

УДК 004.81/.89

Функциональная модель эмоций в обучении с подкреплением

Лятифов Роман Эйнулла оглы^(*)relyatifov@mail.ru
SPIN-код: 6774-2090

Косыгин Егор Владимирович

egor_tech23@mail.ru
SPIN-код: 8561-7383

Савельева Ирина Денисовна

saveleva.i@bmstu.ru

Аннотация. В работе показано, что эмоции могут быть формализованы как внутренние сигналы в архитектурах обучения с подкреплением (RL), влияющие на адаптацию и выбор стратегий поведения агента. Исследование направлено на построение когнитивных моделей, связывающих эмоции с ошибками предсказания, функцией ценности, неопределенностью и мотивационными факторами.

Ключевые слова: когнитивное моделирование, эмоции, обучение с подкреплением, адаптивное поведение, эмоциональный ИИ

A functional model of emotions in reinforcement learning

Latifov Roman Einulla oglu^(*)relyatifov@mail.ru
SPIN-code: 6774-2090

Kosygin Egor Vladimirovich

egor_tech23@mail.ru
SPIN-code: 8561-7383

Savelieva Irina Denisovna

saveleva.i@bmstu.ru

BMSTU, Moscow, Russia

Abstract. The paper shows that emotions can be formalized as internal signals in reinforcement learning (RL) architectures that influence the adaptation and choice of agent behavior strategies. The research aims to build cognitive models linking emotions with prediction errors, value function, uncertainty, and motivational factors.

Keywords: cognitive modeling, emotions, reinforcement learning, adaptive behavior, emotional AI

Введение. Эмоции эволюционно возникли как адаптивные механизмы, усиливающие выживание организма. Современные когнитивные теории (predictive coding, теория мотивационно-потребностной природы эмоций) показывают, что эмоции выполняют функцию сигнала несоответствия ожиданий и реальности, корректируя поведение [1, 2]. В машинном обучении эмоции чаще рассматриваются как задача классификации, однако их можно описывать и как архитектурные механизмы RL, определяющие стратегии агента [3].

Материалы и методы. Исследование основано на анализе работ по когнитивным наукам, психологии эмоций и алгоритмам RL. В качестве базовых концепций использованы: модели RL (функция ценности, prediction error, не-

определенность) [4]; психоэволюционная теория эмоций Р. Плутчика [5]; теория потребностно-информационной природы эмоций П.В. Симонова [6]; Байесовская механика и теория марковских одеял [2]; теория двойной обработки информации (Канемана) К. Канемана [7].

Формализация эмоций в терминах RL может быть представлена следующим образом:

- ошибка предсказания (δ) интерпретируется как источник положительных эмоций (радость, предвкушение) при высоких положительных значениях и отрицательных эмоций (страх, гнев, печаль) при высоких отрицательных значениях;

- функция ценности $V(s)$ отражает эмоциональную оценку состояния: высокие значения соответствуют доверию и радости, низкие — страху и печали;

- уровень неопределенности связан с эмоциями: высокая неопределенность соответствует удивлению и любопытству, низкая — доверию и спокойствию;

- внутренние награды трактуются как эмоциональные сигналы, изменяющие параметры обучения и способствующие адаптивному поведению.

Результаты. Разработана концептуальная модель, раскрывающая соответствия эмоций RL-механизмам (например, страх $\leftrightarrow$ состояния с низкой Q-ценностью, радость $\leftrightarrow$ положительное подкрепление, удивление $\leftrightarrow$ ошибка предсказания). Рассмотрены теоретические сценарии экспериментов: «страх» приводит к избеганию и замедлению траекторий; «удивление» инициирует остановку и переоценку модели; «любопытство» стимулирует исследование новых зон. Такая архитектура демонстрирует более адаптивное и выразительное поведение по сравнению с классическим RL.

Обсуждения. Полученные результаты подтверждают гипотезу о том, что эмоции можно трактовать как вычислительные механизмы адаптации. Это согласуется с теориями предсказательного кодирования и когнитивными моделями ценности. В отличие от традиционных моделей RL, использование «эмоциональных» сигналов улучшает устойчивость агента к неожиданным событиям и делает его поведение более интерпретируемым для человека.

Заключение. Эмоции могут быть формализованы в терминах обучения с подкреплением как сигналы ошибки предсказания, изменения ценности и уровня неопределенности. Такая интеграция позволяет создавать агентов с более гибким, адаптивным и «человеко-понятным» поведением, открывая перспективы для разработки эмоционального ИИ и когнитивной робототехники.

Список источников

- [1] Friston K. A theory of cortical responses. Philosophical transactions of the Royal Society B: Biological sciences, 2005, vol. 360, no. 1456, pp. 815–836.
- [2] Ramstead M.J.D. et al. On Bayesian mechanics: a physics of and by beliefs. Interface Focus, 2023, vol. 13, no. 3, art. no. 20220029.

-
- [3] Ляtifов Р.Э. Обучение с подкреплением как универсальный механизм когнитивного моделирования. *Студенческая научная весна. Всерос. студенческая конф., посв. 85-летию со дня рождения академика И.Б. Федорова: сб. тез. докл.* Москва, Издательский дом «Научная библиотека», 2025, с. 555–556.
 - [4] Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. MIT Press, 2018.
 - [5] Plutchik R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 2001, vol. 89, no. 4, pp. 344–350.
 - [6] Симонов П.В. *Мотивированный мозг*. Москва, Наука, 1987, 266 с.
 - [7] Канеман Д. *Думай медленно... решай быстро*. Москва, АСТ, 2018, 653 с.