Silviu Pitis, PIN 542667

Using Modular Abstractions in Reinforcement Learning for Objective Specification and Reasoning

Outline of Proposed Research The field of reinforcement learning (RL) seeks to design agents that learn to interact with a partially observed environment [34]. Although RL agents have enjoyed significant success—learning to solve challenging continuous control tasks from scratch [19, 24] and achieving super human performance on Atari [22] and Go [30, 31]—the current state-of-the-art agents are highly task specific and do not yet generalize well to unseen environments, goals, or states [10]. If we are to build general purpose RL agents, I argue that at least two fundamental and closely related questions must be addressed: (1) what objectives should a "general-purpose" agent optimize and how can they be specified, and (2) how can we endow agents with the ability to reason about the world at multiple levels of abstraction, using multiple modalities (especially natural language)? My thesis is that the use of modular, high-level abstractions over objectives and states is critical to both questions. To explore this, I propose to build upon ideas from RL [2, 18, 28, 35], planning [12, 27], social choice [3, 6], and causality [23, 33]. Below I outline related work, followed by four specific angles I would like to pursue in the course of my research.

Background & Related Work The typical RL setup involves an agent acting in a Markov Decision Process (MDP) to maximize the long-term sum of a scalar reward (a "value function") [34]. Traditionally, this reward signal was specified by a human designer and given to the agent explicitly. While this works in the case of games (e.g., Atari & Go [22, 30]), the limitations of this approach become clear once we consider more general purpose agents; after all, what reward signal do humans (or societies) optimize?

A pertinent line of work considers agents that can predict, and act to optimize, *multiple* "general" value functions (GVFs) [28, 35]. GVFs have been used to improve agent performance [14] and transfer knowledge across tasks [5], as well as for multi-goal RL [2, 24], hierarchical RL [15, 36], and planning [11, 25].

A key view of GVFs is that they represent abstract knowledge about the world [35], a perspective closely aligned with research on state abstraction [1, 18]. One useful type of GVF, which measures temporal distance between states and goals, is based on state abstractions: in continuous spaces individual states (usually) have measure zero, so that the goal *must* be a state abstraction, or set of success states. GVFs and state abstractions each provide a discrete summary of a high-dimensional underlying space—reward signals and states, respectively—and might be used as a foundation for augmenting RL agents with symbolic reasoning (colloquially known as Good Old-Fashioned AI, or "GOFAI") [4, 27]. For instance, abstractions might enable constraint satisfaction algorithms and partial supervision [46], and the design of interpretable agent interfaces [13]. Though this research direction remains largely unexplored, I believe it is particularly promising as it may enable humans to communicate objectives to RL agents in abstract terms [9, 21].

Abstractions are also fundamental to the intersections of RL with (1) social choice and (2) causation. For RL agents to optimize the right objective from a societal perspective, they must correctly model human values. A foundational idea in social choice is that human preferences can only be practically communicated in *ordinal* or discrete terms [29]. Since RL agents model the world in *cardinal* or continuous terms, they must be able to do the inverse of abstraction: turn (multiple) abstractions (e.g., votes) into an inference about underlying "social preference" [51]. In the study of causation, we are almost always interested in causal relationships between *variables*—does *smoking* cause *cancer*?—which are defined as abstractions over an underlying probability space [23]. A handful of researchers have applied counterfactual reasoning to RL [2, 8, 20], but there is still much to explore and few have applied RL to causal discovery [39] .

Hypotheses, Research Objectives, and Methodology (*in no particular order*)

1. Integrating multiple goal spaces and state abstractions. Humans can accomplish arbitrarily defined goals, whether expressed by an image or via language, and can translate image goals into language goals and vice versa. This ability is fundamental to expressing general preferences and solving novel tasks. In our 2019 extended abstract [50] we propose to enable GVF-based, multi-goal RL agents [2, 24, 28] to do the same by learning a Prototype Goal Encoding ("ProtoGE") that maps goals (or state abstractions) from one space to a more specific "prototype" goal in another space. I hypothesize that if developed, this idea will be instrumental in designing agents capable of achieving a variety of goals, and reasoning using a variety of abstractions, irrespective of modality (vision, language, etc.). To this end, I would like to further formalize our framework in order to develop theoretical performance bounds on the quality of goal translations and to account for goals with disconnected success states (e.g., "travel to one of four corners"). I would also like to experiment with a generative ProtoGE map, rather than the current rule-based mechanism. A successful outcome would be an agent that quickly learns to achieve novel goal specifications by translating them into its native space (cf. humans translating a second language into their native tongue). This might lead to an RL agent that can autonomously master an environment, and quickly adapt to commands in *any* language (English, Chinese, etc.). It may also help us to design an agent that can integrate *multi-modal* social cues from multiple principals (e.g., explicit manual feedback, verbal feedback, or a simple nod).

2. Top-down search with GVFs. Humans have the ability to visualize a set of landmarks on the way to a destination; we can generate landmarks that are nearby, close to the goal, or in an ad hoc fashion, and we can hold and compare multiple candidate plans in our head (e.g., alternative driving routes). By contrast, most modern RL agents are either "model-free" [19, 22] (even hierarchical agents [17, 37]) or use a model to rollout trajectories forward in time [7] (*forward* search). In a 2017 workshop paper [46] I proposed to endow RL agents with the ability to do *top-down* search, which is closely related to an older idea—DG learning—proposed by Leslie Kaelbling in 1993 [16]. This approach has been slow to develop as many of its parts are still subject to active research. But recent work toward better distance models (represented by GVFs) [28, 49], state abstractions [1] and discrete representations [13] has set the stage for combining RL with planning [11]. I would like to build on these ideas to develop an agent that can plan top-down, similar to the way humans do, by dynamically generating landmarks and subgoals, and comparing and combining multiple proposed plans. Though this will require significant engineering, it also entails several interesting theoretical questions, which I intend to explore: e.g., (1) is learning and planning with short-horizon goals provably more efficient and/or accurate than learning long-term goals? (hypothesis: yes), and (2) can we develop more efficient RL algorithms via good landmark selection? (hypothesis: yes).

3. Unifying causal reasoning and RL. Where do useful abstractions come from? An appealing idea is to think of the world as a collection of *independent* causal mechanisms [23]—for instance, the sun rises regardless of when and where I drink my morning coffee. Then each mechanism can be reasoned about individually and entails an abstraction over the underlying system: the sun's movement can be predicted in isolation, which suggests it exists as a discrete entity, independent of my coffee habits. I would like to investigate two specific hypotheses at the intersection of causality and RL. First, noting that goal relabeling [2], an effective technique for multi-goal RL, uses counterfactual reasoning to exploit the independence relation between the agent's subjective goal and the environment transitions; I hypothesize that other independences could be similarly used to relabel data and improve sample efficiency and generalization capabilities. Second, as humans naturally intuit causal relations (albeit not always correctly), I hypothesize that so too can RL agents, if equipped with a carefully designed algorithm for doing so.

4. Integrating ordinal signals for cardinal choice. As suggested in the Background, RL agents will need a theoretically justified way of translating *ordinal* feedback from humans into a *cardinal* understanding of the world. Standard machine learning techniques (e.g., modeling a binary signal with a Bernoulli distribution) are not immediately applicable here, as feedback signals may come from multiple correlated, and sometimes adversarial, principals, and are likely not independent and identically distributed. Is there a normatively justified way of integrating ordinal feedback signals? [26]. My 2019 project [51] is a first step toward formalizing this problem in a simplistic, one-shot setting. I would like to further pursue this topic in more complex settings, both theoretically (optimality, complexity, loss bounds, value of diverse feedback) and empirically (comparison to standard voting rules, perceived trustworthiness to humans).

1 Main References

[1] D. Abel, D. E. Hershkowitz, and M. L. Littman. Near optimal behavior via approximate state abstraction. arXiv preprint arXiv:1701.04113, 2017.

[2] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. P. Abbeel, and W. Zaremba. Hindsight experience replay. In Advances in Neural Information Processing Systems, pages 5048–5058, 2017.

[3] K. J. Arrow, A. Sen, and K. Suzumura. Handbook of social choice and welfare, volume 2. Elsevier, 2010.

[4] C. Atkeson. The symbolic robotics manifesto, aka the GOFAI manifesto. http://www.cs.cmu.edu/ cga/gofai/. Accessed: 2019-09-08.

[5] A. Barreto, W. Dabney, R. Munos, J. J. Hunt, T. Schaul, H. P. van Hasselt, and D. Silver. Successor features for transfer in reinforcement learning. In Advances in neural information processing systems, pages 4055–4065, 2017.

[6] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia. Handbook of computational social choice. Cambridge University Press, 2016.

[7] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. A survey of monte carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in games, 4(1):1–43, 2012.

[8] L. Buesing, T. Weber, Y. Zwols, S. Racaniere, A. Guez, J.-B. Lespiau, and N. Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. arXiv preprint arXiv:1811.06272, 2018.

[9] H. Chan, Y. Wu, J. Kiros, S. Fidler, and J. Ba. ACTRCE: Augmenting experience via teacher's advice for multi-goal reinforcement learning. arXiv preprint arXiv:1902.04546, 2019.

[10] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman. Quantifying generalization in reinforcement learning. arXiv preprint arXiv:1812.02341, 2018.

[11] B. Eysenbach, R. Salakhutdinov, and S. Levine. Search on the replay buffer: Bridging planning and reinforcement learning. arXiv preprint arXiv:1906.05253, 2019.

[12] H. Geffner and B. Bonet. A concise introduction to models and methods for automated planning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 8(1):1–141, 2013.

[13] R. T. Icarte, T. Klassen, R. Valenzano, and S. McIlraith. Using reward machines for high-level task specification and decomposition in reinforcement learning. In International Conference on Machine Learning, pages 2112–2121, 2018.

[14] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. arXiv preprint arXiv:1611.05397, 2016.

[15] Y. Jiang, S. Gu, K. Murphy, and C. Finn. Language as an abstraction for hierarchical deep reinforcement learning. arXiv preprint arXiv:1906.07343, 2019.

[16] L. P. Kaelbling. Hierarchical learning in stochastic domains: Preliminary results. In Proceedings of the tenth international conference on machine learning, volume 951, pages 167–173, 1993.

[17] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In Advances in neural information processing systems, pages 3675–3683, 2016.

[18] L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for MDPs. In ISAIM, 2006.

[19] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous

control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.

[20] C. Lu, B. Schölkopf, and J. M. Hernández-Lobato. Deconfounding reinforcement learning in observational settings. arXiv preprint arXiv:1812.10576, 2018.

[21] J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel. A survey of reinforcement learning informed by natural language. arXiv preprint arXiv:1906.03926, 2019.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529, 2015.

[23] J. Peters, D. Janzing, and B. Schölkopf. Elements of causal inference: foundations and learning algorithms. MIT press, 2017.

[24] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, et al. Multi-goal reinforcement learning: Challenging robotics environments and request for research. arXiv preprint arXiv:1802.09464, 2018.

[25] V. Pong, S. Gu, M. Dalal, and S. Levine. Temporal difference models: Model-free deep RL for model-based control. arXiv preprint arXiv:1802.09081, 2018.

[26] A. D. Procaccia. How is voting theory really useful in multiagent systems? available online, URL: http://www. cs. cmu. edu/arielpro/papers/vote4mas. pdf (DOA: 15.01. 2013).

[27] S. J. Russell and P. Norvig. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,, 2016.

[28] T. Schaul, D. Horgan, K. Gregor, and D. Silver. Universal value function approximators. In International Conference on Machine Learning, pages 1312–1320, 2015.

[29] A. Sen. Collective choice and social welfare. Harvard University Press, 2018.

[30] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. nature, 529(7587):484, 2016.

[31] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. Nature, 550(7676):354, 2017.

[32] D. Silver, H. van Hasselt, M. Hessel, T. Schaul, A. Guez, T. Harley, G. Dulac-Arnold, D. Reichert, N. Rabinowitz, A. Barreto, et al. The predictron: End-to-end learning and planning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3191–3199. JMLR. org, 2017.

[33] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. Causation, prediction, and search. MIT press, 2000.

[34] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

[35] R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In The 10th International Conference on Autonomous Agents and Multiagent Systems, 2011.

[36] R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial intelligence, 112(1-2):181–211, 1999.

[37] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3540–3549. JMLR. org, 2017.

[38] M. White. Unifying task specification in reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3742–3750. JMLR. org, 2017.

[39] S. Zhu and Z. Chen. Causal discovery with reinforcement learning. arXiv preprint arXiv:1906.04477, 2019.

2 References to Own Work

[40] K. De Asis, A. Chan, S. Pitis, R. Sutton, and D. Graves. Fixed-horizon temporal difference methods for stable reinforcement learning. arXiv preprint arXiv:1909.03906, 2019.

[41] S. Pitis. Beyond binary: Ternary and one-hot neurons. r2rt.com, 2016.

[42] S. Pitis. Binary stochastic neurons in Tensorflow. r2rt.com, 2016.

[43] S. Pitis. Deconstruction with discrete embeddings. r2rt.com, 2016.

[44] S. Pitis. Written memories: Understanding, deriving and extending the LSTM. r2rt.com, 2016.

[45] S. Pitis. Methods for retrieving alternative contract language using a prototype. In Proceedings of the 16th International Conference on Artificial Intelligence and Law, 2017.

[46] S. Pitis. Reasoning for reinforcement learning. In Hierarchical Reinforcement Learning Workshop at the 31st Conference on Neural Information Processing Systems (HRL@NIPS 2017), 2017.

[47] S. Pitis. Source Traces for temporal difference learning. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence. AAAI Press, 2018.

[48] S. Pitis. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In Proceedings of the 33nd AAAI Conference on Artificial Intelligence. AAAI Press, 2019.

[49] S. Pitis, H. Chan, and J. Ba. Modeling norms and metrics with neural networks: Deep norms, wide norms, and neural metrics. Work in progress, 2019.

[50] S. Pitis, H. Chan, and J. Ba. ProtoGE: Prototype goal encodings for multi-goal reinforcement learning. In The 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM 2019), 2019.

[51] S. Pitis and M. Zhang. Objective social choice with non-i.i.d. votes. Work in progress, 2019.

3 Other References Cited In Application

[52] L. Kaplow. Multistage adjudication. Harvard Law Review, 126:1179, 2012.

[53] L. Kaplow. Likelihood ratio tests and legal decision rules. American Law and Economics Review, 16(1):1–39, 2014.

[54] J. M. Ramseyer. Liability for defective products: Comparative hypotheses and evidence from japan. The American Journal of Comparative Law, 61(3):617–655, 2013.

[55] J. M. Ramseyer. Second-best justice: The virtues of Japanese private law. University of Chicago Press, 2015.