

УДК 004.8

Когнитивные модели мозга и их роль в создании интерпретируемого искусственного интеллекта

Лятифов Роман Эйнулла оглы lyatifov@bmstu.ru

МГТУ им. Н.Э. Баумана, Россия, Москва

Аннотация. Когнитивные модели мозга играют ключевую роль в разработке интерпретируемого и безопасного искусственного интеллекта (ИИ). Эти модели предоставляют возможность формализации абстрактных гипотез о работе разума и их интеграции в когнитивные архитектуры ИИ. В данной работе рассматриваются Байесовская модель мозга, теория предсказательного кодирования и теория двойной системы разума, а также их применение для создания человекоподобных ИИ, способных адаптироваться, учиться и более эффективно взаимодействовать с человеком.

Ключевые слова: когнитивное моделирование, искусственный интеллект, когнитивные архитектуры

Cognitive models of the brain and their role in creating interpreted artificial intelligence

Latifov Roman Einulla oglu lyatifov@bmstu.ru

BMSTU, Moscow, Russia

Abstract. Cognitive models of the brain play a key role in the development of interpretable and secure artificial intelligence (AI). These models provide an opportunity to formalize abstract hypotheses about the workings of the mind and integrate them into cognitive AI architectures. This paper examines the Bayesian model of the brain, the theory of predictive coding and the theory of the dual mind system, as well as their application to create human-like AI capable of adapting, learning and interacting more effectively with humans.

Keywords: cognitive modeling, artificial intelligence, cognitive architectures

Введение. Когнитивные науки изучают работу человеческого мозга, и с развитием нейронных сетей они стали важным связующим звеном между «мягкими» и точными науками. Когнитивные модели не только объясняют методы машинного обучения, но и вдохновляют создание человекоподобных когнитивных архитектур для безопасного ИИ. В данной работе рассматриваются Байесовская модель мозга, теория предсказательного кодирования и теория двойной системы, а также их применение для разработки когнитивных архитектур ИИ.

Материалы и методы; результаты. Особая роль когнитивных наук заключается в объяснении и интерпретации механизмов работы разума, что

открывает возможность для создания интерпретируемых моделей искусственного интеллекта.

Одним из ключевых направлений является Байесовская модель мозга [1], которая описывает нашу систему познания как пространство убеждений, взаимосвязанных между собой и постоянно накапливаемых в процессе взаимодействия с окружающей средой. Другими словами, любая концепция, или модель (объект, явление, идея), в наших представлениях имеют некоторую степень верности (убежденности). И каждый раз при взаимодействии с реальным объектом или явлением, отражающим этот концепт, мы подтверждаем или подвергаем сомнению свои убеждения относительно модели этого концепта.

Общая теория разума описывает мозг как предсказательную машину (predictive coding theory) [2], минимизирующую «свободную энергию» для поддержания устойчивого состояния [3]. То есть, мы предсказываем некоторое будущее на основе нашей модели Мира, и, если между прогнозом и действительностью (наблюдением) существует различие (неожиданность), мы обновляем параметры своей модели Мира, стремясь минимизировать этот диссонанс между ожиданиями и реальностью.

В психологии разума существует теория о двойной системе (dual process theory) [4], согласно которой мозг работает в двух режимах: интуитивном (быстрое, автоматическое принятие решений) и осознанном (глубокий анализ, осмысление новых данных, мышление и сомнение).

Первая система отражает работу современных нейронных предобученных сетей, или прямой проход в обучении (forward pass). Вторая система по сути является режимом обновления весов нашей предсказательной модели. Она активируется при появлении неожиданности в наблюдениях, и, аналогично нейронным сетям, начинает в некотором виде распространять обратную ошибку (backward pass) для корректировки предсказательной модели.

Применение описанных теорий в современных нейронных сетях приводит к созданию систем, способных к непрерывному обучению и адаптации. Например, алгоритмы обучения с подкреплением позволяют системам обновлять свои предсказательные модели на основе полученных наград или ошибок. Это создает когнитивные архитектуры, которые не только эффективно решают задачи, но и улучшают свои предсказательные способности, аналогично человеческому мозгу.

Заключение. Эти и другие модели, и принципы, взятые из теории разума, психологии мышления, нейробиологии, кибернетики и методов машинного обучения (включая обучение с подкреплением), служат источником вдохновения для разработки новых когнитивных архитектур искусственного интеллекта. Стремление к созданию ИИ, который мыслит и действует по-человечески, оправдано не только улучшением его способности понимать человека, но и повышением безопасности взаимодействия между ИИ и людьми — одной из ключевых задач в развитии современных интеллектуальных систем.

Список источников

- [1] 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.
- [2] Friston K. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 2005, vol. 360, no. 1456, pp. 815–836.
- [3] Friston K. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 2010, vol. 11, no. 2, pp. 127–138.
- [4] Канеман Д. *Думай медленно... решай быстро*. Москва, АСТ, 2018, 653 с.