Reflexive Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose the method of reflexive reinforcement learning (RRL), an extension of self-motivated reinforcement learning that enables the generation of autonomous learning behaviours in the absence of explicit external rewards and where actions can be taken in order to obtain information about the environment. This framework includes entropy reduction of distribution over states, reward functions and actions, as well as integration of non-task-specific information such as the theory of empowerment. RRL is a tool with which an agent can use the current representation of the value, of its state, and of the environment to identify regions of a state space that are interesting through a variety of simple, and distinct reward schemes. We continue to show that RRL is naturally applicable in the multi-agent domain, where agents are shown to be able to identify socially valuable regions of the state space.Finally, we discuss the Markov property in the context of RRL and the demonstrate the benefits of its integration with traditional reinforcement learning approaches.

1 Introduction

Intrinsic motivation [5] has been proposed in order to realise reinforcement learning (RL) in the absence of external rewards. It has been characterised [13] to follow either or both of two main goals:

- Encouraging exploration of unseen or interesting states, such as states that are central or risky, represent momentous decisions, or are at the boundary of a currently known domain;
- Reduction of uncertainty by improvement of the prediction of consequences of actions or of the estimation of (joint) probability distributions over states, actions and rewards.

These two goals may not be achievable at the same time, and various methods to negotiate the goals have been studied in the context of autonomous learning and self-organisation of behaviour [3, 6, 11]. It was shown that options or elementary behaviours can be created by these methods that can also be extended for the generation of self-organised plans [4].

We are proposing a new view on intrinsic motivated RL (IMRL) which can summarised by the maxim: *Act such that inverse reinforcement learning works well.* Inverse RL [12] aims at the reconstruction of the reward distribution based on the observed behaviour of a supposed RL agent and thus at the identification of reasons for its behaviour. In this way, we will refer to inverse RL not as a problem to be solved, but as an easily solvable problem to be posed: the self-motivated agent aims at generating a behaviour in its environment that is particularly clear with respect to underlying reasons given the specificity and stochasticity of the environment.

The potential advantages of this approach are obvious: The agent will tend to produce modes of behaviour that are both robust and informative, i.e. it will try to find a compromise to the mentioned dilemma of intrinsically motivated behaviour. The intrinsically generated behaviour may be further modulated by later or intermittently available external rewards, but will be biased towards readily analysable and explainable policies.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

In inverse RL the trivial solution of interpreting the observed behaviour as random needs to be avoided by an additional criterion such as maximal entropy [20]. In the same way, the trivial case of producing trivial behaviours in IMRL is avoided. As there are many aspects to RL and inverse RL, a wide area of *reflexive* RL (RRL) is conceivable that we will trace out by means of a few special cases.

The rest of this paper is organised as follows: After discussion of prior relevant work, we specify the reinforcement learning problem that we are going to study here in Sect. 4, where we also describe the theory underpinning the robotic implementation. Sect. 4 outlines the algorithm that we used, as well as a description of the experimental setup that was utilised. The results of the experiments are provided and analysed also in Sect. 5. The conclusions of the work and future work are given in Sect. 6.

2 Background

2.1 Reinforcement Learning

RL aims at finding solutions to the Bellman problem of reconstructing a value function that obeys

$$V^*(s) = \max_a \left(R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right) \tag{1}$$

in the case of deterministically chosen maximising actions in a Markovian decision problem, where s is the current and s' the subsequent state, and action a is executable in state s. This includes two aspects: the maximisation of the utility over some time horizon and the identification and control of trajectories towards high-utility states.

Classical RL algorithms have focused mainly on the maximisation problem and did not provide theoretical solutions to the exploration problem, also many heuristic approaches exist. Recently, and possibly inspired by the free-energy paradigm [7] but essentially going back to Refs. [17] and [3], the RL problem has been considered in an information-theoretical formulation [18], where information gain can be systematically incorporated into the utility function

It remains open in this approach how to choose realistic global priors, how to avoid local optima, and what metrics are to be used for the computationally demanding function approximation of probability distributions for policy, rewards, and state transitions, and how to resolve the relationship between information available in the state space and the utility, i.e. the problem of how reward-related gains and information-related costs are cleared. The latter problem is solvable in general as the tipping point of this balance depends on the environment and the representation thereof in the algorithm [16], but with that the generality of the information-theoretic approach is limited.

For simplicity, we will stick here with the discrete case and discuss generalisations in the supplementary material.

2.2 Inverse RL

In an RL problem with discrete actions $a \in \mathcal{A}$ we can try to recover the value function given action-conditioned state transitions $P^a(s', s)$ [12] such that

$$\sum_{s'} V\left(s'\right) P\left(s'|s,a)\right) \geq \sum_{s'} V\left(s'\right) P\left(s'|s,b\right) \tag{2}$$

for any $b \in \mathcal{A}$ different from a. If the reward depends only in the current state, then the relation $R(s) = \left(I - \gamma P^a\right) V(s)$ can be used on order to reconstruct the structure of the reward distribution. Because $V \equiv 0$ (or $R \equiv 0$) is a trivial solution of Eq. 2, additional conditions need to be imposed such as a maximal difference between optimal and second-best action [12] or, more generally, maximum entropy reward distribution [20] or adversarial schemes [14].

Inverse RL presupposes that the observed agent is already well-trained, such that its behaviour expressed the rewards structure; however, continuing to generate data after self-learning would not be meaningful in the present case. Additionally, the agent should not aim at exploiting a partcular inverse RL algorithm. Therefore, the agent does not aim at actually solving the inverse RL problem, and we can assume various stages of inverse RL in order to enable autonomous behaviour in the agent. So the agent should set itself rewards that:

- single out a particular state or that appear identical based on a given sensor configuration;
- have a spatial configuration and preproducibility to stick out the noise;
- single out states that are providing optimally controllable options;
- are compatible with a maximum entropy distribution of paths.

In order to make these points more clearly, we will discuss the concept of empowerment next.

2.3 Discounted empowerment

The concept of *empowerment* [11] is a way to express the value of a state without consideration of goal-oriented behaviour. It can be understood as a quantification of an agent's control over its environment or, likewise, of the freedom of choice of actions and of the level of reproducibility of a sequence of actions [15]. The original aim of this concept was not to motivate exploration, but instead to identify *preferred* states in an environment that is already known. If the agent is within the state s_t the n-step empowerment is defined based on mutual information

$$\mathfrak{E}_n\left(s_t\right)=\max_{\pi: s \rightarrow a} \mathcal{I}\left(s_{t+n} ; a_{t+n-1}, \ldots, a_t\right) \tag{3}$$

so that the task is to find a policy for which the mutual information between the next actions and the set of states is maximal.

Empowerment usually requires full prior information as well as the evaluation of all possible time series and states to determine which state or states are best for the agent to occupy over a given n-step time horizon. In this sense it is quite similar to POMDPs [9], but without any external rewards, apart from the information gains. In the context of RL, this level of computational complexity appears unnecessary because the precise value of $\mathfrak{E}_n$ (3) is largely irrelevant, alternatively an approximation of the empowerment can be produced iteratively by considering the entropy gain per step $\mathfrak{E}_1\left(s_t\right)$ and then summing over the time horizon specified by the RL discount factor γ.

$$\mathfrak{E}_\gamma\left(s_t\right)=\sum_{t=t_0}^{\infty} \gamma^{t-t_0} \mathfrak{E}_1\left(s_t\right) \tag{4}$$

The concept of γ-empowerment introduced in Eq. 4 is similar in information provided, but not equivalent to the original concept (3), because it does not have a crisp time horizon and also because it allows for different measures of the local empowerment $\mathfrak{E}_1\left(s_t\right)$ as long as the tendency of the agent to roam without restrictions is captured in an appropriate way, see Sect. 4.1.

3 Reflexive Reinforcement Learning

We propose *reflexive reinforcement learning* (RRL) as a more general approach that combines self-generated rewards and manipulation of the value function or policy in autonomous learning based on the additional principles:

- versatility and empowerment;
- social interaction and emotional learning;
- optimal internal representations;
- optimal substructure.

We are interested in extracting knowledge from these values, which has been done related to temporal fluctuations, but is perhaps more interesting if spatial variations are considered, as, in this way information directly related to the environment can be extracted, whereas temporal fluctuations mainly provide information about the learning process [8]. We will consider, in addition, qualitative effects of actions and loops which can also be used to characterise the environment. These approaches will be considered in more details in the following subsections. [2]

In the case of reflexive reinforcement learning the *reflexive component* informs state valuation through standard reinforcement learning. This additional component allows that we are able to switch between a variety of different components for different needs, as we will show in Sect. 5.

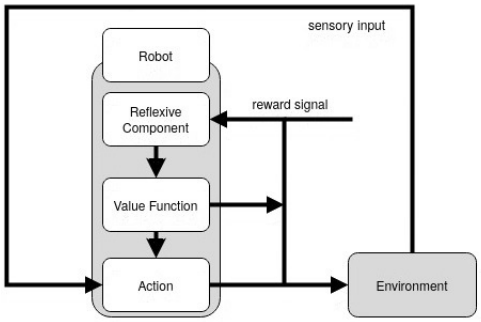


Figure 1: Schematic representation of RRL.

As seen in Fig. 1, the reflexive component will receive external rewards from the environment through the observation of the state, and this adjusted reward is used to inform the state valuation or valuations. In this manner it is possible for the agent to continue to receive information pertinent to potential tasks as it maintains its motivation to explore the environment through the different valuations during periods of no task or when in a state where it is lost.

There are impressively many properties of a problem that are relevant for solving the Bellman optimality equation but which are involve any external evaluative results. additional factors such as, *optimal substructure, hierarchy, controllability, predictability, smoothness, empowerment, Markovianity, the modelling of other agents*, among others. RRL is an extension of intrinsically motivated reinforcement learning that aims at utilising these sources of information that tend to be lost over the course of traditional reinforcement learning approaches.

RRL aims at consolidating the information that is observed over the course of the learning process to indicate regions in the state space that are suitable for continued learning, of high interest, or of high task likelihood, so that an agent is continually motivated to perform, where no task related information is currently visible. RRL utilises simple reflexive rewards related to easily observable quantities and features in a state space to create these valuations in a highly flexible approach.

4 Methods

4.1 Actions and Policy

Each of the tested agents can move in any of the four cardinal directions. The agent is unable to remain in the same state, with the exception that if it attempts to move into an obstacle or wall, on another agent, its position will remain unchanged.

The single agent variants exploration rate is such that $\varepsilon = 0.75$ in order to force the agent to use actions cautiously as errors can often not be corrected in the next or following steps at this level of randomness. The multi agent variant has agents which are identical to the single agent variants with two minor changes, firstly we reduced the exploration rate to $\varepsilon = 0.50$ to allow for an increase in the consistency of movement when agent rewards require complimentary behaviour.

As our aim here is mainly that of illustration if the principle of Reflexive RL (RRL), we opted to use a box function over the entire state action space rather than a reduced number of basis functions, with a traditional ε-greedy policy; however, the approach will work in such a space.

4.2 Rewards

When prioritising the maximisation of entropy, the agent received a reward of -1 if the agent collided with walls, and a reward bonus of

$$R(x,a) = \begin{cases} \mathcal{H}\left(x,a\right) - 1, & \text{if collision} \\ \mathcal{H}\left(x,a\right), & \text{else} \end{cases} \tag{5}$$

where

$$\mathcal{H}\left(x,a\right) = -\sum_{x'} p(x'|x,a)\log p(x'|x,a) \tag{6}$$

In the case of γ-empowerment, the agent is rewarded -1 for remaining in the same state across two subsequent time steps. Where we consider approaches that would lead to prediction error reduction or corner favouring, the agent is rewarded $+1$ if where the agent is in a state that is adjacent to one, or more than one wall respectively.

As we extend to the multi agent variant, we considered the task to be valuation of *social* or *mutual* space that the agents inhabit, as these will be areas of interest in multi agent systems with tasks requiring more than one agent, as such, we reward the agents each with $+1$ when they are within a Manhattan distance of 5 with other agents, and additional schemes in the supplementary matterial.

4.3 Environments

For our experimentation we opted to use four different arena types, set within a 21x21 environment, which exhibit different features through which to view the agents behaviours during entropy maximisation, prediction error minimisation, and the mixed behaviour of the two goals.

Each of the environments were selected based on the features they present. The empty arena was chosen as the base case. The second environment **(b)** presents a winding corridor which ends in a dead end, chosen to observe the effects over corridors of varying size and the effects of the surrounded end. Environment **(c)** observes the effect of large obstacles in the state space and the effects of irregular shapes on the algorithm. The final environment, environment **(d)** consists of a smaller room and a large room, chosen to observe the effects of differently sized regions of the state space, and observe the value the agent places on these objects.

5 Experiments

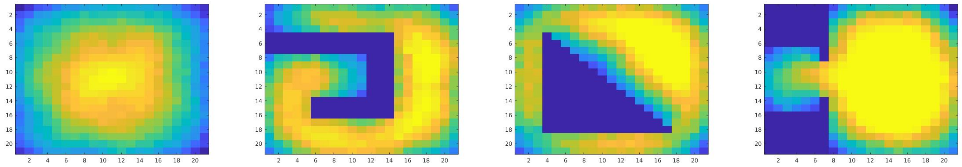
To observe the effects of the various maxims, each agent was run for 10^7 episodes, each being 42 time steps long., $\varepsilon = 0.75$, discount factor $\gamma = 0.9$, learning rate $\alpha = 0.1$.

5.1 Single agent learning

5.1.1 Entropy

For the maximisation of entropy, we considered both computing entropy directly and supplying this as a reward bonus to the agent.

Here we see that purely maximising entropy leads to increased valuations of regions far from walls and relative obstacles. In Fig. 2.(**b**) we see a clear increase in valuation as the agent moves away from either of the "dead-end" regions, with the greatest value being seen in the space on the right hand side, which allows for greatest n-step access to the remainder of the environment.



(a) (b) (c) (d)

Figure 2: These colour maps represent the value of the various arenas the agent was placed in when aiming to purely maximise entropy. **(a)** is an empty arena. **(b)** is a snaking obstacle. **(c)** has a triangular obstacle with two corridors. **(d)** is an arena consisting of two rooms, where the agent initialises in the smaller room.

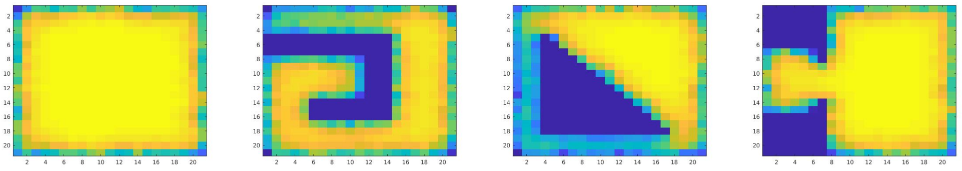
Similarly in Fig. 2.(**d**), the environment containing two different sized rooms, we observe that the greater values in the respective rooms are toward the centre, giving greater access to the remainder of the environment; however, as can be noted, the restricted region of the path between the two rooms also sees a greater valuation than other restricted regions, as this is the area that must be traversed to receive increased entropy.

This is consistent with what is observed in empowerment [11], particularly the cases of the mazes where an agent is in the state of greatest empowerment when it is not enclosed in walls, and over a defined n-step time horizon can actualise the greatest number of future states from the current state.

5.1.2 Discounted empowerment

When considering the case of quantities similar to entropy in our environment we considered the method we respectfully call γ-empowerment, accomplished by providing a negative reward for remaining in the same state between time steps.

As remaining in the same state between time steps is only possible in the case of colliding with an obstacle, and the high level of ε makes this much more likely near corners or walls, we felt this was an appropriate quantity to consider in consort with entropy and empowerment, since under this scheme an agent should more highly value regions that provide future freedom of movement, and as opposed to the entropy case above, requires no calculation, and is easy to work with on-policy.



(a) (b) (c) (d)

Figure 3: These colour maps represent the value of the various arenas the agent was placed in when aiming to purely maximise γ-empowerment (4), with the obstacles, episodes, episode length and parameters as in Fig. 2

The resulting graphs can be seen in Fig. 3. Here we see similar regions of high valuation, with significantly increased value in the surrounding regions. This is consistent with what we would expect in an empowerment case, though with a substantially increased value for n in the traditional case.

We would not expect such a high value directly up to the wall regions, where in the maze variants seen in Klyubin et al. [11] paper there is a smoother gradient between values, with much more distinct regions of increased entropy, closer to what we see in the entropy case.

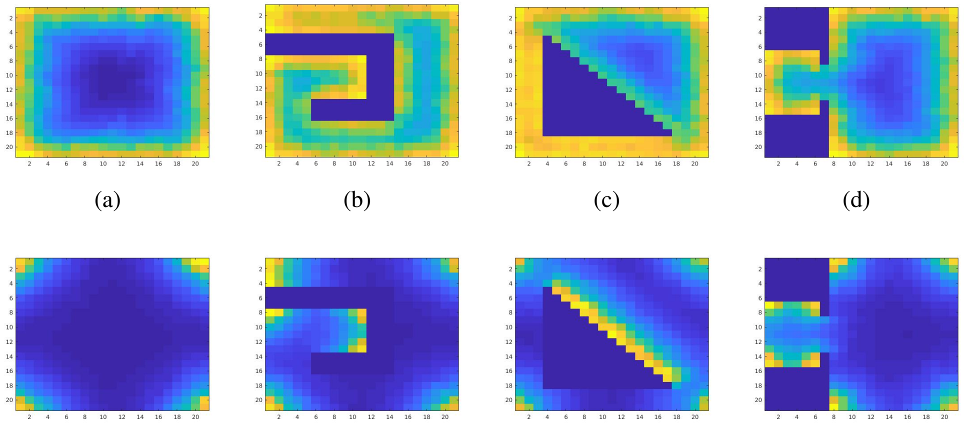
5.1.3 Flexibility of approach

In this section we briefly touch on two other approaches we have used in experimentation to show that this approach is flexible in that it is not only applicable to entropy or entropy-like quantities over a state space. These are variants we have chosen to call *prediction error reduction* and *corner favouring*.

Both sets of images seen in Fig.4 use very simple reward schemes, only receiving rewards when they are located adjacent to walls and corners respectively. These types of regions are highly useful in noisy environments for localisation and reduction in prediction error, but in essence can serve an additional purpose in control schemes, that perhaps where there is no task relevant information, in a highly dynamic environment an agent should look to remove itself to these less task critical areas.

5.2 Multi-agent RRL

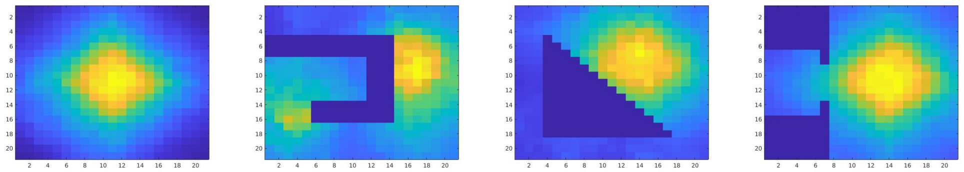
A natural extension of the basic RRL concept we have implemented so far is the extension to multi-agent systems, be they robot and human or multi robot tasks and environments. We are particularly interested in the concept of *socially empowered* robots, given that we can expect that many tasks in logistic and exploratory domains will require robots to work in tandem, or indeed search in tandem for some task related information.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4: These colour maps represent the value of the various arenas the agent was placed in when being rewarded for obtaining sensory values of obstacles or walls in the environment, with the obstacles, episodes, episode length and parameters as in Fig. 2. Subfigures **(a)-(d)** represent the prediction error reduction variants, and the subfigures **(e)-(h)** represent the corner favouring variants.



(a) (b) (c) (d)

Figure 5: These colour maps represent the value of the various arenas the agent were placed in when aiming to purely maximise entropy, where there are 10^7 episodes, each being 42 time steps long, $\varepsilon = 0.50, \gamma = 0.9, \alpha = 0.1$. **(a)** is an empty arena. **(b)** is a snaking obstacle. **(c)** has a triangular obstacle with two corridors. **(d)** is an arena consisting of two rooms, where the agent initialises in the smaller room.. Here both agents are contributing to the same Q function and Value function at each update.

What we see here is a high valuation of mutual space where joint activity is more likely, consistent with what we see in the empowerment-like variants, this has clear use cases in multi-agent tasks, where an agent perhaps moves to these high mutual value regions to recruit other agents where it has identified a task that it is otherwise unable to complete alone.

5.3 Remarks

Rewarding an agent based on entropy leads to valuation of states which is commensurate with what one would expect in an empowerment approach, with the benefit that it can be computed on policy, without the necessity to exhaustively compute over all policies and time series.

Similarly, considering the easier to compute γ-empowerment, we are able to obtain a valuation similar to what we would expect from an empowerment approach; however, we observe that the valuation remains very similar over the open regions with no clear peak value in the environment, which may not prove as useful to the goal of intrinsic motivation over potentially dynamic environments as it shows a tendency to prefer vast regions in the state space, which will make isolating the most interesting or free regions over a complex or dynamic space much more unlikely.

We have also shown that the approach is flexible, highlighting regions such as walls and corners, regions that may be preferable in various existing control architectures. Additionally, we have shown that the approach is consistent in mulit-agent systems, and when rewarding agents for being in a proximity with other agents, what emerges is akin to the empowerment-like case, and should be viewed as a socially empowered state.

All of these variants here serve as a compliment to traditional reinforcement learning approaches through the use of the reflexive component, and indeed, can also be considered in tandem with one another. An intrinsically motivated, agent should seek out regions which are interesting or surprising as in Sect. 5.1, where no task relevant information is available in these identified regions of interest, the agent should return to regions where prediction error can be minimised, and relocalisation is possible, and the cycle should repeat, in a control system, perhaps after sufficient searching of the state space, the agent should move to regions which have minimal impact on a potentially dynamic environment, such as a corner, and wait to search again later.

Alternatively, we consider that if there is no task dependent information available, a continually learning agent should seek out these surprising regions, with the aim of learning more about the environment and correcting the model, by more accurately learning state transition probabilities, or learning about features present in these high interest subspaces, to better perform tasks in the future when this information becomes available.

6 Discussion

6.1 Exploration vs. exploitation

The exploration-exploitation dilemma is not a solved problem in any of the various domains that it has been researched in. We presented here a use case for utilising entropy as a reward on its own or in conjunction with other rewards to highlight the best regions available to an agent in an environment where there is no clear goal. In doing so we have found that these regions which are considered highly valued are similar to those found in empowerment, where the agent more highly values regions from which it is able to access a larger subset of the state space over any given discrete time frame.

As opposed to having to create sophisticated models for task location, the use of entropy maximisation may enable agents to *find* a task or location when lost in a changing environment. In future work we intend to consider the problem using actor-critic algorithms where the actor and critic with two different state-action value functions, and to employ this in a simulated dynamic environment, as well as employing this as a hierarchical model alongside other functions or goals to observe the possibility of robotic self motivation.

6.2 Intrinsic motivation

The search for an intrinsic motivation for an agent to perform any given task, develop new behaviours, or learn its own embodiment is and take advantage of that is a key task in the development of continually learning, adaptable agents which are capable of working in highly dynamic environments. It is essential that an agent is able to identify important or interesting regions in the sensorimotor space, both to learn the model, or in fact learn a goal where no clear goal is immediately visible.

Empowerment seeks to do this [15] by defining an empirical measure which can be performed over the state-action space to definitively state the best possible states for an agent to be in to have sufficient future degrees of freedom. This valuable concept is unfortunately subject to the curse of dimensionality, and as such, other approaches to estimating empowerment have been sought [19]. We believe that we have shown that entropy maximisation allows for an agent to approximate such a position utilising an on-policy approach over varying environments, by instead considering more interesting or surprising regions of the state space to be the most valuable.

This can more concisely be thought of as a form of information empowerment, where, as opposed to the mantra “All else being equal, be empowered” [10], we consider that perhaps the notion in a adaptive learning agent should be “all else being equal, be interesting”.

7 Conclusion

Reflexive reinforcement learning (RRL) is a new direction in machine learning. It is based on the observation that in learning problem where no direct gradient can be used in order to adapt to a particular, the representation of information from the environment (such as state information or evaluative information) requires not only guiding principles (such as smoothness, consistency and locality), but also provides information that can be used to decide about the actions of an agent.

The advantage of reflexive reinforcement learning is that an agent can learn even in the absence of an evaluative signal (reward and punishment), it can bootstrap elementary actions (as in homeokinesis [6]) or can learn about options in the environment (as in empowerment [11]), and obtain more meaningful and generalisable representations (see [16]).

The unavoidable difficulty in reflexive reinforcement learning consists in the fact that the use of quantities that are eventually based on the reward as a reward, introduces a feedback loop which can lead to instabilities or divergences. This is not unknown in RL, where e.g. an often visited source of low reward can dominate a better source of reward that is rarely found, or in cases where correlations among basis functions lead to divergences as notice already in Ref. [1].

In reflexive reinforcement learning such feedback is even more typical, but can also be used to introduce structure the state space by self-organised pattern formation or to identify hierarchical relationships as will be studied in future. In order to keep the effects of self-referentiality under control and to make use of their potential a dynamical systems theory of reinforcement learning is required that does not only consider the agent as a dynamical system, but the full interactive system formed by the agent, its environment and its internal representations.

Finally, we should stress the importance of the presented concept of RRL for explainable AI and its relation to information processing in the brain.

Broader Impact

We believe that RRL has a variety of potential applications in terms of control architecture for an autonomous agents, as well as less lofty pursuits. As we see in Fig. 2 and Fig. 3, the agent shows highlighted regions of preference around the obstacles, preferring to avoid walls and obstructions. We believe there is potential here to consider the notion of “curiosity path planning", where an agent plans the route on the basis of interesting regions within a known environment to better learn about, or be available for future tasks.

Acknowledgements

t.b.a.

References

[1] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

[2] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Incorporating behavioral constraints in online ai systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3–11, 2019.

[3] William Bialek and Naftali Tishby. Predictive information, 1999. arXiv/cond-mat/9902341.

[4] Sebastian Blaes, Marin Vlastelica Pogančić, Jiajie Zhu, and Georg Martius. Control what you can: Intrinsically motivated task-planning agent. In *Advances in Neural Information Processing Systems*, pages 12520–12531, 2019.

[5] Nuttapong Chentanez, Andrew G Barto, and Satinder P Singh. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1281–1288, 2005.

[6] Ralf Der and Georg Martius. *The playful machine: Theoretical foundation and practical realization of self-organizing robots*, volume 15. Springer Science & Business Media, 2012.

[7] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1-3):70–87, 2006.

[8] M Herrmann and R Der. Efficient Q-learning by division of labour. In *Proceedings ICANN*, volume 95, pages 129–134, 1995.

[9] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[10] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. All else being equal be empowered. In *European Conference on Artificial Life*, pages 744–753. Springer, 2005.

[11] Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135. IEEE, 2005.

[12] Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *IMCL*, pages 663–670, 2000.

[13] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

[14] David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

[15] Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment – An introduction. In *Guided Self-Organization: Inception*, pages 67–114. Springer, 2014.

[16] Simón C Smith and J Michael Herrmann. Evaluation of internal models in autonomous learning. *IEEE Transactions on Cognitive and Developmental Systems*, 11(4):463–472, 2019.

[17] Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

[18] Alexander Tschantz, Beren Millidge, Anil K Seth, and Christopher L Buckley. Reinforcement learning through active inference. *arXiv preprint arXiv:2002.12636*, 2020.

[19] Ruihan Zhao, Stas Tiomkin, and Pieter Abbeel. Learning efficient representation for intrinsic motivation. *arXiv preprint arXiv:1912.02624*, 2019.

[20] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proc. 23$^{\text{rd}}$ AAAI Conference on Artificial Intelligence*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Supplementary material

- Generalisation to continuous state and action spaces
- Additional multi agent figures and variants