outer loop consists of feedback or signals from the robots to the humans and vice versa. The two inner loops consist of actions from robots and humans in the environment and feedback or rewards from the environment to robots and humans. Here, we use the term "feedback" to mean any type of information transfer between the two interacting elements.

Interactive behaviors can lead to better SRRL by enabling bi-directional information transfer between the interacting elements. It is a core technology to improve the dialogue performance of a Large Language Model (LLM), e.g., ChatGPT (OpenAI, 2023) by leveraging interactive behaviors. For example, in ChatGPT, on the one hand, high-quality data from human feedback is collected to design the reward model (Stiennon et al., 2020; Gao et al., 2022). On the other hand, after having human feedback, the agent model will be trained to align human values, and then give safe feedback to humans. In most conventional SRRL approaches, there is no human interaction, e.g., CPO (Achiam et al., 2017) and ATACOM (Liu P. et al., 2022). With interactive behaviors, the robot can learn about human behaviors and convey its features and decision-making processes to humans. Those approaches that consider the human-in-the-loop of the robot's learning process do not utilize feedback from the robot learner to the human teacher.

In this paper, we investigate interactive behaviors to achieve three stages of human-centered SRRL: "safe exploration," "safety value alignment," and "safe human-robot collaboration." In the "safe exploration" stage, the robot must explore the unknown state space while preserving safety. In the "safety value alignment" stage, the robot has to align its intentions with the humans. Finally, in the "collaboration" stage, the robot should contribute toward achieving shared goals with humans.

2. Related work on safe robot-reinforcement learning

Safe robot learning has received substantial attention over the last few decades (Turchetta et al., 2019; Baumann et al., 2021; Kroemer et al., 2021; Marco et al., 2021; Kaushik et al., 2022). SRRL methods focus on robot action and state optimization and modeling to ensure the safety of robot learning. For instance, Gaussian models are used to model the safe state space (Akametalu et al., 2014; Sui et al., 2015, 2018; Berkenkamp et al., 2016; Turchetta et al., 2016; Wachi et al., 2018); formal methods are leveraged to verify safe action and state space (Fulton and Platzer, 2018; Kochdumper et al., 2022; Yu et al., 2022); control theory is applied to search safe action space (Chow et al., 2018, 2019; Koller et al., 2018; Li and Belta, 2019; Marvi and Kiumarsi, 2021).

Akametalu et al. (2014) introduced a reachability-based method to learn system dynamics by using Gaussian models, in which the agent can adaptively learn unknown system dynamics and the maximal safe set. Berkenkamp et al. (2016) proposed a region of attraction method to guarantee safe state space based on Gaussian processes and Bayesian optimization. Although the method provided a theoretical analysis for safe robot learning, the assumptions in the study are quite strong and may not be applicable in practical environments. Sui et al. (2015) present a safe exploration method considering noise evaluations, in which the safety is ensured with a high likelihood via a Gaussian process confidence bound. Moreover, the sample complexity and convergence of the method are analyzed, and real-world applications, such as movie recommendations and therapeutic spinal cord stimulation, are used to test how well the system works. Nonetheless, they considered the safe exploration as a bandit setting without any constraints (Turchetta et al., 2016). Turchetta et al. (2016) developed an algorithm in which Gaussian processes are leveraged to model the safe constraints. However, a set of starting safe states is required from which the agent can begin to explore in the work of Sui et al. (2015) and Turchetta et al. (2016).

Some recent works have investigated the application of formal methods for SRRL. Fulton and Platzer (2018) used formal verification to check the correctness of the state transition model and select safe action during RL. They allow unsafe actions if the model is incorrect. Researchers have proposed three types of provably safe RL methods for hard safety (Krasowski et al., 2022): action mask, action replacement, and action projection. Action masking approaches apply a safety layer to restrict the agent's actions to safe actions only (Krasowski et al., 2020). Action replacement approaches replace unsafe actions with safe actions (Hunt et al., 2021). Action projection methods project the unsafe actions to close safe actions (Kochdumper et al., 2022).

Control theory based SRRL approaches have utilized Lyapunov functions and Model Predictive Control (MPC). Chow et al. (2018, 2019) introduced Lyapunov functions based on discrete and continuous control methods for the global safety of behavior policy in RL. However, designing Lyapunov functions for different environments may be difficult, and the requirement of a baseline policy may be challenging to satisfy during real-world applications. Koller et al. (2018) proposed an MPC based method to ensure safe exploration using a statistical model of the system. Marvi and Kiumarsi (2021) introduced a control barrier function method for safe off-policy robot learning without needing to have a thorough understanding of system dynamics, in which the cost functions are augmented by a control barrier function.

Most of the above methods do not consider human-robot interaction. There are some prior works that investigate human-centered SRRL. For instance, Kazantzidis et al. (2022) introduced a mechanism to ensure safety during exploration by harnessing human preferences. Reddy et al. (2020) present a method to learn the model of human objectives by leveraging human feedback based on hypothetical behaviors, and then the model can be used to ensure the safety of robot learning. Saunders et al. (2018) try to guarantee reinforcement learning safety by human interventions, where human interventions are learned through a supervised learning model. However, most of these works do not consider the mutual influence of humans and robots in shared environments. Thus, interactive behaviors between humans and robots that leverage bi-directional information transfer, as shown in Figure 2, still need further investigation to ensure SRRL.

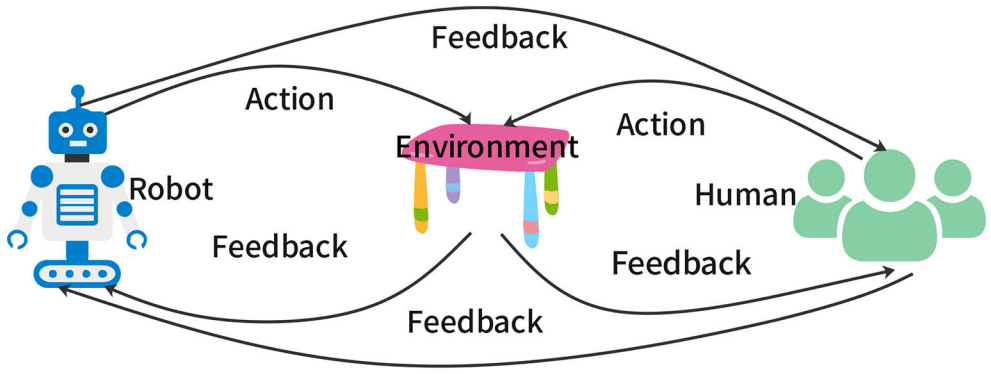

FIGURE 1
Schematic of interactive behaviors.

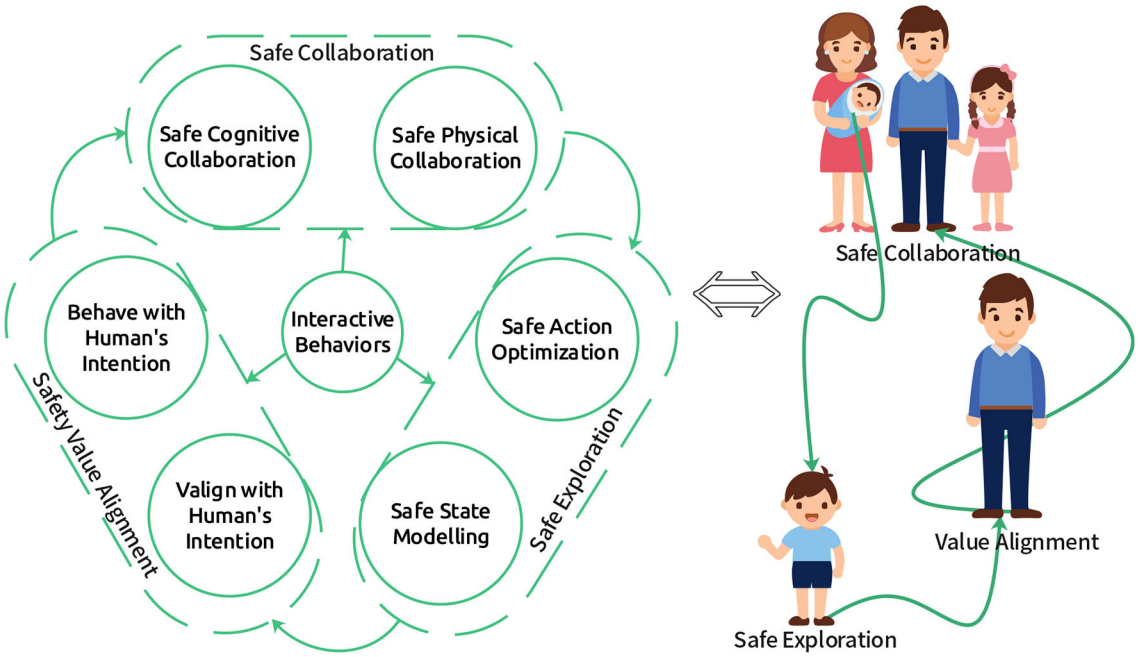

FIGURE 2
Three stages of human-centered SRRL. When a newborn infant enters a new environment, if it wants to survive in the new environment, it must first learn about the environment via interactive behaviors (**Safe Exploration**), e.g., bi-directional feedback with its parents to learn how to walk. Then it should find something helpful to survive better via interactive behaviors (**Value Alignment**), e.g., bi-directional feedback with its teachers to learn human values. Finally, it has to collaborate with others to maximize the society's reward via interactive behaviors (**Safe Collaboration**), e.g., bi-directional feedback with its collaborators to create new things.

3. Human-centered safe robot reinforcement learning framework

Our proposed human-centered SRRL framework, as shown in Figure 2, consists of three stages: safe exploration, safety value alignment, and safe collaboration. When a robot enters a new environment, it must explore it through safe action optimization and state modeling. Then, the robot has to learn the safety value alignment from the human-robot interaction. Finally, the robot should be able to achieve safe collaboration with humans. Interactive behaviors can help successfully achieve each of these three stages.

3.1. Safe exploration

A robot should explore and find helpful information when entering a new environment. During the exploration, the robot must ensure its and the environment's (including other agents') safety. Therefore, the first stage of human-centered SRRL, safe exploration, focuses on exploring new environments and getting information about the environment model through safe actions. Uncertainty about the environment makes it difficult to determine safe actions. For example, newborn infants sometimes end up putting harmful objects in their mouths during oral exploration. It is challenging to ensure safe actions using conventional RL methods

that learn new skills through trial and error. Especially with model-free RL, a robot cannot avoid a destructive action unless it has already tried it (or a similar action; Saunders et al., 2018).

Several methods have been proposed to address the challenge of safe exploration for RL (Garcia and Fernández, 2012; Liu et al., 2020; Bharadhwaj et al., 2021; Turchetta, 2021; Xiong and Diao, 2021; Liu P. et al., 2022; Liu Z. et al., 2022). For example, Garcia and Fernández (2012) introduced a safe exploration method, in which a predefined baseline policy is required, and the baseline policy is approximated by behavioral cloning methods (Anderson et al., 2000). Nonetheless, derived from the method, it would be hard to search for the optimal exploration policy, and the capabilities of the baseline policy could severely limit the method's performance. A conservative safety critic (Bharadhwaj et al., 2021) is proposed to guarantee safety with high probability during robot exploration processes. Garcia and Fernández (2012) introduced a smoother risk function for safe robot exploration, which can achieve monotonic reward improvement and ensure safety. This method needs a predefined baseline policy to explore the safe space. Liu P. et al. (2022) developed a safe exploration method for robot learning by constructing a constrained manifold. This method can guarantee safety for robot exploration using model-free RL. However, it requires an accurate robot model or a perfect tracking controller, which may hinder their method's real-world applications.

Interactive behaviors can be used to transfer expert knowledge from the human to the robot for safe exploration, just like infants are safeguarded by their parents. Human intervention has been investigated to avoid catastrophes in RL (Saunders et al., 2018), and human feedback can be a reference for RL to ensure robot safety during safe exploration (Frye and Feige, 2019). However, the amount of human labor required for complex real-world applications is prohibitive. The bi-directional feedback in interactive behaviors can reduce the human-time cost during safe exploration. On the one hand, the robot can actively query the human and provide explanations of its behavior. On the other hand, the human can also actively query the robot to learn about the robot's model, in addition to intervening for safe exploration.

3.2. Safety value alignment

In the second stage of safe robot learning, to train a robot to perform a task safely, we need to evaluate how well it performs in terms of its performance on safety. Whether it is a form of costs, rewards, or labels, we need some form of signals to guide the safety policies for robot learning. In some scenarios, the safety performance can be evaluated automatically, such as bumping into other vehicles of autonomous driving or breaking the arms in robot manipulation. In these cases, training signals can be straightforwardly defined for safe robot learning. Furthermore, to facilitate the safe deployment of agents in real-world tasks, agents also need to be compatible with users' ethical judgment. Taking automated vehicles as an example, should one autonomous vehicle cut in line to maximize its reward in achieving a goal? Instead of maximizing only the reward, the agents need the capacity to abide by human moral values, which is essential but lacks effort.

Current robot learning algorithms rely on humans to state these training signals (Gu et al., 2022b; Yuan et al., 2022) and assume that humans understand the dangers. For example, imitation learning inferences a reward function from human demonstrations; preference-based learning guides the robot based on human judgments. These classes of tasks involve "human" training signals. The related problem of how to align has been discussed in earlier literature (Christiano et al., 2018; Leike et al., 2018; Kazantzidis et al., 2022; Liu R. et al., 2022) on how to align agents with user intentions, in which meaningful training signals can be hard to obtain, due to the unpredictable long-term effect of the behaviors, or potential influence to other agents and environments in large multi-agent systems.

Designing AI agents that can achieve arbitrary objectives, such as minimizing some cost or penalties, can be deficient in that the systems are intrinsically unpredictable and might result in negative and irreversible outcomes for humans. In the context of interactive learning, we consider how a robot can behave safely or align with the user's intentions whilst maintaining safety under interactive behaviors with humans, and we frame this as the **safety value alignment** problem: *how to create robots that behave safely and align with the human's intentions?*

Interactive behaviors allow agents to infer human values. While agents infer human values from their feedback, bi-directional feedback enables the agent to explain its decision-making process. One early attempt (Yuan et al., 2022) studied bi-directional communication in tabular-based navigation tasks without considering more practical scenarios. The next step aims to study the generalization of such results to large-scale problems via more efficient algorithms.

3.3. Safe human-robot collaboration

The third stage of our safe-robot learning framework aims to accomplish safe physical and cognitive collaboration between robots and humans. Collaboration has enabled humans to achieve great evolutionary success. Therefore, safe human-robot collaboration is essential for successful human-robot co-existence.

The four main categories of human-robot collaboration tasks explored in the literature are collaborative assembly, object handling, object handovers, and collaborative manufacturing (Semeraro et al., 2023). RL has been used for tuning impedance controllers in physical human-robot collaboration tasks such as lifting objects (Roveda et al., 2020) and guided trajectory following (Modares et al., 2015). However, these works evaluated the controllers in simplified scenarios and did not evaluate the generalizability of the learned policies. RL has also been used for performing robot-to-human object handovers (Kupcsik et al., 2018) and human-to-robot object handovers (Chang et al., 2022). Nevertheless, in some cases, the spatial generalizability of learned policies is low (Kshirsagar et al., 2021). Ghadirzadeh et al. (2020) used deep q-learning to generate proactive robot actions in a human-robot collaborative packaging task. However, they only evaluated a specific task scenario with a highly engineered reward function. Also, they did not test the trained policy in the real world and for different human participants than the training set.

Deep RL methods have been applied in real-world learning scenarios for tasks like quadrupedal walking, grasping objects, and varied manipulation skills (Ibarz et al., 2021). One of the

desired features of these works is the ability to perform training with little or no human involvement. However, scenarios of human-robot collaboration typically involve multiple humans in the robot's learning process. Multi-agent RL methods such as self-play or population-play do not perform very well with human partners (Carroll et al., 2019). One proposed solution called Fictitious Co-Play (FCP) involves training with a population of self-play agents and their past checkpoints taken throughout training (Strouse et al., 2021). However, FCP was evaluated only in a virtual game environment.

Interactive RL (IRL) approaches involve a human-in-the-loop to guide the robot's RL process. IRL has been applied to various human-computer interaction scenarios (Arzate Cruz and Igarashi, 2020). In addition, human social feedback in the form of evaluation, advice, or instruction has also been utilized for several robot RL tasks (Lin et al., 2020). However, more research is needed toward utilizing IRL for safe human-robot collaboration. Also, while some works have explored non-verbal cues to express the robot's uncertainty during the learning process (Matarese et al., 2021), most existing IRL approaches do not involve feedback from robots to humans. As depicted in Figure 2, evaluative feedback from robots to humans could help improve human-robot collaboration.

4. Open challenges

In this section, we describe four key open challenges toward utilizing interactive behaviors for SRRL. These open challenges are related to the robustness, efficiency, transparency, and adaptability of SRRL.

1. How can the robot learn robust behaviors with potential human adversaries?
2. How to improve data efficiency of SRRL for effective utilization of interactive behaviors?
3. How to design "transparent" user interfaces for interactive behaviors?
4. How to enhance adaptability of SRRL for handling multiple scenarios of interactive behaviors?

The first challenge is to achieve robust SRRL with respect to unintentional or intentionally erroneous human conduct in interactive behaviors. In existing SRRL methods, it is often neglected that humans might misstate the safety signals. Also, due to the potential involvement of multiple humans with different values, robots need to learn to strike a balance between them. In the extreme case, adversaries may intentionally state their signal to mislead the training of robots to achieve malicious objectives. Training robust agents against such malicious users needs further research. Adversarial training may be useful to ensure safety in such scenarios (Meng et al., 2022). However, adversarial training is not yet ready for real-world robot learning (Lechner et al., 2021).

The second challenge is to improve the data efficiency of SRRL, given that real-world interactive behaviors are expensive. Data efficiency can determine how quickly robots learn new skills and adapt to new environments and how effectively interactive behaviors can be utilized in the learning process. The success of machine learning can be attributed to the availability of large datasets and simulation environments. Therefore, one possible solution to reduce the need for real-world interactive data is to build large datasets or simulations of interactive behaviors. For example, Lee et al. (2022) present a mixed reality (MR) framework in which humans can interact with virtual robots in virtual or augmented reality (VR/AR) environments. It can serve as a platform for collecting data in various human-robot interaction and collaboration scenarios. However, this framework suffers from the limiting aspects of MR, such as the inconvenience of wearable interfaces, motion sickness, and fatigue. Improvements in MR technology will be crucial for the widespread use of such MR environments.

The third challenge is to maintain transparency during SRRL. Transparency is important for the effective utilization of interactive behaviors. MacGlashan et al. (2017) provide empirical results to demonstrate human feedback and robot policy can be interactively influenced by each other, and indicate that the assumption is that the feedback from a human is independent of the robot's current policy, may be incorrect. Transparency can be achieved in the context of interactive behaviors by explaining the robot's decision-making process to the human and exposing the robot's internal state. Several solutions have been proposed to improve the explainability of robot RL (Hayes and Shah, 2017; MacGlashan et al., 2017; van der Waa et al., 2018; Likmeta et al., 2020; Atakishiyev et al., 2021a,b; Matarese et al., 2021). For example, Likmeta et al. (2020) introduced an interpretable rule-based controller for the transparency of RL in transportation applications. Nonetheless, the policy that the method provides may be too conservative, since it severely depends on the restricted rules. Matarese et al. (2021) present a method to improve robot behaviors' transparency to human users by leveraging emotional-behavior feedback based on robot learning progress. However, further research is needed toward communicating the robot's internal state. Non-verbal communication in the form of gazes and gestures can be leveraged to communicate the robot's internal state. For example, if the robot is uncertain about its decisions, it can show hesitation gestures.

The fourth challenge is to enhance the adaptability of safe robot learning to handle a variety of settings involving interactive behaviors. Particularly, a robot can encounter different environments. For instance, completely trusted environments, in which the environment information is known and the environment is stable; **U**nstable environments for which environment information is known; **U**ncertain environments, where the information about the environment is uncertain and partial; **U**nknown environments, in which the environment information is completely unknown. Safety can be ensured in trusted environments, e.g., robots can safely grasp an object in a static environment. However, in "3U" environments, ensuring SRRL is challenging. For instance, during sim-to-real transfer, the discrepancies between the simulation models and real-world models are inevitable (Mitsch and Platzer, 2016), and the real-world environments are replete with uncertain disturbances and unknown information, e.g., in a multi-agent system, guaranteeing each agent's safety may be difficult. Some works provide a potential direction to ensure multi-robot learning safety in unstable environments, for example, Multi-Agent Constrained Policy Optimization and Multi-Agent Proximal Policy Optimization Lagrangian (Gu et al., 2023). Nevertheless, the exploration

involving interactive behaviors between agents and environments can be intricate and time-intensive. The incorporation of human insights and value alignment in the exploratory phase is instrumental in enhancing the adaptability of these agents within a human-in-the-loop learning system. As we envisage future trajectories of research, a salient focus is the attainment of SRRL characterized by interactive behaviors in "3U" environments. In this vein, game theory (Fudenberg and Tirole, 1991) emerges as a pivotal tool, offering nuanced strategies and frameworks for optimizing agent-environment interactions. Concurrently, the integration of advancements in cognitive science is anticipated to play a quintessential role. Specific areas of interest encompass the optimization of information management protocols between humans and robotic agents and the augmentation of robotic cognitive faculties through the effective utilization of perception devices. These integrated approaches aim to engender a more seamless, efficient, and adaptive interaction paradigm, catalyzing enhanced performance and adaptability in complex, dynamic environments.

However, the exploration with interactive behaviors between agents and environments may be time-consuming. Involving human knowledge and value alignment in the exploration can improve its adaptability in a human-loop learning system. Future work to achieve SRRL with interactive behaviors in "3U" environments can leverage game theory (Fudenberg and Tirole, 1991) and advances in cognitive science, for instance, how to manage the information between humans and robots, and robots how to leverage perception devices to enhance its cognitive abilities.

5. Conclusion

Robots need to ensure safety while leveraging RL in human environments. However, conventional SRRL algorithms do not consider the mutual influence between the robot and the human. In this paper, we proposed a human-centered SRRL framework consisting of three stages: safe exploration, value alignment, and human-robot collaboration. We discussed how these stages can leverage mutual influence or bidirectional information transfer between the robot and the human through interactive behaviors. We also described four key open challenges related to the robustness, efficiency, transparency, and adaptability of SRRL for effective utilization of interactive behaviors.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SG: Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing. AK: Methodology, Writing—original draft, Writing—review & editing. YD: Methodology, Writing—original draft, Writing—review & editing. GC: Supervision, Writing—review & editing. JP: Project administration, Supervision, Writing—review & editing. AK: Funding acquisition, Project administration, Resources, Supervision, Conceptualization, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was partially supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3); this work was also supported by The Adaptive Mind, funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art.

Acknowledgments

We would like to thank Yaodong Yang for his suggestions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). "Constrained policy optimization," in *International Conference on Machine Learning*, 22–31.

Akametalu, A. K., Fisac, J. F., Gillula, J. H., Kaynama, S., Zeilinger, M. N., and Tomlin, C. J. (2014). "Reachability-based safe learning with Gaussian

processes," in *53rd IEEE Conference on Decision and Control*, 1424–1431. doi: 10.1109/CDC.2014.7039601

Anderson, C. W., Draper, B. A., and Peterson, D. A. (2000). "Behavioral cloning of student pilots with modular neural networks," in *ICML*, 25–32.

Arzate Cruz, C., and Igarashi, T. (2020). "A survey on interactive reinforcement learning: design principles and open challenges," in *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 1195–1209. doi: 10.1145/3357236.3395525

Atakishiyev, S., Salameh, M., Yao, H., and Goebel, R. (2021a). Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*.

Atakishiyev, S., Salameh, M., Yao, H., and Goebel, R. (2021b). Towards safe, explainable, and regulated autonomous driving. *arXiv preprint arXiv:2111.10518*.

Baumann, D., Marco, A., Turchetta, M., and Trimpe, S. (2021). "Gosafe: globally optimal safe robot learning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 4452–4458. doi: 10.1109/ICRA48506.2021.9560738

Berkenkamp, F., Moriconi, R., Schoellig, A. P., and Krause, A. (2016). "Safe learning of regions of attraction for uncertain, nonlinear systems with Gaussian processes," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, 4661–4666. doi: 10.1109/CDC.2016.7798979

Bharadhwaj, H., Kumar, A., Rhinehart, N., Levine, S., Shkurti, F., and Garg, A. (2021). "Conservative safety critics for exploration," in *International Conference on Learning Representations (ICLR)*.

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., et al. (2021). Safe learning in robotics: from learning-based control to safe reinforcement learning. *Annu. Rev. Control Robot. Auton. Syst*. 5, 411–444. doi: 10.1146/annurev-control-042920-020211

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., et al. (2019). "On the utility of learning about humans for human-ai coordination," in *Advances in Neural Information Processing Systems, Vol*. 32.

Chang, P.-K., Huang, J.-T., Huang, Y.-Y., and Wang, H.-C. (2022). "Learning end-to-end 6dof grasp choice of human-to-robot handover using affordance prediction and deep reinforcement learning," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*.

Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. (2018). "A Lyapunov-based approach to safe reinforcement learning," in *Advances in Neural Information Processing Systems, Vol. 31*.

Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., and Ghavamzadeh, M. (2019). Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*.

Christiano, P., Shlegeris, B., and Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*.

Du, Y., Han, L., Fang, M., Dai, T., Liu, J., and Tao, D. (2019). "LIIR: learning individual intrinsic reward in multi-agent reinforcement learning," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, 4403–4414.

Frye, C., and Feige, I. (2019). Parenting: safe reinforcement learning from human input. *arXiv preprint arXiv:1902.06766*.

Fudenberg, D., and Tirole, J. (1991). *Game Theory*. MIT Press.

Fulton, N., and Platzer, A. (2018). "Safe reinforcement learning via formal methods: toward safe control through proof and learning," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32*. doi: 10.1609/aaai.v32i1.12107

Gao, L., Schulman, J., and Hilton, J. (2022). Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*.

Garcia, J., and Fernández, F. (2012). Safe exploration of state and action spaces in reinforcement learning. *J. Artif. Intell. Res*. 45, 515–564. doi: 10.1613/jair.3761

Ghadirzadeh, A., Chen, X., Yin, W., Yi, Z., Björkman, M., and Kragic, D. (2020). Human-centered collaborative robots with deep reinforcement learning. *IEEE Robot. Autom. Lett*. 6, 566–571. doi: 10.1109/LRA.2020.3047730

Gu, S., Chen, G., Zhang, L., Hou, J., Hu, Y., and Knoll, A. (2022a). Constrained reinforcement learning for vehicle motion planning with topological reachability analysis. *Robotics* 11, 81. doi: 10.3390/robotics11040081

Gu, S., Kuba, J. G., Chen, Y., Du, Y., Yang, L., Knoll, A., et al. (2023). Safe multi-agent reinforcement learning for multi-robot control. *Artif. Intell*. 319:103905. doi: 10.1016/j.artint.2023.103905

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., et al. (2022b). A review of safe reinforcement learning: methods, theory and applications. *arXiv preprint arXiv:2205.10330*.

Han, L., Sun, P., Du, Y., Xiong, J., Wang, Q., Sun, X., et al. (2019). "Grid-wise control for multi-agent reinforcement learning in video game AI," in *International Conference on Machine Learning (ICML)*, 2576–2585.

Hayes, B., and Shah, J. A. (2017). "Improving robot controller transparency through autonomous policy explanation," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 303–312. doi: 10.1145/2909824.3020233

Hunt, N., Fulton, N., Magliacane, S., Hoang, T. N., Das, S., and Solar-Lezama, A. (2021). "Verifiably safe exploration for end-to-end reinforcement learning," in *Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control*, 1–11. doi: 10.1145/3447928.3456653

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., and Levine, S. (2021). How to train your robot with deep reinforcement learning: lessons we have learned. *Int. J. Robot. Res*. 40, 698–721. doi: 10.1177/0278364920987859

Kaushik, R., Arndt, K., and Kyrki, V. (2022). Safeapt: safe simulation-to-real robot learning using diverse policies learned in simulation. *IEEE Robot. Autom. Lett*. 7, 6838–6845. doi: 10.1109/LRA.2022.3177294

Kazantzidis, I., Norman, T. J., Du, Y., and Freeman, C. T. (2022). "How to train your agent: active learning from human preferences and justifications in safety-critical environments," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1654–1656.

Knox, W. B., and Stone, P. (2009). "Interactively shaping agents via human reinforcement: the tamer framework," in *Proceedings of the Fifth International Conference on KNOWLEDGE CAPTURE*, 9–16. doi: 10.1145/1597735.1597738

Kochdumper, N., Krasowski, H., Wang, X., Bak, S., and Althoff, M. (2022). Provably safe reinforcement learning via action projection using reachability analysis and polynomial zonotopes. *arXiv preprint arXiv:2210.10691*. doi: 10.1109/OJCSYS.2023.3256305

Koller, T., Berkenkamp, F., Turchetta, M., and Krause, A. (2018). "Learning-based model predictive control for safe exploration," in *2018 IEEE Conference on Decision and Control (CDC)*, 6059–6066. doi: 10.1109/CDC.2018.8619572

Kong, X., Ma, K., Hou, S., Shang, D., and Xia, F. (2018). Human interactive behavior: a bibliographic review. *IEEE Access* 7, 4611–4628. doi: 10.1109/ACCESS.2018.2887341

Krasowski, H., Thumm, J., Müller, M., Wang, X., and Althoff, M. (2022). Provably safe reinforcement learning: a theoretical and experimental comparison. *arXiv preprint arXiv:2205.06750*.

Krasowski, H., Wang, X., and Althoff, M. (2020). "Safe reinforcement learning for autonomous lane changing using set-based prediction," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–7. doi: 10.1109/ITSC45102.2020.9294259

Kroemer, O., Niekum, S., and Konidaris, G. (2021). A review of robot learning for manipulation: challenges, representations, and algorithms. *J. Mach. Learn. Res*. 22, 1395–1476.

Kshirsagar, A., Hoffman, G., and Biess, A. (2021). Evaluating guided policy search for human-robot handovers. *IEEE Robot. Autom. Lett*. 6, 3933–3940. doi: 10.1109/LRA.2021.3067299

Kupcsik, A., Hsu, D., and Lee, W. S. (2018). "Learning dynamic robot-to-human object handover from human feedback," in *Robotics Research. Springer Proceedings in Advanced Robotics, Vol. 2*, eds A. Bicchi and W. Burgard (Cham: Springer).

Lechner, M., Hasani, R., Grosu, R., Rus, D., and Henzinger, T. A. (2021). "Adversarial training is not ready for robot learning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 4140–4147. doi: 10.1109/ICRA48506.2021.9561036

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, B., Brookshire, J., Yahata, R., and Samarasekera, S. (2022). "Towards safe, realistic testbed for robotic systems with human interaction," in *2022 International Conference on Robotics and Automation (ICRA)*, 11280–11287. doi: 10.1109/ICRA46639.2022.9811766

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. (2018). Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Li, X., and Belta, C. (2019). Temporal logic guided safe reinforcement learning using control barrier functions. *arXiv preprint arXiv:1903.09885*.

Likmeta, A., Metelli, A. M., Tirinzoni, A., Giol, R., Restelli, M., and Romano, D. (2020). Combining reinforcement learning with rule-based controllers for transparent and general decision-making in autonomous driving. *Robot. Auton. Syst*. 131, 103568. doi: 10.1016/j.robot.2020.103568

Lin, J., Ma, Z., Gomez, R., Nakamura, K., He, B., and Li, G. (2020). A review on interactive reinforcement learning from human social feedback. *IEEE Access* 8, 120757–120765. doi: 10.1109/ACCESS.2020.3006254

Liu, A., Shi, G., Chung, S.-J., Anandkumar, A., and Yue, Y. (2020). "Robust regression for safe exploration in control," in *Learning for Dynamics and Control*, 608–619.

Liu, P., Tateo, D., Ammar, H. B., and Peters, J. (2022). "Robot reinforcement learning on the constraint manifold," in *Conference on Robot Learning*, 1357–1366.

Liu, R., Bai, F., Du, Y., and Yang, Y. (2022). "Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning," in *Advances in Neural Information Processing Systems (NeurIPS)*.

Liu, Z., Guo, Z., Cen, Z., Zhang, H., Tan, J., Li, B., et al. (2022). On the robustness of safe reinforcement learning under observational perturbations. *arXiv preprint arXiv:2205.14691*.

Liu, Z., Yang, D., Wang, Y., Lu, M., and Li, R. (2023). EGNN: graph structure learning based on evolutionary computation helps more in graph neural networks. *Appl. Soft Comput*. 135, 110040. doi: 10.1016/j.asoc.2023.110040

Lou, X., Gu, J., Zhang, J., Wang, J., Huang, K., and Du, Y. (2023). "Pecan: leveraging policy ensemble for context-aware zero-shot human-AI coordination," in *Proceedings of the 22st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1654–1666.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., et al. (2017). "Interactive learning from policy-dependent human feedback," in *International Conference on Machine Learning*, 2285–2294.

Marco, A., Baumann, D., Khadiv, M., Hennig, P., Righetti, L., and Trimpe, S. (2021). Robot learning with crash constraints. *IEEE Robot. Autom. Lett*. 6, 1439–1446. doi: 10.1109/LRA.2021.3057055

Marvi, Z., and Kiumarsi, B. (2021). Safe reinforcement learning: a control barrier function optimization approach. *Int. J. Robust Nonlin. Control* 31, 1923–1940. doi: 10.1002/rnc.5132

Matarese, M., Sciutti, A., Rea, F., and Rossi, S. (2021). Toward robots' behavioral transparency of temporal difference reinforcement learning with a human teacher. *IEEE Trans. Hum. Mach. Syst*. 51, 578–589. doi: 10.1109/THMS.2021.3116119

Meng, J., Zhu, F., Ge, Y., and Zhao, P. (2022). Integrating safety constraints into adversarial training for robust deep reinforcement learning. *Inform. Sci*. 619, 310–323. doi: 10.1016/j.ins.2022.11.051

Mitsch, S., and Platzer, A. (2016). Modelplex: verified runtime validation of verified cyber-physical system models. *Formal Methods Syst. Des*. 49, 33–74. doi: 10.1007/s10703-016-0241-z

Modares, H., Ranatunga, I., Lewis, F. L., and Popa, D. O. (2015). Optimized assistive human-robot interaction using reinforcement learning. *IEEE Trans. Cybern*. 46, 655–667. doi: 10.1109/TCYB.2015.2412554

OpenAI (2023). *Chatgpt*. Available online at: https://openai.com/blog/chatgpt/ (accessed January 20, 2023).

Reddy, S., Dragan, A., Levine, S., Legg, S., and Leike, J. (2020). "Learning human objectives by evaluating hypothetical behavior," in *International Conference on Machine Learning*, 8020–8029.

Roveda, L., Maskani, J., Franceschi, P., Abdi, A., Braghin, F., Molinari Tosatti, L., et al. (2020). Model-based reinforcement learning variable impedance control for human-robot collaboration. *J. Intell. Robot. Syst*. 100, 417–433. doi: 10.1007/s10846-020-01183-3

Saunders, W., Sastry, G., Stuhlmüller, A., and Evans, O. (2018). "'Trial without error: towards safe reinforcement learning via human intervention," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2067–2069.

Semeraro, F., Griffiths, A., and Cangelosi, A. (2023). Human-robot collaboration and machine learning: a systematic review of recent research. *Robot. Comput. Integr. Manufact*. 79:102432. doi: 10.1016/j.rcim.2022.102432

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* 362, 1140–1144. doi: 10.1126/science.aar6404

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., et al. (2020). "Learning to summarize with human feedback," in *Advances in Neural Information Processing Systems, Vol*. 33, 3008–3021.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., and Everett, R. (2021). "Collaborating with humans without human data," in *Advances in Neural Information Processing Systems, Vol. 34*, 14502–14515.

Sui, Y., Gotovos, A., Burdick, J., and Krause, A. (2015). "Safe exploration for optimization with Gaussian processes," in *International Conference on Machine Learning*, 997–1005.

Sui, Y., Zhuang, V., Burdick, J., and Yue, Y. (2018). "Stagewise safe Bayesian optimization with Gaussian processes," in *International Conference on Machine Learning*, 4781–4789.

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Tamar, A., Di Castro, D., and Mannor, S. (2012). "Policy gradients with variance related risk criteria," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 1651–1658.

Thomaz, A. L., and Breazeal, C. (2006). "Reinforcement learning with human teachers: evidence of feedback and guidance with implications for learning performance," in *Proceedings of the 21st National Conference on Artificial Intelligence, Vol. 1*, 1000–1005.

Turchetta, M. (2021). *Safety and robustness in reinforcement learning* (Ph.D. thesis). ETH Zurich, Zürich, Switzerland.

Turchetta, M., Berkenkamp, F., and Krause, A. (2016). "Safe exploration in finite Markov decision processes with Gaussian processes," in *Advances in Neural Information Processing Systems, Vol. 29*.

Turchetta, M., Berkenkamp, F., and Krause, A. (2019). "Safe exploration for interactive machine learning," in *Advances in Neural Information Processing Systems, Vol. 32*.

van der Waa, J., van Diggelen, J., van den Bosch, K., and Neerincx, M. (2018). "Contrastive explanations for reinforcement learning in terms of expected consequences," in *Proceedings of the Workshop on Explainable AI on the IJCAI Conference, Vol*. 37 (Stockholm).

Wachi, A., Sui, Y., Yue, Y., and Ono, M. (2018). "Safe exploration and optimization of constrained MDPS using Gaussian processes," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32*. doi: 10.1609/aaai.v32i1.12103

Wang, Y., Liu, Z., Xu, J., and Yan, W. (2022). Heterogeneous network representation learning approach for ethereum identity identification. *IEEE Trans. Comput. Soc. Syst*. 10, 890–899. doi: 10.1109/TCSS.2022.3164719

Xiong, H., and Diao, X. (2021). Safety robustness of reinforcement learning policies: a view from robust control. *Neurocomputing* 422, 12–21. doi: 10.1016/j.neucom.2020.09.055

Yu, D., Zou, W., Yang, Y., Ma, H., Li, S. E., Duan, J., et al. (2022). Safe model-based reinforcement learning with an uncertainty-aware reachability certificate. *arXiv preprint arXiv:2210.07553*.

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., et al. (2022). In situ bidirectional human-robot value alignment. *Sci. Robot*. 7, eabm4183. doi: 10.1126/scirobotics.abm4183

Zhao, J., and Lv, Y. (2023). Output-feedback robust tracking control of uncertain systems via adaptive learning. *Int. J. Control Autom. Syst*. 21, 1108–1118. doi: 10.1007/s12555-021-0882-6

Zhao, X., Gu, C., Zhang, H., Yang, X., Liu, X., Liu, H., et al. (2021). "Dear: deep reinforcement learning for online advertising impression in recommender systems," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35*, 750–758. doi: 10.1609/aaai.v35i1.16156