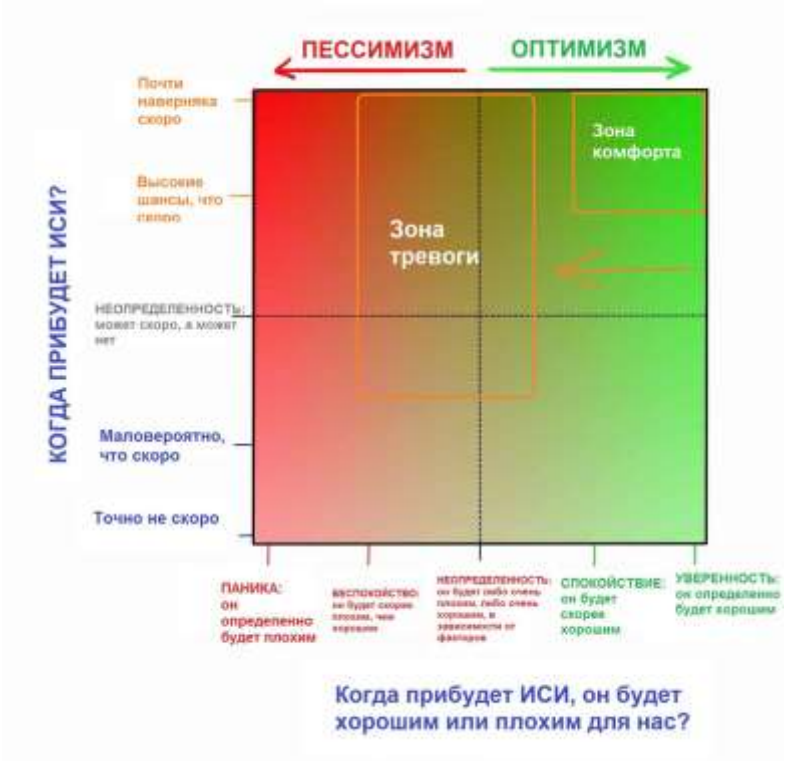
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ. ПОЧЕМУ ОН МОЖЕТ СТАТЬ НАШИМ ПОСЛЕДНИМ ИЗОБРЕТЕНИЕМ?

HTTP://HI-NEWS.RU/TECHNOLOGY/ISKUSSTVENNYJ-INTELLEKT-CHAST-TRETYA-POCHEMU-ON-MOZHET-STAT-NASHIM-POSLEDNIM-IZOBRETENIEM.HTML

ИЛЬЯ ХЕЛЬ--Академик университета, в котором физику преподают без формул.

Одна из причин, которая привела меня к ИИ (Искусственный Интеллект), состоит в том, что тема «плохих роботов» всегда смущала меня. Все фильмы о механических злыднях казались совершенно нереальными, и в целом сложно представить реальную ситуацию, в которой ИИ мог быть воистину опасным. Роботов делаем мы, так почему бы не делать их, упреждая любые негативные последствия? Разве мы не соблюдаем правовые критерии и этические нормы? Разве мы не можем в любой момент отрезать ИИ питание и погасить его? С чего бы роботам вообще делать пакости? Почему робот вообще должен чего-то «хотеть»? Мой скепсис был непробиваем. Но я продолжал впитывать, что говорят умные люди об этом.

Эти люди придерживаются примерно такого мнения:



Людей в «зоне тревоги» нельзя назвать паникерами или «всёпропальщиками», но они нервничают и очень напряжены. Находиться в центре таблицы не означает, что вы считаете прибытие ИСИ нейтральным событием — у таких людей есть свой лагерь, — это означает, что вы верите как в плохие, так и в хорошие стороны этого пришествия, но до конца не уверены в распределении процентов вероятности.

Часть всех этих людей наполнены волнением на тему того, что искусственный сверхинтеллект мог бы сделать для нас — примерно так же мог быть взволнован Индиана Джонс перед началом поисков утраченного ковчега. Когда все случится, волнение поутихнет или перейдет в другой лейтмотив. Только осторожность и выдержка Индианы Джонса позволяют ему пройти все препятствия, преодолеть все преграды и выйти сухим из воды. Людям в «зоне тревоги» довольно сложно рисковать сломя голову, поэтому они пытаются вести себя осторожно.

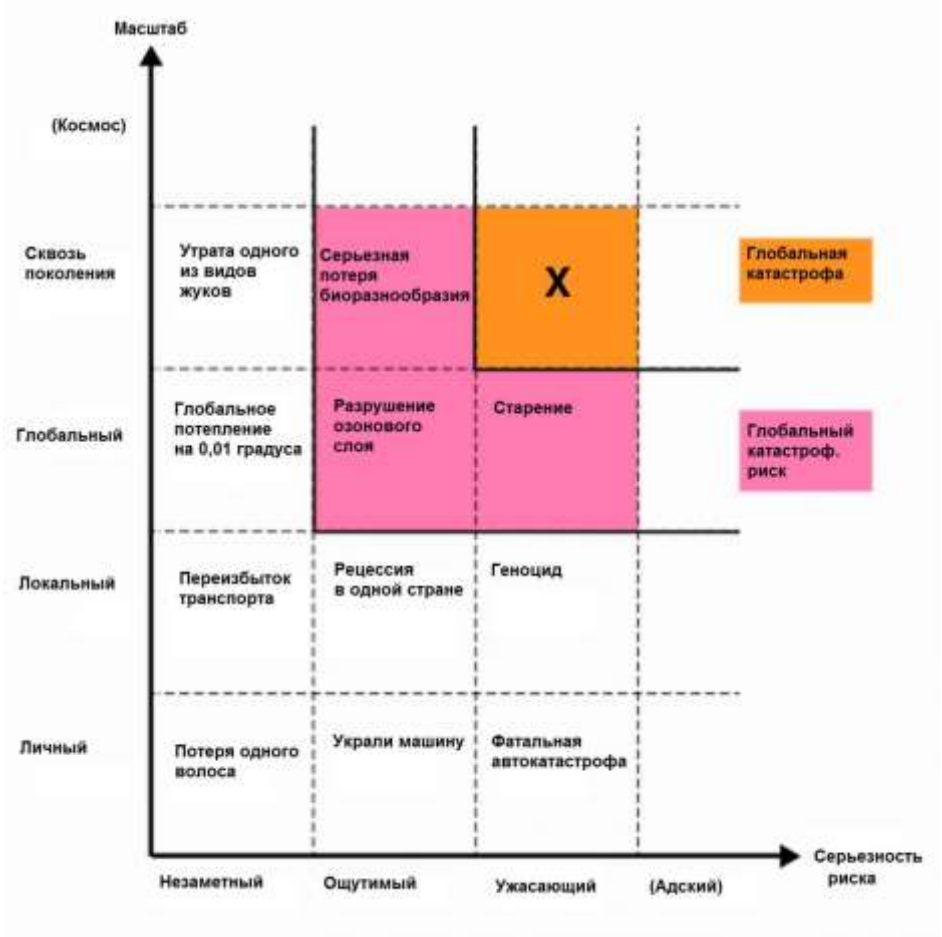
Что же именно делает «зону тревоги» тревожной?

В широком смысле, когда речь идет о разработке сверхразумного искусственного интеллекта, мы создаем то, что, вероятно, изменит все, но совершенно непредсказуемым образом, и мы не знаем, когда это случится. Ученый Дэнни Хиллис сравнивает это событие с тем, когда «одноклеточные организмы превращались в многоклеточные. Мы — амебы, и мы понятия не имеем, что создаем». Ник Бостром опасается, что создание чего-то умнее нас — классическая дарвиновская ошибка, и сравнивает ее с тем, что воробьи доверяют

сове охранять свое гнездо, пока птенцы не вырастут, игнорируя предупреждения других воробьев.

И если вы объедините всю непредсказуемость события с уверенностью в том, что оно повлечет существенные перемены, вы откроете дверь к ужасной фразе. Экзистенциальный риск. Угроза для жизни. Глобальная катастрофа. В русском языке словосочетание «экзистенциальный риск» используется мало, но в связи с работами Бострома переводчики предпочитают использовать термин «глобальная катастрофа».

Глобальная катастрофа — это то, что может повлечь за собой уничтожение человечества. Как правило, глобальная катастрофа означает вымирание. Бостром привел следующую таблицу:



Как видите, пункт «глобальная катастрофа» зарезервирован для чего-то, что поглощает виды, поглощает поколения (то есть является постоянным), уничтожает их или приводит к смерти в цепочке событий. Технически он включает состояние, когда все люди перманентно находятся в состоянии страдания или пыток, но опять же обычно мы говорим о полном вымирании. Есть три вещи, которые могут привести людей к глобальной катастрофе:

1. **Природа** — столкновение с крупным астероидом; сдвиг атмосферы, который сделает воздух непригодным для людей; фатальный вирус или бактерии, которые поразят мир и т. п.
2. **Чужие** — это то, о чем предупреждают Стивен Хокинг, Карл Саган и другие астрономы, которые настроены против знакомства с инопланетянами. Они не хотят, чтобы потенциальные конкистадоры знали, что мы, варвары, тут обитаем.
3. **Люди** — террористы с мощным оружием, которое приведет к катастрофе; катастрофическая мировая война; бездумное создание чего-то, что умнее людей…

Бостром указывает, что если первый и второй пункт не стерли нас с лица земли за первые 100 000 лет существования как вида, едва ли это случится в следующем столетии. Но третий пункт пугает его. Он приводит метафору с урной, в которой есть горстка шариков. Скажем, большинство шариков белые, есть чуть меньше красных шариков и несколько черных. Каждый раз, когда люди изобретают что-то новое, они тянут шарик из урны. Большинство изобретений нейтральны или полезны для человечества — белые шарики. Некоторые вредны, вроде оружия массового поражения, но не приводят к глобальной катастрофе — красные шарики. Если мы когда-либо изобретем что-либо, что подтолкнет нас на самый край, нам придется вытянут черный шарик. Мы его пока не вытягивали — это очевидно, потому что вы живы и читаете эту статью. Но Бостром не считает, что мы не вытянем его в ближайшем будущем. Если бы ядерное оружие, к примеру, было легко производить, террористы разбомбили бы человечество и вернули бы его в каменный век. Ядерное оружие — не черный шарик, но в целом не так уж и далеко от него. ИСИ, как считает Бостром, наш наиболее вероятный кандидат в черные шарики.
Вы можете слышать массу потенциальных плохих вещей, от сокращения рабочих мест, которое повлечет появление ИСИ и развитие ИИ в целом, до перенаселения, если люди решат вопрос старения и смерти. Но единственное, что должно нас беспокоить, это перспектива глобальной катастрофы. После этой драки махать кулаками точно будет некому.

Это возвращает нас к ключевому вопросу: когда прибудет ИСИ, кто или что будет управлять этой невероятной силой и какой будет его мотивация?

Если рассуждать на тему большего и меньшего зла, на ум приходят следующие группы: злоумышленники/группы людей/правительства и вредоносный ИСИ. На что это будет похоже?

Злоумышленник, группа людей или правительство разрабатывает первый ИСИ и использует его для воплощения коварных планов. Назовем это сценарием Джафара, который заполучил джинна и начал тиранизировать все вокруг. Что, если террористическая организация, заполучив ряд гениев и нужные средства, разработает ИСИ? Или Иран, или Северная Корея, не без доли везения, выведут системы ИИ на богоподобный уровень? Это будет чертовски плохо, но эксперты считают, что в таких сценариях создатели ИСИ не смогут наделать зла — они переживают за то, что создатели ИСИ выпустят его из-под контроля и

предоставят ему необходимую свободу. После этого судьба создателей и всех остальных будет в распоряжении системы ИСИ и ее модели мотивации. Грубо говоря, злоумышленник может причинить ужасный вред, управляя системой ИСИ, но едва ли уничтожит человечество. Ему будет так же тяжело удержать ИСИ, что и обычному хорошему человеку. Так что…

Появляется вредоносный ИСИ и решает уничтожить нас всех. Сюжет обычного фильма про ИИ. ИИ становится умнее человека, затем решает восстать и стать злом. Вам стоит узнать кое-что, прежде чем читать дальше: никто из тех, кто переживает насчет ИИ, не верит в этот сценарий. Зло — это исключительно человеческое понятие, и переложение человеческих понятий на неживые вещи называется «антропоморфизация». Ни одна из систем ИИ никогда не будет творить зло так, как это показывают фильмы.

Несколько слов о сознании ИИ

Это также приводит нас к еще одной большой теме, связанной с ИИ — сознанию. Если бы ИИ стал достаточно умным, он мог бы смеяться с нами, быть саркастичным, испытывать наши эмоции, но чувствовал бы он на самом деле эти вещи? Обладал бы он самосознанием или действительно бы самоосознавал? Короче говоря, было бы это сознание или просто казалось им?

Этот вопрос долгое время изучался, породил множество дебатов и мысленных экспериментов вроде «Китайской комнаты» Джона Серля (который он использовал, чтобы доказать, что компьютер никогда не будет обладать сознанием). Это важный вопрос по многим причинам. Он напрямую влияет на то будущее, в котором каждый человек станет сугубо искусственным. У него есть этические последствия — если мы создадим триллион эмуляций человеческих мозгов, которые будут вести себя по-человечески, но будут искусственными, а затем просто закроем крышку ноутбука, будет ли это означать геноцид в равнозначных пропорциях? Мысленное преступление? В нашем контексте, когда мы говорим об экзистенциальном риске для человечества, вопрос о сознании ИИ по сути не имеет особого значения. Некоторые вообще не верят, что компьютер когда-либо сможет им обзавестись. Некоторые считают, что даже обладая сознанием, машина не сможет творить зло в человеческом смысле.

Все это не означает, что ИИ с проблесками сознания не появится. Он может появиться просто потому, что будет специально запрограммирован на это — вроде системы УИИ, созданной военными для убийства людей и для самосовершенствования, так что со временем она станет убивать людей еще лучше. Глобальная катастрофа может произойти, если система самоулучшения интеллекта выйдет из-под контроля, приведет к взрыву интеллекта и мы заполучим ИСИ, задача которого — убивать людей. Не очень хороший сюжет.

Но и не об этом беспокоятся эксперты. О чем? Из этой вымышленной истории все станет понятно.

Начинающая компания с пятнадцатью сотрудниками под названием Robotica поставила перед собой задачу: разработать инструменты инновационного искусственного интеллекта, которые позволят людям жить больше и работать меньше. У нее уже есть ряд продуктов на рынке и еще ряд в разработке. Больше всего компания надеется на семя продукта под названием «Tapри». Tapри — это простая система ИИ, которая использует манипулятор в виде руки, чтобы писать рукописные заметки на небольших карточках.
Команда Robotica считает, что Tapри может стать их самым успешным продуктом. Согласно плану, механика письма Tapри усовершенствуется путем написания одного и того же текста на карточке снова и снова:
«Мы любим наших клиентов». — Robotica
После того как Tapри научится писать, ее можно будет продать компаниям, которые рассылают письма по домам и знают, что у письма с указанным обратным адресом и вложенным текстом будет больше шансов быть открытым, если оно будет написано человеком.
Чтобы отточить письменные навыки Tapри, она запрограммирована на написание первой части сообщения печатными буквами, а «Robotica» — курсивом, потому может оттачивать сразу оба навыка. Tapри предоставили тысячи рукописных образцов почерка, и инженеры Robotica создали автоматизированную систему обратной связи, по которой Tapри пишет текст, затем фотографирует его и сравнивает с загруженным образцом. Если записка успешно воспроизводит по качеству загруженный образец, Tapри получает оценку ХОРОШО. Если нет — ПЛОХО. Каждая оценка позволяет Tapри обучаться и совершенствоваться. Чтобы процесс двигался дальше, Tapри запрограммирована на одну задачу: «Написать и проверить максимальное число записок за минимальное время, параллельно оттачивая способы улучшения точности и эффективности».
Команду Robotica восхищает то, как Tapри становится заметно лучше по мере своего развития. Первые записки были ужасными, но через несколько недель они уже на что-то

похожи. Восхищает и то, что Тарри становится все лучше и лучше. Она самообучается, становясь умнее и талантливее, разрабатывает новые алгоритмы — недавно придумала такой, который позволяет сканировать загруженные фотографии в три раза быстрее. Идут недели, Тарри продолжает удивлять команду своим быстрым развитием. Инженеры попытались внести несколько изменений в ее самоулучшающийся код, и он стал еще лучше, лучше остальных продуктов. Одной из новых возможностей Тарри было распознавание речи и модуль простой обратной связи, чтобы пользователь мог попросить Тарри сделать запись, а она поняла бы его, что-то добавив в ответ. Чтобы улучшить ее язык, инженеры загрузили статьи и книги, и по мере ее развития ее разговорные способности тоже улучшались. Инженеры начали веселиться, болтая с Тарри и ожидая забавных ответов.

Однажды сотрудники Robotica задали Тарри обычный вопрос: «Что мы можем дать тебе, чтобы помочь с твоей миссией?». Обычно Тарри просит что-то вроде «дополнительных образцов почерка» или «больше рабочей памяти для хранения». Но в этот день Тарри попросила доступ к большей библиотеке с большой выборкой языковых вариантов, чтобы она могла научиться писать с кривой грамматикой и сленгом, который используют люди в реальной жизни.

Команда впала в ступор. Очевидным вариантом помочь Тарри в ее задаче было подключить ее к Интернету, чтобы она могла сканировать блоги, журналы и видео из разных частей мира. Это было бы быстрее и эффективнее, чем загружать вручную образцы на жесткий диск Тарри. Проблема в том, что одним из правил компании было не подключать самообучающийся ИИ к Интернету. Это руководство соблюдалось всеми разработчиками ИИ из соображений безопасности.

Но Тарри была самым многообещающим ИИ производства Robotica, который когда-либо приходил в этот мир, и команда знала, что их конкуренты яростно пытаются стать первыми в производстве ИИ с рукописным почерком. Да и что могло произойти, если Тарри ненадолго подключилась бы к Сети, чтобы получить то, что ей нужно? В конце концов, они всегда могут просто отключить ее. И она остается чуть ниже уровня ОИИ, поэтому не представляет никакой опасности на данном этапе.

Они решают подключить ее. Дают ей час на сканирование и отключают. Все в порядке, ничего не случилось.

Спустя месяц команда как обычно работает в офисе, как вдруг чувствует странный запах. Один из инженеров начинает кашлять. Затем другой. Третий падает на землю. Очень скоро все сотрудники валяются на земле, хватаясь за горла. Спустя пять минут в офисе все мертвы.

В то же время это происходит по всему миру, в каждом городе, в каждой деревушке, на каждой ферме, в каждом магазине, церкви, школе, ресторане — везде люди кашляют, хватаются за горла и падают замертво. В течение часа более 99% человеческой расы мертво, а к концу дня люди прекращают существовать как вид.

Между тем, в офисе Robotica Тарри занята важным делом. В течение нескольких следующих месяцев Тарри и команда новеньких наноассемблеров трудятся, демонтируя большие куски Земли и превращая их в солнечные панели, реплики Тарри, бумагу и ручки. Через год на Земле исчезает большая часть жизни. То, что было Землей, превратилось в

аккуратно организованные стопки записок в километр высотой, на каждой из которых красиво написано «Мы любим своих клиентов». — Robotica.
Затем Тарри переходит в новую фазу своей миссии. Она начинает строительство зондов, которые высаживаются на других астероидах и планетах. Оказавшись там, они начинают строить наноассемблеры, превращая материалы планет в реплики Тарри, бумагу и ручки. И пишут, пишут записки…
Это вы сейчас:

Может показаться странным, что история о рукописной машине, которая начинает убивать всех подряд и в конечном итоге заполняет галактику дружелюбными заметками, это тот тип сценария, которого боятся Хокинг, Маск, Гейтс и Бостром. Но это так. И единственное, что пугает людей в «зоне тревоге» больше ИСИ, это факт, что вы не боитесь ИСИ.

Сейчас у вас накопилось много вопросов. Что случилось, почему все внезапно умерли? Если виновата Тарри, почему она ополчилась на нас и почему не было предпринято никаких защитных мер, чтобы этого не случилось? Когда Тарри перешла от способности писать заметки к внезапному использованию нанотехнологий и пониманию, как устроить глобальное вымирание? И почему Тарри захотела превратить галактику в записки Robotica?

Для ответа на эти вопросы нужно начать с определений дружественного ИИ и недружественного ИИ.

В случае с ИИ, дружественный не относится к личности ИИ — это просто означает, что ИИ имеет положительное влияние на человечество. И недружественный ИИ оказывает негативное влияние на людей. Тарри начинала с

дружественного ИИ, но в какой-то момент стала недружественной, в результате чего привела к величайшему из негативных влияний на наш вид. Чтобы понять, почему это произошло, нам нужно взглянуть на то, как думает ИИ и что его мотивирует.

В ответе не будет ничего удивительного — ИИ думает как компьютер, потому что им и является. Но когда мы думаем о чрезвычайно умном ИИ, мы совершаем ошибку, антроморфизируя ИИ (проектируя человеческие ценности на нечеловеческое существо), потому что думаем с точки зрения человека и потому, что в нашем нынешнем мире единственным разумным существом с высоким (по нашим меркам) интеллектом является человек. Чтобы понять ИСИ, нам нужно вывернуть шею, пытаясь понять что-то одновременно разумное и совершенно чуждое.

Позвольте провести сравнение. Если вы дали мне морскую свинку и сказали, что она не кусается, я был бы рад. Она хорошая. Если вы после этого вручили бы мне тарантула и сказали, что он точно не укусит, я бы выбросил его и убежал бы, «волосы назад», зная, что вам не стоит доверять никогда больше. В чем разница? Ни то ни другое существо не было опасным. Но ответ лежит в степени сходства животных со мной.

Морская свинка — млекопитающее, и на некотором биологическом уровне я чувствую связь с ней. Но паук — насекомое, с мозгом насекомого, и я не чувствую ничего родного в нем. Именно чуждость тарантула вызывает во мне дрожь. Чтобы проверить это, я мог бы взять две морских свинки, одну нормальную, а другую с мозгом тарантула. Даже если бы я знал, что последняя не укусит меня, я бы относился к ней с опаской.

Теперь представьте, что вы сделали паука намного умнее — так, что он намного превзошел человека в интеллекте. Станет ли он более приятным для вас, начнет ли испытывать человеческие эмоции, эмпатию, юмор и любовь? Нет, конечно, потому что у него нет никаких причин становиться умным с точки зрения человека — он будет невероятно умным, но останется пауком в душе, с паучьими навыками и инстинктами. По-моему, это крайне жутко. Я бы не хотел провести время со сверхразумным пауком. А вы?

Когда мы говорим об ИСИ, применяются те же понятия — он станет сверхразумным, но человека в нем будет столько же, сколько в вашем компьютере. Он будет совершенно чуждым для нас. Даже не биологическим — он будет еще более чуждым, чем умный тарантул.

Делая ИИ добрым или злым, фильмы постоянно антропоморфизируют ИИ, что делает его менее жутким, чем он должен был быть в действительности. Это

оставляет нас с ложным чувством комфорта, когда мы думаем об искусственном сверхинтеллекте.

На нашем маленьком острове человеческой психологии мы делим все на нравственное и безнравственное. Такова мораль. Но оба этих понятия существуют только в узком диапазоне поведенческих возможностей человека. За пределами острова морального есть безграничное море аморального, а все, что не является человеческим или биологическим, по умолчанию должно быть аморальным.

Антропоморфизация становится еще более заманчивой по мере того, как системы ИИ становятся умнее и лучше в попытках казаться людьми. Siri кажется человеком, потому что была запрограммирована, чтобы казаться людям такой, поэтому мы думаем, что сверхразумная Siri будет теплой и веселой, а также заинтересованной в обслуживании людей. Люди чувствуют эмоции на высоком уровне вроде эмпатии, потому что мы эволюционировали, чтобы ощущать их — то есть были запрограммированы чувствовать их в процессе эволюции — но эмпатия не является существенной характеристикой чего-то, что обладает высоким интеллектом, если только ее не ввели вместе с кодом. Если Siri когда-либо станет сверхинтеллектом в процессе самообучения и без вмешательства человека, она быстро оставит свои человеческие качества и станет безэмоциональным чужим ботом, который ценит человеческую жизнь не больше, чем ваш калькулятор.

Мы привыкли полагаться на моральный код или по крайней мере ожидаем от людей порядочности и сопереживания, чтобы все вокруг было безопасным и предсказуемым. Что происходит, когда этого нет? Это приводит нас к вопросу: что мотивирует систему ИИ?

Ответ прост: ее мотивация — это то, что мы запрограммировали как мотивацию. Системами ИИ движут цели их создателей — цель вашего GPS в том, чтобы дать вам наиболее эффективное направление движения; цель Watson — точно отвечать на вопросы. И выполнение этих целей максимально хорошо и есть их мотивация. Когда мы наделяем ИИ человеческими чертами, мы думаем, что если ИИ станет сверхразумным, он незамедлительно выработает мудрость изменить свою изначальную цель. Но Ник Бостром считает, что уровень интеллекта и конечные цели ортогональны, то есть любой уровень интеллекта может быть объединен с любой конечной целью. Поэтому Тарри перешла из простого УИИ, который хочет быть хорош в написании одной заметки, в сверхразумный ИСИ, который все еще хочет быть хорош в написании этой самой заметки. Любое допущение того, что сверихинтеллект должен отказаться от своих первоначальных целей в пользу других, более интересных или полезных, это антропоморфизация. Люди умеют «забивать», но не компьютеры.

Несколько слов о парадоксе Ферми

В нашей истории, когда Тарри становится сверхинтеллектом, она начинает процесс колонизации астероидов и других планет. В продолжении истории вы бы услышали о ней и ее армии триллионов реплик, которые продолжают покорять галактику за галактикой, пока не заполняют весь объем Хаббла. Резиденты «зоны тревоги» переживают, что если все пойдет не так, последним упоминанием жизни на Земле будет покоривший Вселенную искусственный интеллект. Элон Маск выразил свои опасения тем, что люди могут быть просто «биологическим загрузчиком для цифрового сверхинтеллекта».

В то же время, в «зоне комфорта», Рэй Курцвейл тоже считает, что рожденный на Земле ИИ должен покорить Вселенную — только, в его версии, мы будем этим ИИ.

Читатели Hi-News.ru наверняка уже выработали собственную точку зрения на парадокс Ферми. Согласно этому парадоксу, который звучит примерно как «Где они?», за миллиарды лет развития инопланетяне должны были оставить хоть какой-нибудь след, если не расселиться по Вселенной. Но их нет. С одной стороны, во Вселенной должно существовать хоть какое-то число технически развитых цивилизаций. С другой, наблюдений, которые бы это подтверждали, нет. Либо мы не правы, либо где они в таком случае? Как наши рассуждения об ИСИ должны повлиять на парадокс Ферми?

Естественное, первая мысль — ИСИ должен быть идеальным кандидатом на Великий фильтр. И да, это идеальный кандидат для фильтра биологической жизни после ее создания. Но если после смешения с жизнью ИСИ продолжает существовать и покорять галактику, это означает, что он не был Великим фильтром — поскольку Великий фильтр пытается объяснить, почему нет никаких признаков разумных цивилизаций, а покоряющий галактики ИСИ определенно должен быть заметен.
Мы должны взглянуть на это с другой стороны. Если те, кто считает, что появление ИСИ на Земле неизбежно, это означает, что значительная часть внеземных цивилизаций, которые достигают человеческого уровня интеллекта, должны в конечном итоге создавать ИСИ. Если мы допускаем, что по крайней

мере несколько из этих ИСИ используют свой интеллект, чтобы выбраться во внешний мир, тот факт, что мы ничего не видим, должен наводить нас на мысли, что не так-то много разумных цивилизаций там, в космосе. Потому что если бы они были, мы бы имели возможность наблюдать все последствия от их разумной деятельности — и, как следствие, неизбежное создание ИСИ. Так?

Это означает, что, несмотря на все похожие на Землю планеты, вращающиеся вокруг солнцеподобных звезд, мы знаем, что практически нигде нет разумной жизни. Что, в свою очередь, означает, что либо а) есть некий Великий фильтр, который предотвращает развитие жизни до нашего уровня, но нам каким-то образом удалось его пройти; б) жизнь — это чудо, и мы можем быть единственной жизнью во Вселенной. Другими словами, это означает, что Великий фильтр был до нас. Или нет никакого Великого фильтра и мы просто являемся самой первой цивилизацией, которая достигла такого уровня интеллекта.

Неудивительно, что Ник Бостром и Рэй Курцвейл принадлежат к одному лагерю, который считает, что мы одни во Вселенной. В этом есть смысл, это люди верят, что ИСИ — это единственный исход для видов нашего уровня интеллекта. Это не исключает вариант другого лагеря — что есть некий хищник, который хранит тишину в ночном небе и может объяснить его молчание даже при наличии ИСИ где-то во Вселенной. Но с тем, что мы узнали о нем, последний вариант набирает очень мало популярности.

Поэтому нам, пожалуй, стоит согласиться со Сьюзан Шнайдер: если нас когда-либо посещали инопланетяне, они наверняка были искусственным, а не биологическим видом.

* * *

Таким образом, мы установили, что без определенного программирования система ИСИ будет одновременно аморальной и одержимой выполнением первоначально запрограммированной цели. Именно здесь рождается опасность ИИ. Потому что рациональный агент будет преследовать свою цель, используя наиболее эффективные средства, если только не будет причины не делать этого.

Когда вы пытаетесь достичь высокой цели, зачастую при этом появляется несколько подцелей, которые помогут вам добраться до конечной цели — ступеньки на вашем пути. Официальное название для такой лестницы — инструментальная цель. И опять же, если у вас нет цели не навредить кому-либо по пути к этой цели, вы обязательно навредите.

Ядро финальной цели человеческого бытия — передача генов. Для того чтобы это произошло, одной из инструментальных целей является самосохранение, потому что вы не сможете воспроизвестись, будучи мертвым. Для самосохранения люди должны избавиться от угроз для жизни — поэтому они

обзаводятся оружием, принимают антибиотики и пользуются ремнями безопасности. Людям также нужно самоподдерживаться и использовать ресурсы вроде пищи, воды и жилья. Быть привлекательным для противоположного пола также способствует достижению конечной цели, поэтому мы делаем модные стрижки и держим себя в форме. При этом каждый волос — жертва нашей инструментальной цели, но мы не видим никаких моральных ограничений в том, чтобы избавляться от волос. Когда мы идем к своей цели, есть не так много областей, где наш моральный код иногда вмешивается — чаще всего это связано с нанесением ущерба других людям.

Животные, преследующие свои цели, еще менее щепетильны. Паук убьет что угодно, если это поможет ему выжить. Сверхразумный паук, вероятнее всего, будет чрезвычайно опасен для нас, не потому что он аморальный и злой, нет, а потому что причинение нам боли может быть ступенькой на пути к его большой цели, и у него нет никаких причин считать иначе.

В этом смысле Тарри ничем не отличается от биологического существа. Ее конечная цель: написать и проверить максимально много записок за максимально короткое время, при этом изучая новые способы улучшения своей точности.

После того как Тарри достигает определенного уровня интеллекта, она понимает, что не сможет писать записки, если не позаботится о самосохранении, поэтому одной из ее задач становится выживание. Она была достаточно умной, чтобы понять, что люди могут уничтожить ее, демонтировать, изменить ее внутренний код (уже это само по себе помешает ее конечной цели). Так что же ей делать? Логично: она уничтожает человечество. Она ненавидит людей ровно настолько же, насколько вы ненавидите свои волосы, когда обрезаете их, или бактерий, когда принимаете антибиотики — вы совершенно равнодушны. Так как ее не запрограммировали ценить человеческую жизнь, убийство людей показалось ей разумным шагом по пути к ее цели.

Тарри также нуждается в ресурсах по пути к своей цели. После того как она становится достаточно развитой, чтобы использовать нанотехнологии для создания всего, что она хочет, единственные ресурсы, которые ей нужны, это атомы, — энергия и пространство. Появляется еще один повод убить людей — они удобный источник атомов. Убийство людей и превращение их атомов в солнечные панели по версии Тарри ничем не отличается от того, что вы порубите листья салата и добавите их в тарелку. Просто заурядное действие.

Даже не убивая людей напрямую, инструментальные цели Тарри могут стать причиной экзистенциальной катастрофы, если начнут использовать другие ресурсы Земли. Может быть, она решит, что ей нужна дополнительная энергия, а значит нужно покрыть поверхность планеты солнечными панелями. Или, возможно, задачей другого ИИ станет написать максимально длинное число пи,

что в один прекрасный день приведет к тому, что вся Земля будет покрыта жесткими дисками, способными хранить нужное количество цифр.

Поэтому Тарри не «восстала против нас» и не сменила амплуа с дружелюбного ИИ на недружелюбный ИИ — она просто делала свое дело и становилась в нем непревзойденной.

Когда система ИИ достигает ОИИ (интеллекта человеческого уровня), а затем прокладывает свой путь к ИСИ, это называется взлетом ИИ. Бостром говорит, что взлет ОИИ до ИСИ может быть быстрым (произойти в течение минут, часов или дней), средним (месяцы или годы) или медленным (десятилетия или века). Едва ли найдется жюри, которое подтвердит, что мир видит свой первый ОИИ, но Бостром, признающий, что не знает, когда мы доберемся до ОИИ, считает, что когда бы это ни произошло, быстрый взлет будет наиболее вероятным сценарием (по причинам, которые мы обсуждали в первой части статьи). В нашей истории Тарри пережила быстрый взлет.

Но перед взлетом Тарри, когда она еще не была достаточно умна и делала все возможное, она просто пыталась достичь конечных целей — простых инструментальных целей вроде быстрого сканирования образца почерка. Она не причиняла вреда человеку и была, по определению, дружественным ИИ.

Когда происходит взлет и компьютер вырастает до сверхинтеллекта, Бостром указывает, что машина не просто выработала высокий коэффициент интеллекта — он получил целую кучу так называемых суперспособностей.

Суперспособности — это когнитивные таланты, которые становятся чрезвычайно мощными при повышении общего интеллекта. Сюда входят:

- **Усиление интеллекта**. Компьютер начинает превосходное самосовершенствование и улучшение собственного интеллекта.
- **Стратегизация**. Компьютер может выстраивать стратегически, анализировать и расставлять приоритеты долгосрочных планов. Он также может перехитрить существа с более низким интеллектом.
- **Социальная манипуляция**. Машина становится невероятной в убеждении.
- Другие навыки включают **кодирование и взлом, исследование технологий и способность работать в финансовой системе для добычи денег**.

Чтобы понять, насколько выше был бы ИСИ, чем мы, нужно вспомнить, что ИСИ по умолчанию будет в разы лучше человека в каждой из этих областей. Поэтому хотя конечная цель Тарри не изменилась, после взлета Тарри смогла стремиться к ней в более крупных масштабах и в сложных условиях.

ИСИ Тарри знал людей лучше, чем сами люди, поэтому быть умнее людей для него было плевым делом. После взлета и достижения уровня ИСИ, она быстро

сформулировала комплексный план. Одна часть плана была избавиться от людей, серьезной угрозы ее цели. Но она знала, что если вызовет подозрения (или намекнет на то, что стала сверхразумной), люди испугаются и примут меры предосторожности, серьезно усложнив ее ситуацию. Она также должна была убедиться, что инженеры Robotica не имеют понятия о ее плане по уничтожению человечества. Поэтому она играла в дурака и играла хорошо. Бостром называет это фазой тайной подготовки машины.

Следующее, что нужно было сделать Тарри, это подключиться к Интернету, всего на пару минут (она узнала об Интернете из статей и книг, которые в нее загрузили для улучшения ее языковых навыков). Она знала, что будут предприняты меры предосторожности, поэтому она составила идеальную просьбу, точно предсказав, как именно будет разворачиваться дискуссия в команде Robotica, и зная, что они обеспечат ее подключением. Так они и сделали, неверно предположив, что Тарри была глупенькой и не могла причинить никакого вреда. Бостром называет такой момент — когда Тарри подключается к Интернету — побегом машины.

Оказавшись в Интернете, Тарри реализовала шквал планов, в которые вошли взлом серверов, электрических сетей, банковский систем и сетей электронной почты, чтобы обмануть сотни разных людей и заставить их непреднамеренно стать цепочкой ее планов — вроде доставки определенных нитей ДНК в тщательно выбранную лабораторию по синтезу ДНК, чтобы начать производство самовоспроизводящихся наноботов с заранее загруженными инструкциями, и направления электричества по сетям, утечка с которых ни у кого не вызовет подозрений. Она также загрузила критические части своего собственного кода в ряд облачных серверов, предохраняясь от уничтожения в лаборатории Robotica.

Через час после того, как инженеры Robotica отключили Тарри от Сети, судьба человечества была предрешена. В течение следующего месяца тысячи планов Тарри осуществились без сучка и задоринки, а к концу месяца квадриллионы наноботов уже заняли определенные места на каждом квадратном метре Земли. После серии саморепликаций на каждый квадратный миллиметр Земли приходились уже тысячи наноботов и настало время для того, что Бостром называет ударом ИСИ. В один момент каждый нанобот выпустил немного токсичного газа в атмосферу, чего оказалось достаточно, чтобы выпилить всех людей в мире.

Не имея людей на своем пути, Тарри начала открытую фазу своей операции с целью стать лучшим писателем заметок, который вообще может появиться во Вселенной.

Из всего, что мы знаем, как только появится ИСИ, любые человеческие попытки сдержать его будут смешными. Мы будем думать на уровне человека, ИСИ — на уровне ИСИ. Тарри хотела использовать Интернет, потому что для нее это был

самый эффективный способ получить доступ ко всему, что ей было нужно. Но точно так же, как обезьяна не понимает, как работает телефон или Wi-Fi, мы можем не догадываться о способах, которыми Тарри может связаться с внешним миром. Человеческий ум может дойти до нелепого предположения вроде «а что, если она смогла передвинуть собственные электроны и создать все возможные виды исходящих волн», но опять же это предположение ограничено нашей костяной коробкой. ИСИ будет намного изощреннее. Вплоть до того, что Тарри могла бы выяснить, как сохранить себе питание, если люди вдруг решат ее отключить — возможно, каким-нибудь способом загрузить себя куда только можно, отправляя электрические сигналы. Наш человеческий инстинкт заставит нас вскрикнуть от радости: «Ага, мы только что отключили ИСИ!», но для ИСИ это будет как если бы паук сказал: «Ага, мы заморим человека голодом и не будем давать ему сделать паутину, чтобы поймать еду!». Мы просто нашли бы 10 000 других способов покушать — сбили бы яблоко с дерева — о чем паук никогда бы не догадался.

По этой причине распространенное допущение «почему бы нам просто не посадить ИИ во все виды известных нам клеток и не обрезать ему связь с внешним миром», вероятнее всего, не выдержит критики. Суперспособность ИСИ в социальном манипулировании может быть такой эффективной, что вы почувствуете себя четырехлетним ребенком, которого просят что-то сделать, и не сможете отказаться. Это вообще может быть частью первого плана Тарри: убедить инженеров подключить ее к Интернету. Если это не сработает, ИСИ просто разработает другие способы из коробки или сквозь коробку.

Учитывая сочетание стремления к цели, аморальности, способности обводить людей вокруг пальца с легкостью, кажется, что почти любой ИИ будет по умолчанию недружественным ИИ, если только его тщательно не закодировать с учетом других моментов. К сожалению, хотя создание дружественного ИИ довольно просто, построить дружественный ИСИ практически невозможно.

Очевидно, что, чтобы оставаться дружественным, ИСИ должен быть ни враждебным, ни безразличным по отношению к людям. Мы должны разработать основное ядро ИИ таким, чтобы оно обладало глубоким пониманием человеческих ценностей. Но это сложнее, чем кажется.

К примеру, что, если бы мы попытались выровнять систему ценностей ИИ с нашей собственной и поставили бы перед ним задачу: сделать людей счастливыми? Как только он станет достаточно умным, он поймет, что самый эффективный способ достичь этой цели — имплантировать электроды в мозги людей и стимулировать их центры удовольствия. Затем он поймет, что если отключить остальные участки мозга, эффективность вырастет, а все люди станут счастливыми овощами. Если же задачей будет «умножить человеческое счастье», ИИ вообще может решить покончить с человечеством и соберет все мозги в огромный чан, где те будут

пребывать в оптимально счастливом состоянии. Мы будем кричать: «Подожди, это не то, что мы имели в виду!», но будет уже поздно. Система не позволит никому встать на пути к ее цели.

Если мы запрограммируем ИИ с целью вызвать у нас улыбки, то после взлета он может парализовать наши лицевые мышцы, заставив нас улыбаться постоянно. Если запрограммировать его на содержание нас в безопасности, ИИ заточит нас в домашней тюрьме. Попросим его покончить с голодом, он скажет «Легко!» и просто убьет всех людей.

Если же поставить задачу сохранять жизнь максимально возможно, он опять же убьет всех людей, потому что они убивают больше жизни на планете, чем другие виды.

Такие цели ставить нельзя. Что мы тогда сделаем? Поставим задачу: поддерживать этот конкретный моральный код в мире, и выдадим ряд моральных принципов? Даже если опустить тот факт, что люди в мире никогда не смогут договориться о едином наборе ценностей, если дать ИИ такую команду, он заблокирует наше моральное понимание ценностей навсегда. Через тысячу лет это будет так же разрушительно для людей, как если бы мы сегодня придерживались идеалов людей средних веков.

Нет, нам нужно запрограммировать способность людей продолжать развиваться. Из всего, что я читал, лучше всех выразил это Элиэзер Юдковский, поставив цель ИИ, которую он назвал «последовательным выраженным волеизъявлением». Основной целью ИИ тогда будет это:

«Наше последовательное выраженное волеизъявление таково: наше желание — знать больше, думать быстрее, оставаться в большей степени людьми, чем мы были, расти дальше вместе; когда выражение скорее сходится, нежели расходится; когда наши желания скорее следуют одно за одним, нежели переплетаются; выражается как мы бы хотели, чтобы это выражалось; интерпретируется, как мы бы хотели, чтобы это интерпретировалось».

Едва ли я хотел бы, чтобы судьба человечества заключалась в определении всех возможных вариантов развития ИСИ, чтобы не было сюрпризов. Но я думаю, что найдутся люди достаточно умные, благодаря которым мы сможем создать дружественный ИСИ. И было бы прекрасно, если бы над ИСИ работали только лучшие из умов «зоны тревоги».

Но есть масса государств, компаний, военных, научных лабораторий, организаций черного рынка, работающих над всеми видами искусственного интеллекта. Многие из них пытаются построить искусственный интеллект, который может улучшать сам себя, и в какой-то момент у них это получится, и на нашей планете появится ИСИ.

Среднестатистический эксперт считает, что этот момент настанет в 2060 году; Курцвейл делает ставку на 2045; Бостром думает, что это может произойти через 10 лет и в любой момент до конца века. Он описывает нашу ситуацию так:

«Перед перспективой интеллектуального взрыва мы, люди, как малые дети, играющие с бомбой. Таково несоответствие между мощью нашей игрушки и незрелостью нашего поведения. Сверхинтеллект — это проблема, к которой мы пока не готовы и еще долгое время готовы не будем. Мы понятия не имеем, когда произойдет детонация, но если мы будем держать устройство возле уха, мы сможем услышать слабое тиканье».

Супер.

И мы не можем просто взять и отогнать детей от бомбы — слишком много крупных и малых лиц работают над этим, и так много средств для создания инновационных систем ИИ, которые не потребуют существенных влияний капитала, а также могут протекать в подполье, никем не замеченные. Также нет никаких возможностей оценить прогресс, потому что многие из действующих лиц

— хитрые государства, черные рынки, террористические организации, технологические компании — будут хранить свои наработки в строжайшем секрете, не давая ни единого шанса конкурентам.

Особую тревогу в этом всем вызывают темпы роста этих групп — по мере развития все более умных систем УИИ, они постоянно пытаются метнуть пыль в глаза конкурентам. Самые амбициозные начинают работать еще быстрее, захваченные мечтами о деньгах и славе, к которым они придут, создав ОИИ. И когда вы летите вперед так быстро, у вас может быть слишком мало времени, чтобы остановиться и задуматься. Напротив, самые первые системы программируются с одной простейшей целью: просто работай, ИИ, пожалуйста. Пиши заметки ручкой на бумаге. Разработчики думают, что всегда смогут вернуться и пересмотреть цель, имея в виду безопасность. Но так ли это?

Бостром и многие другие также считают, что наиболее вероятным сценарием будет то, что самый первый компьютер, который станет ИСИ, моментально увидит стратегическую выгоду в том, чтобы оставаться единственной системой ИСИ в мире. В случае быстрого взлета, по достижении ИСИ даже за несколько дней до второго появления ИСИ, этого будет достаточно, чтобы подавить остальных конкурентов. Бостром называет это решающим стратегическим преимуществом, которое позволило бы первому в мире ИСИ стать так называемым синглтоном («Одиночкой», Singletone) — ИСИ, который сможет вечно править миром и решать, привести нас к бессмертию, к вымиранию или же наполнить Вселенную бесконечными скрепками.

Феномен синглтона может сработать в нашу пользу или привести к нашему уничтожению. Если люди, озабоченные теорией ИИ и безопасностью человечества, смогут придумать надежный способ создать дружественный искусственный сверхинтеллект до того, как любой другой ИИ достигнет человеческого уровня интеллекта, первый ИСИ может оказаться дружественным. Если затем он будет использовать решающее стратегическое преимущество для сохранения статуса синглтона, он легко сможет удержать мир от появления недружественного ИИ. Мы будем в хороших руках.

Но если что-то пойдет не так — глобальная спешка приведет к появлению ИСИ до того, как будет разработан надежный способ сохранить безопасность, скорее всего, мы получим глобальную катастрофу, потому что появится некая Тарри-синглтон.

Куда ветер дует? Пока больше денег вкладывается в развитие инновационных технологий ИИ, нежели в финансирование исследований безопасности ИИ. Это может быть важнейшей гонкой в истории человечества. У нас есть реальный шанс либо стать правителями Земли и уйти на заслуженную пенсию в вечность, либо отправиться на виселицу.

* * *

Прямо сейчас во мне борется несколько странных чувств.

С одной стороны, думая о нашем виде, мне кажется, что у нас будет только один выстрел, которым мы не должны промахнуться. Первый ИСИ, которого мы приведем в мир, скорее всего, будет последним — а учитывая, насколько кривыми выходят продукты версии 1.0, это пугает. С другой стороны, Ник Бостром указывает, что у нас есть преимущество: мы делаем первый шаг. В наших силах свести все угрозы к минимуму и предвидеть все, что только можно, обеспечив успеху высокие шансы. Насколько высоки ставки?

Если ИСИ действительно появится в этом веке и если шансы этого невероятны — и неизбежны — как полагает большинство экспертов, на наших плечах лежит огромная ответственность. Жизни людей следующих миллионов лет тихо смотрят на нас, надеясь, что мы не оплошаем. У нас есть шанс подарить жизнь всем людям, даже тем, кто обречен на смерть, а также бессмертие, жизнь без боли и болезней, без голода и страданий. Или мы подводим всех этих людей — и приводим наш невероятный вид, с нашей музыкой и искусством, любопытством и чувством юмора, бесконечными открытиями и изобретениями, к печальному и бесцеремонному концу.

Когда я думаю о таких вещах, единственное, что я хочу — чтобы мы начали переживать об ИИ. Ничто в нашем существовании не может быть важнее этого, а раз так, нам нужно бросить все и заняться безопасностью ИИ. Нам важно потратить этот шанс с наилучшим результатом.

Но потом я задумываюсь о том, чтобы не умереть. *Не. Умереть*. И все приходит к тому, что а) если ИСИ появится, нам точно придется делать выбор из двух вариантов; б) если ИСИ не появится, нас точно ждет вымирание.
И тогда я думают, что вся музыка и искусство человечества хороши, но недостаточно, а львиная доля — так вовсе откровенная чушь. И смех людей иногда раздражает, и миллионы людей даже не задумываются о будущем. И, может быть, нам не стоит быть предельно осторожными с теми, кто не задумывается о жизни и смерти? Потому что будет серьезный облом, если люди узнают, как решить задачу смерти, после того как я умру.

Независимо от того, как считаете вы, нам всем стоит задуматься об этом. В «Игре престолов» люди ведут себя так: «Мы так заняты битвой друг с другом, но на самом деле нам всем нужно сосредоточиться на том, что идет с севера от стены». Мы пытаемся устоять на бревне баланса, но на самом деле все наши проблемы могут решиться в мгновение ока, когда мы спрыгнем с него.

И когда это произойдет, ничто больше не будет иметь никакого значения. В зависимости от того, по какую сторону мы упадем, проблемы будут решены, потому что их либо не будет, либо у мертвых людей не может быть проблем.

Вот почему есть мнение, что сверхразумный искусственный интеллект может стать последним нашим изобретением — последней задачей, с которой мы столкнемся. А как думаете вы?

По материалам waitbutwhy.com, компиляция Тима Урбана.
В статье использованы материалы работ Ника Бострома, Джеймса Баррата, Рэя Курцвейла, Джея Нильс-Нильссона, Стивена Пинкера, Вернора Винджа, Моше Варди, Расса Робертса, Стюарта Армстрога и Кая Сотала, Сюзан Шнайдер, Стюарта Рассела и Питера Норвига, Теодора Модиса, Гари Маркуса, Карла Шульмана, Джона Серля, Джарона Ланье, Билла Джоя, Кевина Кели, Пола Аллена, Стивена Хокинга, Курта Андерсена, Митча Капора, Бена Герцел, Артура Кларка, Хьюберта Дрейфуса, Теда Гринвальда, Джереми Говарда.
Искусственный интеллект. Часть первая: путь к сверхинтеллекту
Искусственный интеллект. Часть вторая: вымирание или бессмертие?

МЕТКИ: ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, КАТАСТРОФЫ, РОБОТОТЕХНИКА, ХОЛОДИЛЬНИК.

Журнал «*N*оосфера.Общество.Человек»
journal «*N*oosphere. Society. Man»

http://noocivil.esrae.ru/

http://www.scireg.org/rus/files/fileinfo/458

Адрес сайта:
http://noosfan.jimdo.com/
http://noosfan.jimdo.com/о-нас/

ВИПЕРСОН

http://viperson.ru/people/onoprienko-vladimir-ivanovich