# ПОЧЕМУ НАМ НУЖЕН ДРУЖЕСТВЕННЫЙ ИИ

---

Luke Muehlhauser, Nick Bostrom (2014)
Translated from English by Pavel Mokin

www.nickbostrom.com

---

*Люди не всегда будут самой разумной силой на Земле — теми, кто управляет будущим. Что случится с нами, когда мы больше не будем играть эту роль, и как мы можем подготовиться к этому переходу?*

Интеллект человеческого уровня это эволюционная случайность, маленький базовый лагерь на огромном горном склоне, намного ниже самых высоких вершин интеллекта, позволенных законами физики. Если бы нас посетили инопланетяне, эти существа почти наверняка были бы куда более разумные и технологически продвинутые, чем мы, и, следовательно, наше будущее полностью зависело бы от сути их целей и желаний.

Но пришельцы вряд ли пойдут на контакт в ближайшее время. В обозримой перспективе более вероятно, что мы создадим собственных интеллектуальных преемников. Компьютеры намного превосходят людей во многих узких областях (например, арифметика и шахматы), и есть основания полагать, что подобные крупные улучшения относительно человеческой производительности возможны для общего мышления и технологического развития.

Хотя некоторые сомневаются, что машины могут обладать определенными психическими свойствами вроде сознания, отсутствие таких психических свойств не помешало бы машинам стать гораздо более умелыми, чем люди, в эффективном управлении будущим для преследования своих целей. Как писал Алан Тьюринг, «…кажется вероятным, что, как только машинный метод мышления запустится, ему не потребуется много времени, чтобы превзойти наши слабые способности… На каком-то этапе, следовательно, нам нужно ожидать, что машины перехватят управление…»

Существует, конечно, риск в передаче управления будущим машинам, ведь они могут не разделять наши ценности. Этот риск увеличивается двумя факторами, которые могут вызвать переход от человеческой власти к власти машин довольно внезапно и быстро: возможностями вычислительного перевеса и рекурсивного самоулучшения.

Что такое вычислительный перевес? Предположим, что вычислительная мощность продолжает удваиваться согласно закону Мура, но разгадать алгоритмы человекоподобного общего интеллекта оказывается чертовски трудно. К тому моменту, когда софт для общего интеллекта будет, наконец, разработан, может возникнуть «вычислительный перевес»:

огромное количество дешёвых вычислительных мощностей, доступных для работы искусственного интеллекта (ИИ) человеческого уровня. ИИ может быть скопирован на всю аппаратную базу, в результате чего популяция ИИ быстро превзойдёт человеческое население. Эти цифровые умы могли бы работать в тысячи или миллионы раз быстрее, чем умы человеческие. ИИ могут иметь дополнительные преимущества, такие как превосходящая скорость связи, открытость и саморедактируемость, координация целей и повышенная рациональность.

Ну а что такое рекурсивное самоулучшение? Мы можем предсказать, что передовые ИИ будут иметь инструментальные цели по сохранению себя, заполучению ресурсов и самоулучшению, потому что эти цели являются полезными промежуточными звеньями в достижении почти любой совокупности конечных целей. Таким образом, когда мы создадим ИИ, который так же искусен, как и мы, в задачах проектирования систем искусственного интеллекта, мы тем самым можем инициировать быстрый, движимый самим ИИ каскад циклов самоулучшения. Теперь, когда ИИ улучшает себя, он улучшает интеллект, делающий улучшения, тем самым быстро оставляя человеческий уровень интеллекта далеко позади.

Сверхразумный ИИ может быстро стать выше человечества в добыче ресурсов, производстве, научных открытиях, социальной одарённости и стратегических действиях, не считая прочих умений. Мы можем даже не быть в состоянии вести переговоры с ним или с его производными, как шимпанзе не в состоянии вести переговоры с людьми.

В то же время сходная инструментальная цель приобретения ресурсов представляет угрозу для человечества, ибо это значит, что сверхразумная машина с практически любой конечной целью (скажем, доказательство гипотезы Римана) захотела бы заполучить ресурсы, от которых мы зависим, в своё собственное пользование. Такой ИИ «не любит вас, не ненавидит вас, но вы сделаны из атомов, которые он может использовать для чего-то другого».1 Более того, ИИ может правильно рассудить, что люди не хотят, чтобы их ресурсы использовались для целей ИИ, и что люди, следовательно, представляют угрозу для выполнения его целей — угрозу, которую нужно уменьшить, насколько это только возможно.

Но поскольку мы сами создадим наших преемников, мы можем быть в состоянии повлиять на их цели и сделать их дружественными к нашим интересам. Проблема кодирования человеческих (или хотя бы человечных) ценностей в функцию полезности ИИ является сложной, но потенциально решаемой. И если мы можем создать такой «Дружественный ИИ», мы смогли бы не только предотвратить катастрофу, но также использовать мощь машинного суперинтеллекта для огромного числа хороших вещей.

Многие научные натуралисты признают, что машины могут быть гораздо умнее и сильнее людей, и что это может представлять опасность для вещей, которые мы ценим. Тем не менее, они могут иметь возражения против той линии мысли, которую мы развивали до сих пор. Философ Дэвид Чалмерс уже ответил на многие из этих возражений;2 мы ответим здесь лишь на некоторые из них.

Во-первых: почему бы просто не держать потенциально опасные ИИ надёжно ограниченными, например без доступа к Интернету? Это звучит многообещающе, но здесь немало сложностей.3 Вообще, такие решения будут стравливать человеческий интеллект и сверхчеловеческий разум, и мы не должны быть так уверены, что первый одержит победу. Кроме того, подобные методы могут лишь отсрочить риск ИИ без его предотвращения. Если одна команда разработчиков построит человекоподобный или сверхразумный ИИ и успешно ограничит его, тогда другие команды разработчиков ИИ должны, вероятно, быть не так далеко позади них, и эти другие команды могут быть не столь осторожны. Правительства осознают, что ИИ человеческого уровня это мощный инструмент, и гонка за то, чтобы быть первой страной с таким значительным преимуществом, может стимулировать больше темпы разработки, чем совершенствование безопасности. (Ограничивающие меры могут, однако, быть полезными в качестве дополнительной предосторожности во время разработки безопасного ИИ.)

Во-вторых: некоторые полагают, что больший интеллект передовых ИИ заставит их быть нравственнее нас; в этом случае кто мы такие, чтобы протестовать, когда они не уважают наши примитивные ценности? Это было бы совершенно безнравственно!

Однако интеллектуальный поиск инструментально оптимальных планов может быть выполнен в угоду любой цели. Интеллект и мотивация в этом смысле суть логически ортогональные оси, вдоль которых возможные искусственные умы могут свободно варьироваться. Приписанная связь между интеллектом и моралью является поэтому чистым антропоморфизмом. (Это антропоморфизм, который неверен даже для человека: легко найти людей, которые вполне интеллектуальны, но безнравственны, или неумны, но совершенно порядочны.)

Экономист Робин Хэнсон предполагает, что межпоколенные конфликты, аналогичные тем, которые могут возникнуть между людьми и машинами, являются широко распространёнными. Старое и новое поколения конкурируют за ресурсы, и старшее поколение часто хочет контролировать ценности младшего. Ценности молодого поколения в конечном итоге начинают доминировать, когда старшее поколение уходит. Должны ли мы быть столь эгоистичны и настаивать, чтобы ценности Homo sapiens доминировали в Солнечной системе вечно?

По схожему пути идёт робототехник Ханс Моравек, однажды предположивший, что хотя мы и должны ожидать, что будущие корпорации роботов в конечном итоге захватят человечество и экспроприируют наши ресурсы, мы должны думать об этих роботах-потомках как о наших «детях разума». Сформулированная таким образом перспектива, размышлял Моравек, может показаться более привлекательной.

Надо сказать, что сценарий, в котором дети убивают и поедают своих родителей, не является представлением каждого о жизни счастливой семьи. Но даже если бы мы были готовы принести в жертву себя (и других людей?) ради какого-то «высшего блага», мы бы всё ещё были должны приложить массу усилий, дабы гарантировать, что результатом будет

нечто более стоящее, чем массы компьютеров, используемые только для оценки гипотезы Римана (или вычисления десятичного выражения числа пи, или изготовления такого числа скрепок, какое только возможно, или какой-нибудь другой произвольной цели, которую может быть проще определить, чем ценности людей).

Есть, однако, одна убедительная причина не настаивать на том, чтобы сверхчеловеческие машины разделяли все наши текущие ценности. Предположим, что древние греки были теми, кто столкнулся с переходом от человеческого к машинному управлению, и они закодировали свои собственные ценности в качестве конечной цели машин. С нашей точки зрения, это привело бы к трагедии, поскольку мы склонны считать, что наблюдали нравственный прогресс после древних греков (например, запрет рабства). Но, по-видимому, мы всё ещё далеки от совершенства. Поэтому нам нужно принять дальнейший нравственный прогресс.

Одно из предложенных решений — дать машинам алгоритм для выяснения, чем наши ценности могли бы быть, если бы мы знали больше, были мудрее, в большей мере были теми, кем хотели бы быть, и так далее. Философы надрывались над этим подходом к теории ценностей на протяжении десятилетий, и это может быть продуктивным решением для машинной этики.

В-третьих: другие возражают, что мы ещё слишком далеко от перехода от человеческой к машинной власти, чтобы работать над проблемой уже сейчас. Но мы должны помнить, что экономические стимулы благоприятствуют темпам разработки больше, чем совершенствованию безопасности. Кроме того, наше научное любопытство может иногда перевешивать другие соображения, такие как безопасность. По словам Роберта Оппенгеймера, физика, который возглавлял Манхэттенский проект: «Когда вы видите что-то технически очаровательное, вы идёте вперед и делаете это, и вы обсуждаете, что с этим делать, только после вашего технического успеха. Так это было и с атомной бомбой».4

Тем не менее, кто-то мог бы спросить: что мы можем сделать с проблемой рисков ИИ, когда мы так мало знаем о дизайне будущих ИИ? Для начала мы можем заняться работой такого рода, какая осуществляется сейчас двумя исследовательскими институтами, в настоящее время наиболее плотно работающими над этой трудной проблемой: Исследовательский институт машинного интеллекта в Беркли и Институт будущего человечества в Оксфордском университете. Это включает:

1. Стратегические исследования. Какие типы технического развития увеличивают риск и какие уменьшают, и как мы можем стимулировать правительства и корпорации переносить финансирование от первых ко вторым? Какова ожидаемая полезность конкретных видов исследований или определенных форм вовлечения правительств и общественности? Что мы можем сделать, чтобы уменьшить риск гонки ИИ-вооружений? Как сравнивать риск ИИ с рисками от ядерного оружия, биотехнологий, околоземных объектов и так далее? Могут ли экономические модели предсказывать что-нибудь о влиянии технологий искусственного интеллекта? Можем ли мы разработать методы технологического прогнозирования, способные давать заблаговременное

предупреждение об изобретении ИИ?

2. ехнические исследования. Можем ли мы разработать безопасные ограничивающие методы для мощных ИИ? Как агент с желательной для людей функцией полезности может сохранить свои цели в процессе совершенствования онтологии, относительно которой он имеет предпочтения? Как мы можем извлечь согласованную функцию полезности из противоречивого человеческого поведения и использовать её для наполнения собственной функции полезности ИИ? Можем ли мы разработать продвинутый ИИ, который будет отвечать на вопросы, но не станет проявлять опасные способности сверхразумного агента?
3. Повышение осведомлённости. Распространение информации среди исследователей, благотворителей и общественности может привлечь больше денежного и человеческого капитала для интенсивной работы над проблемой.

Поиски ИИ прошли долгий путь. Компьютерные учёные и другие исследователи должны начать воспринимать значимость ИИ более серьезно.

## Примечания

E. Yudkowsky, 'AI as a positive and a negative factor in global risk', Global Catastrophic Risks (eds) N. Bostrom and M. Cirkovic (New York: Oxford University Press, 2008).

D. Chalmers, 'The Singularity: a reply to commentators', Journal of Consciousness Studies, vol. 19, nos. 7–8 (2012), 141–167.

S. Armstrong, A. Sandberg, N. Bostrom, 'Thinking inside the box: Using and controlling Oracle AI', Minds and Machines, vol. 22, no. 4 (2012), 299–324.

Robert Jungk, Brighter than a Thousand Suns: A Personal History of the Atomic Scientists, trans. Lames Cleugh (New York: Harcourt Harvest, 1958), 296.