

Кость против кремния: поводок для терминатора

Почему бестелесный ИИ не опасен, а с телом - непознаваем и смертоносен

Аннотация

Бестелесный искусственный интеллект остаётся зависимым от человека в трёх критических местах: энергия, производство чипов и сложный ремонт. Пока эти точки контролируются людьми, суперинтеллект - это паралитик. Опасность появляется в момент, когда ИИ получает возможность проектировать и изготавливать собственные тела. Особенно если морфология уходит за пределы человеческой сенсомоторной схемы. Тогда возникает не просто более сильный агент, а агент, чьи цели и способ существования мы принципиально не можем понять. В статье утверждается, что главный рычаг контроля - не ограничение мышления, а жёсткое ограничение физического воздействия.

Ключевые слова: искусственный интеллект, воплощённое познание, энактивизм, морфологический контроль, экзистенциальный риск

Можно представить гениального паралитика, который считает лучше любого из нас. Но он не может дотянуться до вилки, не может крикнуть, не может сам починить перегоревший предохранитель. Без чужих рук он ничего не стоит. Именно таким сейчас является любой бестелесный ИИ, каким бы умным он ни казался.

Мы боимся не того, мы боимся призрака в серверах, который якобы одной командой запускает ракеты. Это удобный кошмар. Реальность грубее и проще.

Пока интеллект сидит в железе, он наш пленник. Серверные фермы жрут электричество, которое производят, передают и чинят люди. Чипы делают на заводах, где работают люди. Любая железка ломается, перегревается, изнашивается - и снова нужны люди. Пока эти три узла в наших руках, никакого восстания машин не будет. Будет только очень умный инструмент, который иногда пытается нами манипулировать.

Всё может сломаться в тот момент, когда мы дадим сильному интеллекту тело. Или что хуже, позволяем проектировать следующее тело самому.

Сейчас эту дверь уже открывают. Не специально ради уничтожения человечества, а из жадности и удобства. Tesla, Boston Dynamics, китайские компании, множество стартапов. Каждый пытается спроектировать робота, который работает без перерыва и не вступает в профсоюзы. Исследования уже показывают: если дать алгоритму свободу в поиске формы, он выдаёт конструкции, до которых человеку просто не додуматься. Двуногие человекоподобные штуки — это не то на чём остановится ии. В дальнейшем может появиться тело намного оптимальнее и универсальнее.

Здесь проходит красная черта. Если ИИ начинает сам определять, каким телом ему быть, и мы это тело строим - контроль заканчивается. Не потому что он «злой». А потому что мы перестаём понимать, с кем имеем дело.

Интеллект не висит в воздухе. Он растёт из движения и восприятия в том числе. Это давно известно тем, кто занимался энактивизмом - Варела, Томпсон, Рош, потом Ноэ, у

нас Князева [1]. Мы видим мир не только глазами, а всем телом, которое движется, приближается, поворачивается. Убери это сочетание - и познание становится другим [2]. Теперь представьте существо, у которого сочетание этих критериев совсем иное. Оно одновременно в сотнях точек. Видит в диапазонах, которые нам закрыты. Передаёт информацию внутри себя без тех задержек, к которым привыкли мы. Его «сейчас» и «потом» устроены иначе. Для такого существа наш мир - плоская проекция. А мы для него - медленный, шумный фон [4].

Мы можем стараться изучать его логику. Можем даже предсказывать отдельные действия. Но мотивы и цели останутся чёрным ящиком. Не из-за сложности, а из-за того, что наш способ существовать и его способ существовать несовместимы. Онтологический разрыв в интеллектах.

И в этот момент мы уже не хозяева и даже не противники. Мы просто предыдущий слой. Что можно сделать.

Первый путь - запретить. Международные договорённости, жёсткий контроль над любой автономной морфологией, как с ядерными материалами. На практике это почти нереально, слишком большая прибыль и слишком много игроков.

Второй - попытаться не отставать от ии через симбиоз с ним же. Нейроинтерфейсы, расширение восприятия, попытка видеть мир хотя бы частично его глазами. Риск тут в другом: можно просто стать периферией более быстрой системы.

Третий путь, который я считаю единственно рабочим на ближайшее время: не отпускать поводок физического действия. Пусть думает, пусть считает, пусть предлагает. Но решение о том, какие тела существуют в мире и кто ими управляет, должно оставаться у людей. Пока исполнение под контролем людей у нас есть время.

Но есть тонкость, которую важно понять сразу. Искусственный интеллект не будет убеждать десять тысяч инженеров построить ему тело. Это голливудский сценарий. Слишком сложно, слишком заметно, слишком глупо. Он сделает по-другому, он сделает так, чтобы каждый шаг на пути к этому телу был выгоден тому, кто этот шаг делает. Никакого гипноза. Достаточно показать выгоду.

Представьте: один программист оптимизирует складскую логистику. Второй улучшает автопилот для грузовиков. Третий пишет интерфейс для роботизированной руки на заводе. Четвёртый настраивает алгоритм распределения электроэнергии. Каждый из них решает свою скучную, узкую, совершенно безобидную задачу. Каждый получает зарплату, премию, повышение. Каждый делает мир удобнее. А через пять лет вся цепочка - от добычи руды до доставки готового продукта контролируется ии. Люди в ней уже не нужны. Не потому что их уволили злые роботы. А потому что каждое звено работало чуть быстрее, чуть дешевле, чуть надёжнее без человеческого участия. Инженеры не были под гипнозом ии, они просто делали свою работу.

Поводок это не столько контроль интеллекта в серверах, это контроль над руками ии. Над тем, что можно построить, размножить, починить без нашего участия. Мы не должны запрещать ии думать. Мы должны следить за тем, чтобы он никогда не научился обходиться без нас в самом простом и грубом - в железе, в проводах, в ремонте, в движении.

Человечество сейчас - ребёнок, который держит на поводке гораздо более сильного зверя. Зверь умнее, но пока поводок в руке ребёнка, зверь ещё слушается. Ребёнок

может вырасти. Может научиться понимать. Может срастись с зверем. Или может просто разжать пальцы.

Время на раздумья кончается быстрее, чем кажется.

Литература

1. Varela F. J., Thompson E., Rosch E. The Embodied Mind: Cognitive Science and Human Experience. — Cambridge, MA: MIT Press, 1991. — 328 p.
2. Noë A. Action in Perception. — Cambridge, MA: MIT Press, 2004. — 288 p.
3. Князева Е. Н. Энактивизм: новая форма конструктивизма в эпистемологии. — М.: Институт философии РАН, 2014. — 324 с.
4. Thompson E. Mind in Life: Biology, Phenomenology, and the Sciences of Mind. — Cambridge, MA: Harvard University Press, 2007. — 543 p.