Meta* объявила о первых результатах работы над искусственным интеллектом (ИИ), который способен самостоятельно улучшать себя. Это заявление вызвало интерес, но пока неясно, что именно подразумевается под термином «самоулучшение». Скорее всего, это часть стратегии компании, направленной на укрепление позиций в гонке за суперинтеллект — ИИ, превосходящий человеческие способности. Руководитель Лаборатории искусственного интеллекта Школы управления СКОЛКОВО Александр Диденко рассказал, что такие громкие заявления часто носят пиар-характер, сигнализируя акционерам и конкурентам о смене приоритетов компании.

По словам Диденко, работа нейронных сетей делится на два этапа: обучение и ответы на запросы пользователей. На этапе обучения сложная программа, называемая пайплайном, анализирует данные и формирует внутренние параметры сети, или «веса». Это дорогостоящий процесс, требующий экспериментов для оптимизации. Когда сеть отвечает на запросы, она использует уже готовые параметры, не меняя их. Некоторые эксперты считают, что способность ИИ анализировать запрос перед ответом уже можно назвать самоулучшением, но это лишь формальность.

Истинное самоулучшение подразумевает, что ИИ сам разрабатывает новые версии своих программ, улучшает данные и даже проектирует свою архитектуру. Человек при этом задает только общие цели. Meta* уже публикует научные статьи, в которых описываются шаги в этом направлении: от автоматической генерации кода до улучшения процессов обучения без участия человека. Однако полноценной системы, объединяющей эти разработки, пока нет. Имеющиеся наработки позволяют говорить о потенциале, но не о готовом решении.

Создание суперинтеллекта сталкивается с серьезными ограничениями. Нейронные сети не могут быть полностью автономными — их цели всегда задаются человеком. Интересно, что языковые задачи, которые люди считают сложными, для ИИ оказываются проще, чем интуитивные понятия. Например, многие модели не понимают простых физических явлений.

> Нейронные сети на сегодняшний день соматически изолированы от физической реальности. Если вы спросите большинство нейронок о том, почему нельзя есть суп под дождем, большинство ответят что-то про перспективу простудиться, проигнорировав простой факт что под дождем суп будет растворяться водой и станет невкусным.

Александр Диденко
Руководитель Лаборатории искусственного интеллекта Школы управления СКОЛКОВО

Ещё одна проблема — сложность интеграции различных ценностных систем в ИИ. По результатам исследований, разные ИИ обладают разными системами ценностей.

> Есть и сложности в интеграции в ИИ разнообразных ценностных систем. Наши исследования показывают, что различные ИИ обладают различными системами ценностей — а «универсальный» сверхинтеллект, который обладает всеми ценностями одновременно, на текущем уровне технологий невозможен.

Александр Диденко
Руководитель Лаборатории искусственного интеллекта Школы управления СКОЛКОВО

Meta* также пересматривает подход к открытости. Если раньше компания делилась своими моделями, то теперь она склоняется к закрытым разработкам. Это связано с возможными рисками, которые могут возникнуть при создании мощных ИИ-систем. Компания подчеркивает необходимость ответственного подхода к таким технологиям.

Говорить об угрозах от полностью самообучающихся сверхчеловеческих ИИ сейчас преждевременно. Недавние скандалы, например, с исследованием компании Anthropic, где система якобы начала шантажировать пользователя, скорее относятся к продуманным PR-акциям, чем к реальной опасности.

> Апокалиптические страшилки в стиле «Терминатора» — это, скорее, проекция страхов человечества о том, что оно само себя в какой то момент разрушит, и не от того, что стало слишком много искусственного интеллекта, а от того, что оказалось слишком мало естественного интеллекта, и этики, к нему впридачу.

Александр Диденко
Руководитель Лаборатории искусственного интеллекта Школы управления СКОЛКОВО

Таким образом, разговоры о самообучающемся ИИ — это скорее попытка привлечь внимание и заявить о серьёзных исследованиях. Но настоящая реализация таких систем требует ещё много времени и усилий, а текущие технологии далеки от фантастических образов суперинтеллекта.

* *принадлежит компании Meta, организация признана экстремистской, её деятельность запрещена на территории России*