Towards a more Unified, Explicable, and Generalized Representation of Human Utility

Xinyi Yang
Department of Automation
Tsinghua Universtiy
xy-yang21@mails.tsinghua.edu.cn

Abstract

Utility determines our preferences about decisions that involve uncertainty in every aspect of life. But as it is both hierarchical and multidimensional, there exists no unified representation. This essay discusses two basic ways to obtain reward function, reward engineering and inverse learning without reward engineering, under the framework of RL. Based on learning ways, three kinds of representations of human utility, prior knowledge, human feedback, and intrinsic motivation are compared in interpretability and generalization. Finally, we argue that intrinsic motivation could provide more general explanations, and seeking a unified representation of different motivation dimensions is worth attention.

1 Introduction

There are very bountiful and varied expressions of intelligence in human behavior, *e.g.*, perception, planning, imitation, motor control, social cognition, knowledge representation, non-verbal language and verbal language. But what could drive agents to behave intelligently in such a diverse variety of ways? Is there a unified mechanism or a specialized problem formulation needed for each ability? A convictive hypothesis is that the generic objective of maximizing utility is enough to drive behavior that exhibits most intelligent abilities.[10] Rational agents choose a choice from an implicit set of options, which will maximize their own utility, *i.e.*, the factor summed from all the rewards and costs for each state on the way to the goal. Furthermore, from a young age, humans implicitly assume that agents choose goals and actions to maximize the rewards they expect to obtain relative to the costs they expect to incur, which allows both children and adults to observe the behavior of others and infer their beliefs and desires, their longer-term knowledge and preferences and even their character.[5] Therefore, how to learn and represent human utility becomes the key to building agents that can convey such intelligence like humans or even interact with humans.

However, it seems very difficult to model human utility in a unified way in machines for two reasons: Firstly, utility functions are always internal to humans and varied from individual to individual. For example, some people prefer to organize clothes by season while other people prefer by color, but they usually don't know why or have never even noticed such habits themselves. Secondly, reward signals to humans are not only hierarchical but also multidimensional. There are basic needs in daily life, *e.g.*, eating food for feeling full, wearing thick clothes for warmth, and there are also social concerns when interacting with other people, *e.g.*, praise and belief. To tackle these problems, researchers have proposed several methods to learn and represent different kinds of rewards, which are aimed at in this essay. We first start with two ways to obtain reward functions, explicit predesigning rewards and inverse learning without reward engineering, which will be introduced with the reinforcement learning framework in Sec. 2. And then Sec. 3 will go back to the sources of reward signals, discussing advantages and disadvantages of these different representations when learned with two ways mentioned before. After comparing, we draw a conclusion in Sec. 4 that one representation, intrinsic motivation is found higher generalization when explaining different behaviors, which may deserve more attention in future studies.

2 How to Learn the Reward Function

2.1 Reinforcement Learning Formulation

Reinforcement Learning (RL) formalizes the problem of goal-seeking with a wide and realistic range of goals and worlds corresponding to different reward signals to maximize in different environments.[11] Thus, we utilize this framework to illustrate two ways to obtain reward functions, which are consistent with "forward" RL and inverse RL respectively. Before this, the formulations of RL should be elaborated first:

2.1.1 Agent

An agent selects an action A_t at time t, given the sequence of observed states and actions $H_t = S_1, A_1, ..., S_{t-1}, A_{t-1}, S_t$, that have occurred in the history of interactions between the agent and the environment.

2.1.2 Environment

An environment is a system that the next state S_{t+1} is determined by transition rules $p(S_{t+1}|S_t, A_t)$.

2.1.3 Trajectory

A trajectory is a sequence of actions performed under the corresponding environment states $\pi(a|s)$.

2.1.4 Rewards

A reward is a special observation R_t emitted at every time-step t by a reward signal in the environment, which provides an instantaneous measurement of progress towards a goal. The RL problem is defined by a cumulative objective to maximize, commonly called the reward function $r(s, a)$, such as a sum of rewards or the average reward per time-step or other calculation forms.

2.2 Reward Engineering

Reward engineering, or "forward" reinforcement learning in the RL framework, needs designing rewards, *i.e.*, R_t at every time t and the reward function $r(s, a)$, by hand first. Given states $S \in s$, actions $A \in a$, transition rules $p(S_{t+1}|S_t, A_t)$, rewards and reward function, the system learns trajectories $\pi(a|s)$ that will maximize the reward function $r(s, a)$. Examples can be easily found in games, where scores are the most direct rewards and maximizing accumulated these scores is the goal. Zhong et al. [16] has designed rewards for a cooperative-competitive multi-agent active tracking game, enabling tracker to perform desired distraction-robust active visual tracking that can be well generalized to unseen environments. However, as stated earlier that rewards are always internal to people, it is hard to define reward function for most real world activities. Although manually designed reward functions are very interpretable, the challenge of designing has limited RL to applications with simple reward functions, and has been restricted to users who speak this language of mathematically-defined reward functions. On top of that, functions designed for a specific task are also difficult to generalize to other similar tasks.

2.3 Inverse Learning without Reward Engineering

Since reward engineering is not enough, inverse RL provides an approach to infer reward functions from existing examples. Such methods without reward engineering are usually performed through two steps: First given $S \in s$, actions $A \in a$, transition rules $p(S_{t+1}|S_t, A_t)$ and samples τ_i from $\pi(\tau)$, learn a separate model to represent rewards and reward functions $r(s, a)$; And then optimize this reward function with standard RL algorithms. This process is more common in human life, for example, Tom in BIGAI always sets the desk higher than everyone next to him, which demonstrates that the height of the desk is a reward signal to him when working and within a certain range the higher the desk is the greater the reward function value becomes. There are also methods (*e.g.*, Eysenbach et al. [3]) that do not learn a separate reward function, but learn to predict future success directly from transitions and success examples. If Mary saw Tom at KFC every time for dinner, the next time Tom asks Mary to take dinner for him, Mary will know KFC may be the best choice to maximize Tom's reward function. However, problems also exist for inverse learning. There are not always enough successful examples for all tasks and successful examples are not always precisely optimal. Secondly, it is difficult to evaluate a learned reward because what it exactly means is unclear.

3 Where do these Reward Signals come from

After understanding the two basic ideas of acquiring reward functions and how to model decision-making through reward functions, we can return to the question of where these environment reward signals come from. As utility is both hierarchy and multidimensional,34 main approaches for representation, prior knowledge, human feedback, and intrinsic motivation will be discussed about their interpretability and generalization when being learned.

3.1 Prior Knowledge

Prior knowledge is information of rewards and reward functions, that will be given before the whole learning process. Such information can be provided as all reward engineering is completed. But more commonly, only part of the information will be provided in advance, *e.g.* types of rewards, forms of reward functions or some rules obeyed by rewards changing. For example, Wu et al. [15] has defined several kinds of sorting criteria, such as *Category* (sort objects based on general categories, *e.g.*, put clothes here and toys there), *Attribute* (sort objects based on object attributes, *e.g.*, put plastic items here and mental items there) and so on. Then learn preferences for these criteria. The most significant advantage of prior knowledge is full interpretability, as anything learned has been given a meaning in the real world. However, it is hard to provide completely accurate and comprehensive knowledge beforehand so its generalization is poor. Still take sorting as an example, maybe many people sort objects by categories or attributes, but there are just a few people who sort by color, so it will be not practical to stick to prior criteria.

3.2 Human Feedback

To address the problem that just demonstration is not enough, researchers have switched to inverse learning rewards from various forms of human feedback, mainly corrections, preferences and rankings. Learning from corrections is learning from the process that the person modifies an existing trajectory and changes it to a new configuration. The agent has to estimate what trajectory the person might have intended given its current trajectory and the corrected configuration.[7] And learning from preferences (rankings) is through querying the person for preferences between two trajectories (ranking a set of trajectories in the order of preference). But such algorithms are effective only in smaller action spaces as it takes a long time to narrow in on the rough region of space where the true reward function lies.[8] To overcome their respective limitations, several approaches are also proposed to combine two or three kinds of human feedback or integrate prior demonstrations, *e.g.*, Bıyık et al. [2] has integrated demonstrations and preferences, and Mehta and Losey [7] has proposed a formalism that unites demonstrations, corrections, and preferences. Although feedback is easy for a person to provide, what exactly does the agent learn lacks interpretability. In other words, the agent could learn that the person prefers or does not prefer one trajectory, but couldn't know why.

3.3 Intrinsic Motivation

Psychology experiments have shown that human behaviors are driven by intrinsic motivation that is rooted early in infants' cognition, *e.g.*, curiosity, surprise, prosociality, morality and so on. Take prosociality as an example, it can serve as the motivation for social intelligence such as helping, sharing, and collaborating. infants as young as 18 months of age quite readily help others to achieve their goals in a variety of different situations, even when they receive no immediate benefit and the person helped is a stranger.[14] And infants selectively avoid helping those who cause or even intend to cause others harm because their prosocial behavior is mediated by the intentions behind the actor's moral behavior, irrespective of outcome.[13] Moreover, Tomasello et al. [12] has proposed that humans are biologically adapted for participating in collaborative activities inolving shared goals and socially coordinated action plans. And the fairness and equity drive children to share a resource produced through collaboration equitably at an earlier age.[4] Such intrinsic motivation provides a more abstract and more generalized explanation for behaviors in everyday life.

4 Towards a Unified Represetation of Intrinsic Motivation

After comparing the interpretability and generalization of these 3 reward signal sources, it could be concluded that intrinsic motivation can generally explain the vast majority of behavioral phenomena in humans. Since it could not only provide several dimensions of internal reward beforehand, but the degree of attention towards different dimensions also could be learned inversely, representing human utility by intrinsic motivation is a worthy research direction. There have also been some studies trying to model one dimension. Pathak et al. [9] formulates curiosity as the error in an agent's ability to predict the consequence of its own actions in a visual feature space learned by a self-supervised inverse dynamics model. Surprise minimizing reinforcement learning (SMiRL) [1] formalizes the phenomenon that every organism struggles against disruptive environmental forces to carve out and maintain an orderly niche into an unsupervised RL method. Furthermore, Ma et al. [6] allows agents to learn collaborative behaviors without any external reward but with expectation alignment with their neighbors. However, we can see that different dimensions are also modeled separately, and it lacks a unified representation of these different dimensions although they always influence each other and work together, which is what we need to focus more on in the future.

References

[1] Glen Berseth, Daniel Geng, Coline Devin, Nicholas Rhinehart, Chelsea Finn, Dinesh Jayaraman, and Sergey Levine. Smirl: Surprise minimizing reinforcement learning in unstable environments. *arXiv preprint arXiv:1912.05510*, 2019. 4

[2] Erdem Bıyık, Dylan P Losey, Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. *The International Journal of Robotics Research*, 41(1):45–67, 2022. 3

[3] Ben Eysenbach, Sergey Levine, and Russ R Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. *Advances in Neural Information Processing Systems*, 34:11541–11552, 2021. 2

[4] Katharina Hamann, Felix Warneken, Julia R Greenberg, and Michael Tomasello. Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, 476(7360):328–331, 2011. 4

[5] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604, 2016. 1

[6] Zixian Ma, Rose Wang, Fei-Fei Li, Michael Bernstein, and Ranjay Krishna. Elign: Expectation alignment as a multi-agent intrinsic reward. *Advances in Neural Information Processing Systems*, 35:8304–8317, 2022. 4

[7] Shaunak A Mehta and Dylan P Losey. Unified learning from demonstrations, corrections, and preferences during physical human-robot interaction. *arXiv preprint arXiv:2207.03395*, 2022. 3

[8] Malayandi Palan, Nicholas C Landolfi, Gleb Shevchuk, and Dorsa Sadigh. Learning reward functions by integrating human demonstrations and preferences. *arXiv preprint arXiv:1906.08928*, 2019. 3

[9] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017. 4

[10] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. 1

[11] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 2

[12] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691, 2005.

[13] Amrisha Vaish, Malinda Carpenter, and Michael Tomasello. Young children selectively avoid helping people with harmful intentions. *Child development*, 81(6):1661–1669, 2010.

[14] Felix Warneken and Michael Tomasello. Altruistic helping in human infants and young chimpanzees. *science*, 311(5765):1301–1303, 2006.

[15] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *arXiv preprint arXiv:2305.05658*, 2023.

[16] Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. Towards distraction-robust active visual tracking. In *International Conference on Machine Learning*, pages 12782–12792. PMLR, 2021.