**Intuitions Regarding Transferable Predictive Knowledge**

**Introduction**

Reinforcement learning is one of the most promising approaches to solve artificial intelligence problem. However, reinforcement learners learn feature as function of given state-action-reward, and any change to those makes transfer hard, even if the nature of new task is similar.

Even with hundreds of trillions of neural connections, we humans often struggle with reasoning over the overwhelming reality. But while a lone human is less likely to survive in an isolated island than a lone chimpanzee, humans can do wonders in large numbers (Harari, 2014). Human to human transfer learning (e.g. education) and large-scale collaboration has made us the most sophisticated surviving species in this constrained environment.

Knowledge can be represented using predictions about the future. Predictions can also be used to represent states (Littman et al., 2002). Predictive knowledge is autonomously verifiable, can be represented by general value functions, which can be learned by reinforcement learning approaches. However, predictive knowledge has no public semantics, and is personal and subjective (Sutton et al., 2005; White, 2015).

The goal here is to discuss assumptions regarding learning transferable predictive knowledge. Transfer in reinforcement learning has a rich literature, but the discussion here focuses on how humans transfer their knowledge.

Core reinforcement learning literature mainly focuses on optimal decision making. Here, the focus is on discussing design of internal environment of the agent, that computes state, generates rewards based on observation, executes actions given decisions etc. (Singh et al., 2010).

**Objectives**

The objectives here are to discuss –

1. Transferable PK learning.
2. Using PK as inductive bias in solving RL task and updating PK based on learning of different agents.
3. Discussing approaches that imitative and non-imitative approaches to transferable PK.

**Human transferability**

3 important reasons for which humans can transfer their learnings well:

1. Common basis reward function (maximizing chances of survival and produce maximum descendants)

All behaviors are for reducing primary deficit or drivers (survival and reproduction needs – hunger, sex, pain avoidance) or learned secondary drivers – which is reinforced by predicting reduction of primary drivers. Animals get motivated to take action to restore balance of internal environment (Skinner & Hull, 1944). Intrinsic motivation attributes are more or less universal not only for humans, but also for other animals, which also indicates the goal of intrinsic motivation is ensuring survival of species.

Basis fitness function gives rise to primary, intrinsic and learned rewards. Like direct drivers, intrinsic rewards like curiosity are also inherent, but they don’t directly reduce primary drivers. However, on the long run, they help us gather experiences and skills that increases our chances of primary driver reduction across different scenarios. They can also act as reinforcement to secondary drivers. In fact, all activities that reduce primary drivers can be thought as intrinsically motivated, as all such activities indirectly increases chance of survival (Singh et al., 2010).

All human has similar goal of survival, and they act on similar environment with similar constraints and challenges. Thus, humans have learned to work together and value their states similarly (states with food are good, states with harm are bad), and learn similar policy towards survival. This also applies today, as we highly value states that predicts long term survival.

1. Common action space

Options framework deals with learning higher level hierarchical decisions that execute many low-level decisions (Sutton et al., 2005). In case of humans, basic apparent low-level actions (use of hands and feet) can be thought as common. We abstract over or build complex new hierarchical actions over the basic ones – for example driving includes controlling the steering with hand and idea of movement. Thus, we have shared understanding of ways of impacting our environment. Common basic action ideas make the learned knowledge relatable for human.

1. Common observation and thought reference

Like humans, many animals have similar basis reward (survival) and basis actions (e.g. movement) sense, but a significant human innovation is common reference to express and think about reality and fiction – language. We think of similar ideas with common words of language, and we share knowledge with others using language. The location of certain ideas or words can be traced to regions of cerebral cortex (Huth et al., 2016). Similar environment challenge, common basis reward and action and makes us learn common approaches to survival, and we can transfer our learnings with common language.

**Learning Transferable PK**

Since most of the tasks we want artificial agents to solve is related to human need, we can consider designing agents that tries to approach human intelligence. If human significance is possible due to transfer learning and collaboration, we can think of designing agents that can generate 'human imitating and autonomously verifiable transferable knowledge'. To enable learning of transferable predictive knowledge, human intelligence imitating artificial agents can be designed to have:

*Common agent space*: including common basic capabilities to impact external environment, with possibilities of hierarchical actions learning.

*Common human imitating internal balance and reward function*: a single question can be used to value every state – probability of survival from that state. Internal environment of human rewards interim reduction of primary drivers like food consumption in case of imbalance of the internal environment. A human imitating reward function and internal environment balance can lead to common valuation of states. All secondary rewards will be treated as derivatives of the basis reward.

*Common abstraction reference*: human imitating supervised learner can detect objects or ideas of the environment (e.g. semantic segmentation) and represent observation ideas as graph. Semantic segmentation is a popular research topic, and it is used in many RL tasks (e.g. learning to drive in a driving simulation task). In case of human intelligence imitating artificial agent, supervised semantic segmentation of observation can act as common reference for different learning agents.

*Environment imitating challenges faced by humans*: competition over constrained resource while ensuring survival in human imitating environment can allow artificial agents to find human imitating approaches and skills.

*External knowledge base to update the world model learned by distributed agents*: literature has addressed using external memory with artificial agents (Graves et al., 2016). If distributed agents have access to transferable knowledge base, it can be used as inductive bias for problem solving. If the agent’s learnings are similarly transferable, it can be used to update the knowledge base.

*Nature of PK*: RL literature deals with diverse kinds of knowledge to learn and transfer. Reward function is calculated in the inner environment of the agent. It can be learned from a distribution of reward functions, which helps development of skills that are helpful across multiple tasks (Barto et al., 2004; Singh et al., 2005). Reward function can also be reused to transfer knowledge for future tasks (Guo et al., 2013).  
General value function can be used as predictive knowledge (White, 2015), to represent states (Littman et al., 2002). Value function can be transferred as parameters for new task (Taylor & Stone, 2005). Values can be common for many SMDP, and it can be used to initialize value for new SMDP (Mehta et al., 2008).  
Relational state representation is a compact way to represent state (van Otterlo, 2012). It can be used to learn predictive state representation (Wingate et al., 2007). Relation can be learned with graph (Battaglia et al., 2018). Graphs can also be used in reinforcement learning (Jiang et al., 2018; Madjiheurem & Toni, 2020).  
Graph can be a compact way to represent diverse relations of ideas of the state with capability of continual learning and good search and merging capability.

*Bottom up training*: TD networks are used to build higher level predictions from lower level predictions, hierarchical options are composed of many routine low level decisions (Sutton et al., 2005). If we consider high-level options learning from backwards, agent already has some understanding of the component low-level action outcomes.

Humans don’t try to solve complex tasks like driving just after birth. We start with basic skills like movement and navigation and slowly learn more complex and hierarchical skills like driving. Even if a human plays a racing game for the first time, given s/he knows settings of the game, s/he already has knowledge of movement, constrained time. S/he already knows that the car needs to be driven on the road while avoiding collision etc. Often, a hierarchical problem might not seem to have any connection with basis goal of survival, but with bottom up training, it should be possible to draw the relation.

When an RL agent tries to solve a complex RL task from scratch, it tries to do so without any understanding of the parts of the problem. It does often allow the agent to find novel solutions to the task, but it also creates problems like sparse reward and requirement of many samples.

But if the agent is trained in a bottom up process – solving basic problems before complex ones, not only it should be able to learn good cognitive map and transfer low-level skills to solve the given racing game task more effectively, but it should also be able to solve related tasks like autonomous vehicle simulation with same low-level skills.

*Top down state value approximation*: Humans often value a state in terms of relation of ideas of that state, possible actions and the current goal (which depends on the balance of internal environment - we value food more when we're hungry). This can be considered as top down breakdown of the observation and state. And after breaking down the ideas of the observation, we usually get a sense of the value of the state from the relationship of those ideas. The value of states is often different for different people given internal environment balance (a thirsty person will value water more) or awareness of complex reward structure (one might not value education if s/he isn't sure about its long-term benefits). But generally, the of states that more directly reduces primary drivers is positively valued.

Similarly, given internal environment balance driven goal (e.g. hungry) and allowed actions, top down supervised breakdown of observation into finite ideas, and representing those ideas as graph should reduce variables and allow artificial agent to formulate value of state from past histories of similar state ideas with less calculation, and at the same time make the new history relevant for other agents that takes observation input with similar graph representation. We can compare this with ‘Goal directed learning’, where old learnings are used to solve new task. Using old experience to solve new task is also discussed in RL literature (Niekerk et al., 2019; Saxe et al., 2017).

*Model learning*: model-based search and planning methods like Monte Carlo Tree Search can be applied to find related experience in the known world model to update value of given state. Here, the Knowledge is being used as inductive bias for solving a new problem. The learning from new trajectories can further be used to update the knowledge base.

Transfer learning allows to keep updating value of complex world, with a goal to maximize chances of survival. If the internal environment of the agent gives rise to motivation to reduce primary driver, the agent can autonomously update value of world and take actions according to greedy policy to reach states that restores internal environment balance. If many policies lead to survival, optimal policy can be used for valuing states.

**Type of approaches**

Depending on the internal environment design, we can think of human imitating or natural reinforcement learner. As described earlier, human imitating transferable knowledge can be possible if knowledge is gathered with reference to human imitating external and internal environment – reward function, basis actions and observation reference, learning process etc.

It should be possible to instill basis actions in the agent, with the capacity to abstract higher-level actions. Learning observation reference can be achieved with supervised learning methods like semantic segmentation that can detect both concrete and abstract ideas, and also deconvolute each idea into their component ideas – like the idea of book can break down into paper, knowledge etc. Current semantic segmentation methods focus on 2D vision, but full spectrum 5 senses classifiers including 3D vision should help human imitating learning.

If the agent had perfect causal relationship of the world, then just providing the goal of survival would suffice (Samuelson & Swinkels, 2006). But since the agent doesn't have such understanding to start with, we need to design reward functions that reward interim states like eating food, mating or exploration. Inverse reinforcement learning can be used to approximately estimate reward function of a subject from provided trajectories (Abbeel & Ng, 2004). It can be used to understand human imitating reward function for diverse scenarios.

Human imitating environments is necessary where the artificial agents can act and learn. Designing such environment can be challenging. Artificial systems can learn from interactions in the real world (with similar impact and feedback on the environment as humans). In future, it might also be possible to directly access the histories of human brain to train artificial agents.

One of the motivations of the study of machine learning was to allow artificial systems to learn the rules for solving a problem instead of hand designing them ourselves. But human imitating intelligence will require hand designed reward function, environment, observation reference etc. Instead, we can think of new kind of problem solution skills that won’t require significant human input.

We can initiate agents that can learn transferable knowledge in a constrained environment that is not human imitating. Like an environment with increasing entropy, while the goal of artificial agents will be to minimize entropy with competition for constrained resources. Thus, the agents will have more reward if they can reduce more entropy with less resources. Human intelligence is nothing but the best approach to solve problem of human survival, and problem of artificial survival can give rise to different kind of intelligence.

For allowing transferable knowledge, the agents can have a common unsupervised representation learner. Unsupervised representation learning for a number of related tasks (Espeholt et al., 2018), and unsupervised object segmented observation input (Burgess et al., 2019; Greff et al., 2019; Watters et al., 2019) is discussed in RL literature. Unsupervised learning can allow NN to capture more accurate representation with increasing training.

We can combine RL and evolutionary methods – the agents will generate knowledge in their lifetime with RL methods, and variations of the most successful agents will be passed on to next generations. Upon generations of knowledge buildup, we can expect the agents to start figuring out novel approaches to solve given problems, which'll be a kind of intelligence, although not human imitating.

**Conclusion**

With technologies like 'Neuralink', we're aiming to build physical connection between human and artificial systems. One day it might be possible for artificial agents to learn knowledge that is directly transferable with humans. Magnus Carlsen has gathered years of experience to become world champion in chess, perhaps one day the learnings can be suited and transferred to another human or artificial agent instantly. At the same time, it might be possible to suit and transfer objective approaching knowledge of artificial agents directly to humans – like transferring Alpha Zero's chess expertise directly to Carlsen. Agent-Human-Agent transferable knowledge can allow both humans and artificial agents to instantly build on top of each other’s knowledge to solve harder problems. That might lead to more and more objective approximation of causal relations of reality.

At the same time, we can develop totally new kind of intelligence that tries to solve altogether different problems or same problems differently than humans. It's not clear what'll be the use of such intelligence, but the opportunities demand exploration.

**Reference**

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, *346*, 1–8. https://doi.org/10.1145/1015330.1015430

Barto, A. G., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. *Proceedings of the 3rd International Conference on Development and Learning*, 112–119.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks*. 1–40. http://arxiv.org/abs/1806.01261

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., & Lerchner, A. (2019). *MONet: Unsupervised Scene Decomposition and Representation*. 1–22. http://arxiv.org/abs/1901.11390

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Yotam, B., Vlad, F., Tim, H., Dunning, I., Legg, S., & Kavukcuoglu, K. (2018). IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. *35th International Conference on Machine Learning, ICML 2018*.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476. https://doi.org/10.1038/nature20101

Greff, K., Lopez, R., Rishabh, K., Nick, K., Chris, W., Daniel, B., Matthey, L., Botvinick, M., & Lerchner, A. (2019). *Multi-Object Representation Learning with Iterative Variational Inference*.

Guo, X., Singh, S., & Lewis, R. (2013). Reward mapping for transfer in long-lived agents. *Advances in Neural Information Processing Systems*, 1–9.

Harari, Y. N. (2014). *Sapiens: A Brief History of Humankind.* Harper.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Jiang, J., Dun, C., Huang, T., & Lu, Z. (2018). *Graph Convolutional Reinforcement Learning*. 1–13. http://arxiv.org/abs/1810.09202

Littman, M. L., Sutton, R. S., & Singh, S. (2002). Predictive representations of state. *Advances in Neural Information Processing Systems*.

Madjiheurem, S., & Toni, L. (2020). Representation learning on graphs: A reinforcement learning application. *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*.

Mehta, N., Natarajan, S., & Tadepalli, P. (2008). *Transfer in variable-reward hierarchical reinforcement learning*. 289–312. https://doi.org/10.1007/s10994-008-5061-y

Niekerk, B. Van, James, S., Earle, A., & Rosman, B. (2019). *Composing Value Functions in Reinforcement Learning*.

Samuelson, L., & Swinkels, J. (2006). Information, evolution and utility. *Theoretical Economics*, *1*(1), 119–142. https://econtheory.org/ojs/index.php/te/article/viewFile/20060119/452/16

Saxe, A. M., Earle, A. C., & Rosman, B. (2017). Hierarchy through composition with multitask LMDPS. *34th International Conference on Machine Learning, ICML 2017*, *6*, 4616–4629.

Singh, S., Barto, A. G., & Chentanez, N. (2005). Intrinsically motivated reinforcement learning. *Advances in Neural Information Processing Systems*.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 70–82. https://doi.org/10.1109/TAMD.2010.2051031

Skinner, B. F., & Hull, C. L. (1944). Principles of Behavior. *The American Journal of Psychology*, *57*(2), 276. https://doi.org/10.2307/1416955

Sutton, R. S., Rafols, E. J., & Koop, A. (2005). *Temporal Abstraction in TD Networks*.

Taylor, M. E., & Stone, P. (2005). Behavior transfer for value-function-based reinforcement learning. *Proceedings of the International Conference on Autonomous Agents*, *July*, 201–207. https://doi.org/10.1145/1082473.1082482

van Otterlo, M. (2012). Solving relational and first-order logical markov decision processes: A survey. *Adaptation, Learning, and Optimization*, *12*, 253–292. https://doi.org/10.1007/978-3-642-27645-3\_8

Watters, N., Matthey, L., Bosnjak, M., Burgess, C. P., & Lerchner, A. (2019). *COBRA: Data-Efficient Model-Based RL through Unsupervised Object Discovery and Curiosity-Driven Exploration*. 1–24. http://arxiv.org/abs/1905.09275

White, A. (2015). *Developing a predictive approach to knowledge*.

Wingate, D., Soni, V., Wolfe, B., & Singh, S. (2007). Relational knowledge with predictive state representations. *IJCAI International Joint Conference on Artificial Intelligence*, 2035–2040.