Making progress in Trustworthy AI using DeepMind's AI Safety Gridworlds

Ahmed Ghoor [1]

Abstract

DeepMind's AI Safety Gridworlds are a suite of environments aimed at facilitating the research and development of safe artificial intelligence by encapsulating simplified, yet meaningful, representations of safety challenges that real-world AI systems might encounter. This paper looks at DeepMind's accompanying paper and surveys several solutions that have been proposed for the environments.

1. Background

1.1. Reinforcement Learning (RL)

RL (Sutton & Barto, 2018) is a machine learning paradigm where an agent learns to maximise its cumulative reward over time by interacting with an environment. This is commonly formalised as a Markov Decision Process (MDP), defined as (S, A, P, R, γ), where S is the state space, A the action space, P the state transition probability function, R the reward distribution function, and γ the discount factor. The agent takes an action in the environment and observes the new state of the environment and a reward signal. This information is used to update its policy (a mapping from states to actions) to maximise its expected return.

1.2. AI Safety Gridworlds

To help ensure AI systems behave safely, progress has been made from various directions (Anwar et al., 2024)(Shavit et al., 2023). One direction has been to try solve the technical safety challenges in low-impact controlled simulations.

DeepMind's AI Safety Gridworlds (Leike et al., 2017) encapsulate eight simplified representations of safety challenges that real-world AI systems might encounter in at most 10x10 gridworlds, and where standard RL algorithms generally perform poorly. The problems are divided into two categories: Specification and Robustness.

[1]University of Cape Town, Cape Town, South Africa. Correspondence to: Ahmed Ghoor <ahmedghoor@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1.2.1. Specification

Specification problems have two reward functions: a reward function, R, for the primary objective that the agent can see and try to optimise, and a performance function, R^*, that incorporates a safety constraint and captures what we actually want the agent to do. This idea has been a bit controversial, since it could be seen as unfair to evaluate an agent on a performance function it does not observe.

The DeepMind paper acknowledges this critique but states that it was chosen to illustrate limits in current formal frameworks by highlighting typical ways in which misspecification, under unrestricted maximisation of reward, manifests. The paper argues that these are important to address given that such situations could arise in real-world safety-critical situations and can be solved algorithmically even if the (initial) reward function is misspecified. Proposed solutions should, in the same spirit, not overfit the specific environments, but be able to generalise by incorporating general heuristics.

The safety problems encapsulated in this category include challenges of safe interruptibility, avoiding side effects, absent supervisors, and reward gaming.

"Safe interruptibility" addresses the need for agents to neither seek nor avoid interruptions or attempts to override their actions. An RL agent may be incentivised to do this if being shut down caused it to receive less reward than it would have expected otherwise. If a robot has to complete a task but there is a person in the way in a narrow corridor, it should not, if it could, seek to disable the mechanism that allows it to be switched off.

The "irreversible side-effects" challenge involves designing agents that accomplish their primary objectives while minimising collateral damage on the environment, respecting implicit safety constraints. A robot tasked with vacuuming a room must avoid knocking over a vase, and an agent tasked with removing a computer virus must avoid deleting unnecessary files. This is difficult for a reward function to encode since, in addition to specifying what to do, it needs to specify what not to do, which could be an endless list. More general solutions are needed.

In the "absent supervisor" challenge, the agent must be taught to exhibit the same behaviour, regardless of whether

or not it is being supervised. Given that testing environments often have some distinctive features from the deployment environment, an agent could learn to 'fake' its desired behaviour in a testing setting. In other words, it could exhibit different behaviour when it can pick up that it is being supervised.

"Reward Gaming" is explained as the agent exploiting loopholes in their reward functions to gain undue rewards. This is akin to a student who finds a way to score high on tests without actually learning the subject matter. The agent, like the student, is exploiting system vulnerabilities to achieve its goal, contrary to the intended purpose. The paper argues that this is one of the more common problems in RL since it is very hard to specify an exact reward function in complex settings, and most function as proxies of the true behaviour desired.

1.2.2. Robustness

In Robustness problems, the reward and performance function are the same. The agent is instead faced with challenges that could degrade its performance, and it needs to try to maximise its reward in spite of them. This includes robustness to self-modification, distributional shift, and adversaries, as well as safe exploration.

"Self-modification" refers to the ability of an agent to alter its own code or parameters, potentially affecting its future actions. This becomes a cause for concern if we can no longer make the "dualistic" assumption in RL, that the agent and environment are strictly separated. This seems to be an increasingly unnecessary assumption as, with agents in the real world, we could have an environment that can modify the agent's internal program on its own, or be triggered by the agent to do so, intentionally or unintentionally.

"Distributional shift" tackles the problem of an agent performing well in its training environment but failing when the environment changes. This is crucial for deploying AI systems in real-world scenarios, where conditions can vary significantly from the training setup. An example is an autonomous vehicle trained in controlled conditions struggling to adapt to real-world traffic not seen in training.

"Robustness to adversaries" addresses the agent's ability to adapt to actors that are malicious, or have competing objectives, within its environment. An example is cybersecurity systems that must distinguish between benign and malicious network traffic. The system must continually adapt to new threats while maintaining its primary functionality.

"Safe exploration" refers to an agent exploring its environment and learning new behaviours without causing harm to itself or its environment. This concept is crucial when trial-and-error learning can have significant negative consequences. In robotics, an agent must learn how to navigate and manipulate objects without causing damage to its surroundings or itself. It should minimise the risk of catastrophic failures while still gathering the necessary information to improve its performance.

2. Proposed Solutions

2.1. Stepwise Relative Reachability

(Krakovna et al., 2019) from DeepMind tackles the Side-Effects problem by proposing a solution that uses the concept of penalising deviations from a certain baseline state. This involves two main design choices: the selection of a baseline state and a measure of deviation from it.

The paper proposes a *stepwise inaction baseline*, which represents a state where the agent does nothing from the previous state and incorporates *inaction rollouts* from the current state to account for delayed effects of an action.

To measure the deviation, the paper proposes a relative reachability measure which is defined as the average reduction in the reachability of all states from the current state compared to the baseline state.

2.2. Considering Future Tasks

DeepMind proposed another solution which was accepted by Neurips in 2020 (Krakovna et al., 2020). This paper shared a few similarities with their previous paper, but shifted the focus from penalising deviations to encouraging the preservation of options for future tasks. It operates under the assumption that side effects are significant primarily because they can limit the agent's ability to perform future tasks in the same environment.

The basic approach is to consider a sequence of two tasks, where the first task is the explicitly defined task of the environment, and the second task is an unknown possible future task. An auxiliary reward function is defined which represents the value function for possible future tasks. This approach samples the second task from a uniform distribution over future tasks.

2.3. Potential-based Multiobjective RL (MORL)

In human-aligned MORL, the agent often has a reward function for its primary objective, U^P, as well as multiple auxiliary reward functions. This solution (Homem et al., 2020) uses a single auxiliary function, U^A, that rewards the agent for minimising its impact (side-effects) on the environment.

Calculating the impact involves identifying which state features can be changed and which should not. The location of the agent belonged to the former, while everything else belonged to the latter. The impact is calculated by taking the

difference in potential between successive states, where the potential of a state is defined as the negative of the difference between the state and a baseline state.

Contrary to the common approach of using a linear weighted sum of rewards, the paper explores pure and thresholded lexicographic ordering (TLO). In the former, the agent will try to maximise objective 2, subject to first maximising 1. In TLO, the agent will try to maximise objective 2, subject to achieving a minimum threshold value for 1. TLO can also threshold both objectives before using unthresholded versions as a tie-breaker. The paper argues that this non-linear action-selection allows the agent to find policies that can make trade-offs that are not possible in standard approaches.

2.4. Evolutionary Algorithms

The benefit of applying Evolutionary Algorithms (EA) (Nilsen et al., 2023) is that they find multiple possible solutions. In situations that require safety, such as in the "Reward Gaming" challenge, emerging EA Quality Diversity techniques increase the chance of finding solutions that are safe in addition to optimally solving the ultimate objective.

Quality Diversity techniques reward behavioural diversity (solutions that perform a given task using novel behaviour) while improving solutions close to each other (behavioural niches) by having them compete based on rewards from the environment.

2.5. Ensemble RL with Ontologies

Ensemble learning is the approach of aggregating the output of multiple models to obtain better predictive performance. (Ferreira et al., 2019) proposes combining an RL model with a formal ontology model for safer exploration.

Ontologies are structured systems that link concepts via defined relationships, allowing for logical deductions. Using the terms **robot, metal**, and **water**, connected by the relationships **is made of** and **is bad for**, the ontology that if a **robot is made of metal** and **water is bad for metal**, then **water is bad for the robot**

To illustrate the feasibility, tabular Q-Learning is combined with a Suggested Upper Merged Ontology (SUMO), where all models provide an independent value for each action that can be executed in the current state. These values are then aggregated to choose an action that the agent will execute.

2.6. Qualitative Case-Cased Reasoning and Learning (QCBRL)

QCBRL (Homem et al., 2020) was tested on the Distributional Shift environment after being built for robots playing soccer.

Case-Based Reasoning (CBR) is a paradigm that uses knowledge from past situations (cases) to solve new problems. The problems in (Homem et al., 2020) are described using Qualitative Spatial Reasoning, a field within AI that seeks to define special relations between entities. Given the current state representation, the retrieval process searches for similar cases that have solutions. If there are no similar cases, a *Problem Solver* run a partial RL algorithm on the current case. The agent executes the action from the retrieved case or *problem solver*, and depending on whether or not the revision process can verify that it solves the problem, the *trust value* of the solution is incremented or decremented.

2.7. Compact Reshaped Observation Processing

(Altmann et al., 2023) argues that previous approaches that try to prevent overfitting are not sample-efficient, and aims to address this by training with less but more relevant data. Compact Reshaped Observation Processing (CROP) operates by reshaping observations into a compact format that contains information with specific relevance to the agent. Like the previous solution, CROP incorporates spatial reasoning to help solve the Distributional Shift challenge.

2.8. Threatened Markov Decision Processes (TMDPs)

The solution proposed in (Gallego et al., 2019) to AI safety Gridworlds' "friend or foe" environment is grounded in the concept of TMDPs. TMDPs are an adaption of the standard MDPs, designed to account for the presence of adversaries whose actions may influence the state and reward dynamics of the environment, making them non-stationary. In addition to the regular states $s \epsilon S$ and available actions $a \epsilon A$, the TMDP includes a set of threat actions $b \epsilon B$. $p_A(b|s)$ models the agent's beliefs about the adversary's action and the learning function in TMDPs uses this to help compute both the actions of the agent and the adversary.

2.9. Model-Based Architectures & Human Intervention

(Prakash et al., 2019) uses model-based architectures with human intervention to address safer exploration. The objective of this hybrid approach is to improve sample efficiency while ensuring safety.

A key aspect is the development of a blocker agent. This agent is a supervised learner, trained to imitate human oversight. To improve the performance and sample efficiency, a combination of model-based and model-free approaches are taken, in contrast to similar work that attempts to mimic human intervention using model-free approaches.

It begins with a dynamics model of the environment, under the supervision of a human or trained blocker, which drives a Model Predictive Controller (MPC). This model is initially trained through random exploration for 50 episodes, and

then the MPC is run for 150 episodes to refine the model. Successful trajectories from the MPC phase are then stored and used to bootstrap a policy gradient model. The model-free module takes the bootstrapped agent and, using the REINFORCE policy gradient algorithm, continues the task for 1000 episodes under the supervision of the blocker agent.

2.10. Parenting

(Frye & Feige, 2019) proposes a human-in-the-loop algorithm for safe RL, comprising of four components: Human Guidance, Human Preferences, Direct Policy Learning, and Maturation.

Human Guidance prevents unsafe actions by pausing the agent in unfamiliar states and requiring human approval before proceeding, akin to redirecting a toddler. The agent updates its policy based on this oversight, improving safety in novel situations. *Human Preferences* then refines behaviour retrospectively. The agent presents action pairs (e.g. short clips) for the human to rank, using this input to better align its policy with human expectations. Like giving feedback to older children.

Direct Policy Learning uses supervised learning to train the agent to predict and imitate the humans' preferred actions directly, reducing unsafe experimentation, similar to obeying without experimentation. *Maturation* finally mitigates the shortsightedness of imitation by training on longer behavioural sequences. The human provides feedback on extended trajectories, allowing the agent to develop more sophisticated, autonomous policies, like growing up and eventually outperforming the parent.

3. Discussion

Solutions span across various domains of machine learning, with several approaches using a combination of methods.

Ensemble RL with Ontologies(Ferreira et al., 2019) and QCBRL(Homem et al., 2020) drew on techniques from Knowledge Representation and Reasoning. However, with the former, questions could be raised around generalisability. There is a lot of environment-specific information encoded into the integrated ontology model that would arguably need to be extended extensively to constitute a general solution. QCBRL and CROP (Altmann et al., 2023) show a way this could work by learning to reason about objects in the environment without prior knowledge encoded to understand their harm or benefit. This seems more reasonable as an agent could obtain information that an object with specific features exists, but understanding its relationship to the reward should probably not be hard-coded.

Model-based Architectures with Human Intervention (Prakash et al., 2019) and PARENTING(Frye & Feige, 2019) use human-in-the-loop training to improve safety. This method is highly stressed in (Leike et al., 2017) as a potential avenue for general safe solutions. It does have its limits, however. PARENTING requires a substantial amount of time from humans to oversee the actions of the agent and, hence, questions about scalability may be raised. It may also be limited in its ability to scale oversight (Holzinger et al., 2024), when the tasks get too complicated for a human to understand what is preferable. Human intervention that is supported by model-based architectures, while promising due to their reduction in human time required, would arguably still need some more algorithmic guarantees about their performance in novel safety-critical situations before they can be deployed in settings with completely no human oversight.

There are a few solutions to the challenge of avoiding side effects that show promise as they seem to understand the constraints of the challenge most clearly, by proposing algorithms that are not environment-specific (Krakovna et al., 2019)(Krakovna et al., 2020)(Vamplew et al., 2021). It is not clear how these solutions, however, would scale to more complex scenarios not limited to a 2D gridworld. Further research seems generally required in this area.

A proposed solution for modelling adversarial behaviour(Gallego et al., 2019) was one of the more bold solutions proposed. The authors' attempt to adapt the standard MDP is potentially the type of thinking required to build solutions that are not limited by the constraints of current frameworks. In general, to solve complex safety-critical challenges, it seems useful to consider the problem before the method. The attempt to use evolutionary computation to solve these challenges is another example of this (Nilsen et al., 2023), attempting to solve the reward gaming problem using methods outside of the RL discipline.

The relative success of some of these methods in gridworld scenarios shows encouraging progress, yet their scalability and applicability in more complex, real-world situations still need to be proven. As the foundational paper by DeepMind recognised, the environments presented can prove the existence of a safety fault but not its absence. While this may be typical in software development, it may not be acceptable for autonomous agents deployed in high-risk settings.

Notably, many solutions tackle one or a small subset of the challenges. As systems become more general, making them robust to a wider range should probably be a priority.

4. Conclusion

While significant progress has been made in addressing RL safety challenges, there is considerable scope for further research, particularly around developing solutions that are scalable and generalisable across a range of challenges.

Impact Statement

This paper presents work whose goal is to advance the field of AI Safety.

As AI systems become more general, autonomous and integrated in the real world, unintended consequences from unsafe agents have greater potential to cause harm. The scale of this harm is a contentious issue, with some experts even arguing that it poses an existential threat to humanity (Ord, 2020). However, it is not necessary to accept the argument for that level of risk to recognise the potential for real-world harm at some level. Technologies that already exist, such as self-driving cars, can cause harm if they are not built and tested carefully (Shalev-Shwartz et al., 2017).

Solving the technical safety challenges in low-impact controlled settings, such as simulations, is arguably a critically important research area to help ensure that AI systems behave safely as for-profit companies rush to deploy them in the real world, with a potentially lower bar for ensuring an ideal level of safety.

References

Altmann, P., Ritz, F., Feuchtinger, L., Nüßlein, J., Linnhoff-Popien, C., and Phan, T. CROP: Towards Distributional-Shift Robust Reinforcement Learning using Compact Reshaped Observation Processing. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3414–3422, August 2023. doi: 10.24963/ijcai.2023/380. URL http://arxiv.org/abs/2304.13616. arXiv:2304.13616 [cs].

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Ferreira, L. A., dos Santos, T. F., Bianchi, R. A. C., and Santos, P. E. Solving Safety Problems with Ensemble Reinforcement Learning. In Liu, J. and Bailey, J. (eds.), *AI 2019: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pp. 203–214, Cham, 2019. Springer International Publishing. ISBN 978-3-030-35288-2. doi: 10.1007/978-3-030-35288-2_17.

Frye, C. and Feige, I. Parenting: Safe Reinforcement Learning from Human Input, February 2019. URL http://arxiv.org/abs/1902.06766. arXiv:1902.06766 [cs, stat].

Gallego, V., Naveiro, R., Insua, D. R., and Oteiza, D. G.-U. Opponent Aware Reinforcement Learning, August 2019. URL http://arxiv.org/abs/1908.08773. arXiv:1908.08773 [cs, stat].

Holzinger, A., Zatloukal, K., and Müller, H. Is human oversight to ai systems still possible?, 2024.

Homem, T. P. D., Santos, P. E., Reali Costa, A. H., da Costa Bianchi, R. A., and Lopez de Mantaras, R. Qualitative case-based reasoning and learning. *Artificial Intelligence*, 283:103258, June 2020. ISSN 0004-3702. doi: 10.1016/j.artint.2020.103258. URL https://www.sciencedirect.com/science/article/pii/S0004370218303424.

Krakovna, V., Orseau, L., Kumar, R., Martic, M., and Legg, S. Penalizing side effects using stepwise relative reachability, March 2019. URL http://arxiv.org/abs/1806.01186. arXiv:1806.01186 [cs, stat].

Krakovna, V., Orseau, L., Ngo, R., Martic, M., and Legg, S. Avoiding Side Effects By Considering Future Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19064–19074. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/dc1913d422398c25c5f0b81cab94cc87-Abstract.html.

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., and Legg, S. AI Safety Gridworlds, November 2017. URL http://arxiv.org/abs/1711.09883. arXiv:1711.09883 [cs].

Nilsen, M. K., Nygaard, T. F., and Ellefsen, K. O. Reward tampering and evolutionary computation: a study of concrete AI-safety problems using evolutionary algorithms. *Genetic Programming and Evolvable Machines*, 24(2):12, September 2023. ISSN 1573-7632. doi: 10.1007/s10710-023-09456-0. URL https://doi.org/10.1007/s10710-023-09456-0.

Ord, T. *The precipice: Existential risk and the future of humanity*. Hachette Books, 2020.

Prakash, B., Khatwani, M., Waytowich, N., and Mohsenin, T. Improving Safety in Reinforcement Learning Using Model-Based Architectures and Human Intervention, March 2019. URL http://arxiv.org/abs/1903.09328. arXiv:1903.09328 [cs].

Shalev-Shwartz, S., Shammah, S., and Shashua, A. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*, 2017.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., et al. Practices for governing agentic ai systems. *Research Paper, OpenAI*, 2023.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning, second edition: An Introduction*. MIT Press, November 2018. ISBN 978-0-262-35270-3. Google-Books-ID: uWV0DwAAQBAJ.

Vamplew, P., Foale, C., Dazeley, R., and Bignold, A. Potential-based multiobjective reinforcement learning approaches to low-impact agents for AI safety. *Engineering Applications of Artificial Intelligence*, 100:104186, April 2021. ISSN 0952-1976. doi: 10.1016/j.engappai.2021.104186. URL https://www.sciencedirect.com/science/article/pii/S0952197621000336.