Ник Бостром: Живем ли мы в компьютерной симуляции (2001)

“Are you living in a computer simulation?”

by Nick Bostrom [Published in Philosophical Quarterly (2003) Vol. 53, No. 211, pp. 243-255. (First version: 2001)]

В данной статье утверждается, что по крайней мере одно из трёх следующих предположений является истинным:

(1) весьма вероятно, что человечество вымрет до того, как достигнет «постчеловеческой» фазы;
(2) каждая постчеловеческая цивилизация с крайне малой вероятностью будет запускать значительное число симуляций своей эволюционной истории (или ее вариантов) и
(3) мы почти определенно живём в компьютерной симуляции.

Из этого следует, что вероятность нахождения в фазе постчеловеческой цивилизации, которая сможет запускать симуляции своих предшественниц, равна нулю, если только не принять как истину тот случай, что мы уже живём в симуляции. Обсуждаются также другие следствия этого результата.

1.Введение

Многие произведения научной фантастики, а также прогнозы серьёзных футурологов и исследователей технологий предсказывают, что в будущем будут доступны колоссальные объёмы вычислительных мощностей. Предположим, что эти предсказания верны. Например, последующие поколения со своими сверхмощными компьютерами смогут запускать детальные симуляции своих предшественников или людей, подобных своим предшественникам. Поскольку их компьютеры будут настолько сильны, они смогут запускать много подобных симуляций. Предположим, что эти симулированные люди обладают сознанием (а они будут обладать им, если симуляция будет высокоточной и если определённая широко принятая в философии концепция сознания является верной). Из этого следует, что наибольшее число умов, подобных нашему, не принадлежат оригинальной расе, а, скорее, принадлежат людям, симулированным продвинутыми потомками оригинальной расы. Основываясь на этом,

можно утверждать, что разумно ожидать, что мы находимся среди симулированных, а не среди исходных, натуральных биологических умов. Таким образом, если мы не считаем, что сейчас живём в компьютерной симуляции, то мы не должны считать, что наши потомки будут запускать много симуляций своих предков. В этом и есть основная идея. В оставшейся части работы мы рассмотрим ее более подробно.

Помимо интереса, который может представлять данный тезис для тех, кто вовлечён в футуристические дискуссии, здесь есть и сугубо теоретический интерес. Данное доказательство является стимулом для формулирования некоторой методологической и метафизической проблематики, и так же предлагает некоторые естественные аналогии традиционным религиозным концепциям, и эти аналогии могут показаться удивительным или же наводящими на размышления.

Структура этой статьи такова: в начале мы сформулируем некое предположение, которое нам надо импортировать из философии сознания для того, чтобы это доказательство заработало. Затем мы рассмотрим некоторые эмпирические причины для того чтобы полагать, что запуск огромного множества симуляций человеческих умов будет доступен будущей цивилизации, которая разовьёт многие из тех технологий, относительно которых было прояснено, что они не противоречат известным физическим законам и инженерным ограничениям.

Эта часть не является необходимой с философской точки зрения, но однако побуждает обратить внимание на основную мысль статьи. Затем последует изложение доказательства по сути, с использованием некоторых простых приложений теории вероятности, и раздел, обосновывающий слабый принцип равнозначности, который данное доказательство использует. В итоге мы обсудим некоторые интерпретации альтернативы, о которой говорится в начале, и это и будет заключением доказательства о проблеме симуляции.

2. Предположение о независимости от носителя

Распространённым предположением в философии сознания является предположение о независимости от носителя. Идея состоит в том, что ментальные состояния могут возникать в любом носителе из широкого класса физических носителей. При том условии, что в системе воплощается правильный набор вычислительных структур и процессов,

в ней могут возникать осознанные переживания. Сущностным свойством не является воплощение в основанных на углероде биологических нервных сетях внутричерепных процессов: основанные на силиконе процессоры внутри компьютеров могут проделывать абсолютно тот же трюк. Аргументы в пользу этого тезиса выдвигались в существующей литературе, и, хотя он не является полностью непротиворечивым, мы будем принимать его здесь как данность.

Доказательство, которое мы здесь предлагаем, однако, не зависит от какой-нибудь очень сильной версии функционализма или компьютализма. Например, мы не должны принимать то, что тезис о независимости от носителя является с необходимостью истинным (как в аналитическом, так и в метафизическом смысле) – а должны принимать только то, что, в действительности, компьютер под управлением соответствующей программы мог бы обладать сознанием. Более того, мы не должны предполагать, что для того, чтобы создать сознание в компьютере, нам пришлось бы запрограммировать его таким образом, чтобы он вёл себя во всех случаях как человек, проходил бы тест Тюринга и т. д. Нам нужно только более слабое допущение о том, что для создания субъективных переживаний достаточно того, чтобы вычислительные процессы в человеческом мозгу были бы структурно скопированы в соответствующих высокоточных деталях, например, на уровне индивидуальных синапсов. Эта уточненная версия независимости от носителя является весьма широко принятой.

Нейротрансмиттеры, факторы роста нервов и другие химические вещества, которые меньше синапсов, очевидным образом играют роль в человеческом познании и обучении. Тезис о независимости от носителя состоит не в том, что эффект от этих химических веществ является малым или пренебрежимым, но в том, что они влияют на субъективный опыт только через прямое или непрямое воздействие на вычислительную активность. Например, если не существует субъективных различий без того, чтобы имела места также разница и в синаптических разрядах, то тогда требуемая детализация симуляции находится на синаптическом уровне (или выше).

3.Технологические пределы вычислений

На нынешнем уровне технологического развития у нас нет ни достаточно эффективного мощного оборудования, ни соответствующего программного обеспечения, чтобы создавать сознательные умы на компьютере. Однако были выдвинуты серьёзные

аргументы в пользу того, что если технологический прогресс будет продолжаться без остановок, то тогда эти ограничения будут в конечном счёте преодолены. Некоторые авторы утверждают, что эта фаза наступит всего через несколько десятилетий. Однако для целей нашей дискуссии не требуется никаких предположений о временной шкале. Доказательство симуляции работает столь же хорошо и для тех, кто полагает, что потребуются сотни тысяч лет, чтобы достичь «постчеловеческой» фазы развития, когда человечество обретёт большую часть тех технологических способностей, которые, как сейчас можно показать, согласуются с физическими законами и с материальными и энергетическими ограничениями.

Эта зрелая фаза технологического развития сделает возможным превращать планеты и другие астрономические ресурсы в компьютеры колоссальной силы. В настоящий момент трудно быть уверенным относительно каких-либо предельных границ компьютерной мощности, которая будет доступна постчеловеческим цивилизациям. Поскольку у нас до сих пор нет «теории всего», мы не можем исключить возможности того, что новые физические феномены, запрещённые современными физическими теориями, могут быть использованы для преодоления ограничений, которые, согласно нашему нынешнему представлению, накладывают теоретические пределы на обработку информации внутри данного куска материи. С гораздо большей степенью надёжности мы можем установить нижние границы постчеловеческих вычислений, предполагаю реализацию только тех механизмов, которые уже понятны. Например, Эрик Дрекслер дал набросок устройства системы, размером с кубик сахара (за исключением системы охлаждения и питания), которая могла бы выполнять 1021 операций в секунду. Другой автор дал грубую оценку в 1042 операций в секунду для компьютера размером с планету. (Если мы научимся создавать квантовые компьютеры, или научимся строить компьютеры из ядерной материи или плазмы, мы сможем приблизиться ещё ближе к теоретическим пределам. Сет Ллойд вычислил верхний предел для компьютера весом в 1 кг в 5*1050 логических операций в секунду, выполняемых над 1031 бит. Однако, для наших целей достаточно использовать более консервативные оценки, которые подразумевают только известные сейчас принципы работы.)

Количество компьютерной мощности, необходимое для того, чтобы эмулировать человеческий мозг, поддаётся точно такой же грубой оценке. Одна оценка, основанная на том, насколько затратно в вычислительном смысле было бы скопировать функционирование кусочка нервной ткани, который мы уже поняли и чья функциональность была уже скопирована в кремнии (а именно, была

скопирована система усиления контрастности в сетчатке глаза), даёт оценку в примерно 1014 операций в секунду. Альтернативная оценка, основанная на числе синапсов в мозге и частоте их срабатывания, даёт величину в 1016-1017 операций в секунду. Соответственно, и даже ещё больше вычислительных мощностей может потребоваться, если бы нам захотелось симулировать в деталях внутреннюю работу синапсов и ветвей дендритов. Однако, весьма вероятно, что центральная нервная система человека имеет определённую меру избыточности на микроуровне, чтобы компенсировать ненадёжность и шум своих нейронных компонентов. Следовательно, можно было бы ожидать значительного прироста эффективности при использовании более надёжных и гибких небиологических процессоров.

Память является не более значительным ограничением, чем процессорная мощность. Более того, поскольку максимальный поток сенсорных данных человека имеет порядок в 108 бит в секунду, то симулирование всех сенсорных событий потребовало бы пренебрежимо малой стоимости в сравнении с симулированием кортикальной активности. Таким образом, мы можем использовать процессорную мощность, необходимую для симулирования центральной нервной системы, как оценку общей вычислительной стоимости симуляции человеческого ума.

Если окружающая среда входит в симуляцию, это потребует дополнительной компьютерной мощности – количество которой зависит от размеров и подробности симуляции. Симуляция всей вселенной с точностью до квантового уровня, очевидно, невозможна, за исключением того случая, если будет открыта некая новая физика. Но для того, чтобы получить реалистическую симуляцию человеческого опыта, требуется гораздо меньше – только столько, сколько нужно, чтобы убедиться, что симулированные люди, взаимодействующие обычным человеческим образом с симулированной окружающей средой, не заметят никаких различий. Микроскопическая структура внутренних областей Земли может быть легко опущена. Удалённые астрономические объекты могут подвергнуты очень высокому уровню сжатия: точное сходство должно быть только в узком диапазоне свойств, которые мы можем наблюдать с нашей планеты или с космического аппарата внутри Солнечной системы. На поверхности Земли макроскопические объекты в необитаемых местах должны быть непрерывно симулированы, но микроскопические феномены могут заполняться ad hoc, то есть по мере необходимости. То, что вы видите через электронный микроскоп должно выглядеть неподозрительным, но у вас обычно нет никаких способов проверить его согласованность с ненаблюдаемыми частями

микромира. Исключения возникают, когда мы нарочно проектируем системы, чтобы запрячь ненаблюдаемые микроскопические феномены, которые действуют в соответствии с известными принципами, чтобы получить результаты, которые мы можем независимым образом проверить. Классическим примером этого является компьютер. Симуляция, таким образом, должна включать в себя непрерывные имитации компьютеров вплоть до уровня индивидуальных логических элементов. Это не представляет проблем, так как наша нынешняя вычислительная мощность является пренебрежимо малой по постчеловеческим стандартам.

Более того, постчеловеческий творец симуляции будет иметь достаточно вычислительной мощности, чтобы отслеживать в деталях состояние мыслей во всех человеческих мозгах всё время. Таким образом, когда он обнаружит, что какой-то человек готов сделать некое наблюдение о микромире, он может заполнить симуляцию с достаточным уровнем детализации настолько, насколько это нужно. Если какая-то ошибка случится, режиссер симуляции может легко отредактировать состояния любого мозга, который узнал об аномалии до того, как он разрушит симуляцию. Либо режиссер может отмотать симуляцию на несколько секунд назад и перезапустить ее таким образом, чтобы избежать проблемы.

Из этого следует, что наиболее дорогостоящим при создании симуляции, которая неотличима от физической реальности для находящихся в ней человеческих сознаний, будет создание симуляций органических мозгов вплоть до нейронного или субнейронного уровня. Хотя невозможно дать очень точную оценку цены реалистической симуляции человеческой истории, мы можем использовать оценку в 1033-1036 операций в качестве грубой оценки.

По мере того, как мы будем иметь больше опыта в создании виртуальной реальности, мы достигнем лучшего понимания вычислительных требований, которые необходимы для того, чтобы такие миры выглядели реалистичными для их посетителей. Но даже если наша оценка неверна на несколько порядков величины, это не имеет большого значения для нашего доказательства. Мы отметили, что грубая оценка вычислительной мощности компьютера массой с планету составляет 1042 операций в секунду, и это только принимая во внимание уже известные нанотехнологические конструкции, которые, вероятно, далеки от оптимальных. Один такой компьютер может симулировать всю ментальную историю человечества (назовём это симуляцией предков) используя только одну миллионную своих ресурсов за 1 секунду. Постчеловеческая цивилизация может в

конечном счёте построить астрономическое количество таких компьютеров. Мы можем заключить, что постчеловеческая цивилизация может запустить колоссальное количество симуляций предков, даже если она потратит на это только малую долю своих ресурсов. Мы можем придти к этому заключению, даже допуская значительную погрешность во всех наших оценках.

Постчеловеческие цивилизации будут иметь достаточно вычислительных ресурсов, чтобы запустить огромное множество симуляций-предков, даже используя очень малую долю своих ресурсов для этих целей.

4. Ядро доказательства симуляции

Основная идея этой статьи может быть выражена в следующем: если есть существенный шанс, что наша цивилизация когда-нибудь достигнет постчеловеческой стадии и запустит множество симуляций-предков, то как мы можем доказать, что мы не живём в одной такой симуляции?

Мы разовьём эту мысль в виде строгого доказательства. Давайте введём следующие обозначения:

f_p – доля от всех цивилизаций человеческого уровня, которые доживают до постчеловеческой стадии;
N – среднее число симуляций предков, которые запускает постчеловеческая цивилизация;
H – среднее число людей, которые жили в цивилизации до того, как она достигла постчеловеческой стадии.

Тогда реальная доля всех наблюдателей с человеческим опытом, которые живут в симуляции:

$$f_{sim} = \frac{f_p HN}{f_p HN + H}$$

Обозначим как долю постчеловеческих цивилизаций, которые заинтересованы в запуске симуляций-предков (или которые содержат по крайней мере некоторое количество отдельных существ, которые заинтересованы в этом и имеют значительные ресурсы, чтобы

запускать значительное число симуляций) и как среднее число симуляций-предков, запускаемых такими заинтересованными цивилизациями, мы получаем:

$$N = N_I f_I$$

И следовательно:

$$f_{sim} = \frac{f_p f_i N_I}{f_p f_I N_I + 1} \qquad (*)$$

По причине колоссальной вычислительной силы постчеловеческих цивилизаций является крайне большой величиной, как мы видели в предыдущем разделе. Рассматривая формулу (*) мы можем видеть, что по крайней мере одно из трёх следующих предположений является истинным:

(1) $f_p \approx 0$
(2) $f_I \approx 0$
(3) $f_{sim} \approx 1$

5. Мягкий принцип равнозначности

Мы можем сделать шаг дальше и заключить, что при условии истинности пункта (3) можно быть почти наверняка уверенным, что вы находитесь в симуляции. Говоря в общем, если мы знаем, что доля x всех наблюдателей с опытом человеческого типа живёт в симуляции, и мы не имеем никакой дополнительной информации, которая показывает, что наш собственный частный опыт является с большей или меньшей степенью вероятности воплощённым in machine, а не in vivo, чем другие виды человеческого опыта, и тогда наша уверенность, что мы находимся в симуляции, должна быть равна x:

$$\mathrm{Cr}(\mathrm{SIM}|f_{sim}=x)=x$$

Этот шаг оправдан очень слабым принципом равнозначности. Давайте разделим два случая. В первом случае, который является более простым, все исследуемые умы подобны вашему, в том смысле, что они в точности качественно соответствуют вашему уму: у них есть та же самая информация и те же самые переживания, что у вас. Во втором случае умы только подобны друг другу лишь в широком смысле, будучи тем сортом умов, которые типичны для человеческих существ, но качественно отличаются друг от друга и каждый имеет различный набор опыта. Я утверждаю, что даже в том случае, когда умы качественно различны, доказательство симуляции по-прежнему работает, при условии, что у вас нет никакой информации, которая отвечает на вопрос о том, какие из различных умов симулированы и какие реализованы биологически.

Детальное обоснование более строго принципа, которое включает оба наших частных примера как тривиальные частные случаи, была дана в литературе. Недостаток места не даёт возможности привести здесь обоснование целиком, но мы можем привести здесь одно из интуитивных обоснований. Представим, что x% популяции имеют определённую генетическую последовательность S внутри определённой части своего ДНК, которая обычно называется «мусорной ДНК». Предположим, далее, что нет никаких проявлений S (за исключением тех, которые могут проявиться при генетическом тестировании) и нет никаких корреляций между обладанием S и какими-либо внешними проявлениями. Тогда вполне очевидно, что до того, как ваша ДНК будет секвенсирована, является рациональным приписать уверенность в x% гипотезе, что у вас есть фрагмент S. И это является вполне независимым от того факта, что люди, у которых есть S, имеют умы и переживания качественно отличающиеся от тех, что имеют люди, у которых нет S. (Они различны просто потому что у всех людей есть различный опыт, а не потому что есть какая-то прямая связь между S и тем видом опыта, который переживает человек.)

Те же самые рассуждения применимы, если S не является свойством обладания определённой генетической последовательностью, а вместо этого фактом нахождения в симуляции, в предположении, что у нас нет информации, которая позволяет нам предсказать какие-либо различия между переживаниями симулированных умов и между переживаниями исходных биологических умов.

Следует подчеркнуть, что мягкий принцип равнозначности подчёркивает только равнозначность между гипотезами, каким именно из наблюдателей вы являетесь, когда вы не имеете информации о том, которым из наблюдателей вы являетесь. Он в общем случае не приписывает равнозначность между гипотезами, когда у вас нет

конкретной информации о том, какая из гипотез является истинной. В отличие от Лапласова и других более сильных принципов равнозначности, он, таким образом, не подвержен парадоксу Бертрана и другим подобным затруднениям, которые осложняют неограниченное применение принципов равнозначности.

Читатели, знакомые с доказательством Конца света (Doomsday argument, DA)(J. Leslie, “Is the End of the World Nigh? ” Philosophical Quarterly 40, 158: 65-72 (1990)), могут испытывать беспокойство, что принцип равнозначности, применяемый здесь, опирается на те же предположения, что ответственны за выбивание почвы из-под DA, и что контринуитивность некоторых выводов последнего бросает тень на достоверность рассуждения о симуляции. Это не так. DA опирается на гораздо более строгую и противоречивую предпосылку о том, что человек должен рассуждать так, как если бы он был случайной выборкой из всего множества людей, которые когда-либо жили и будут жить (в прошлом, настоящем и будущем), несмотря на то, что мы знаем, что мы живём в начале XXI века, а не в какой-то точке в далёком будущем. Мягкий принцип неопределённости обращается только к тем случаям, когда у нас нет дополнительной информации о том, к какой группе людей мы принадлежим.

Если делание ставок является неким основанием для рациональной веры, то тогда, если все сделают ставку на то, находятся ли они в симуляции или нет, то, если люди будут использовать мягкий принцип неопределённости и будут делать ставку на то, что они в симуляции, опираясь на знание о том, что большая часть людей в ней находится, то тогда почти все выиграют свои ставки. Если они будут ставить на то, что они не в симуляции, то почти все проиграют. Кажется более полезным следовать принципу мягкой равнозначности. Далее, можно представить себе последовательность возможных ситуаций, в которых всё большая часть людей живёт в симуляциях: 98%, 99%, 99.9%, 99.9999%, и так далее. По мере приближения к верхнему пределу, когда все живут в симуляции (откуда можно дедуктивно вывести, что каждый находится в симуляции), кажется разумным требование, чтобы достоверность, которую некто приписывает тому, что он находится в симуляции, плавно непрерывно приближалось к лимитирующему пределу полной уверенности.

6. Интерпретация

Возможность, указанная в пункте (1), вполне понятна. Если (1) верно, то человечество почти наверняка не сможет достичь постчеловеческого уровня; ни один вид на нашем уровне развития не становится

постчеловеческим, и трудно обнаружить какие-либо оправдания для мысли, что наш собственный вид обладает какими-либо преимуществами или особой защитой от будущих катастроф. При условии (1), таким образом, мы должны приписать высокую достоверность Гибели (DOOM), то есть гипотезе о том, что человечество исчезнет до того, как достигнет постчеловеческого уровня:

$$CR(Doom|f_p \approx 1) \approx 1$$

Можно представить гипотетическую ситуацию, в которой мы имеем такие данные, которые перекрывают наши знания о fp. Например, если мы обнаружим, что в нас вот-вот врежется гигантский астероид, мы можем предположить, что мы оказались исключительно невезучими. Мы можем в этом случае приписать гипотезе о Гибели большую достоверность, чем наше ожидание доли цивилизаций человеческого уровня, которые не смогут достичь постчеловечества. В нашем случае, однако, у нас, судя по всему, нет никаких оснований думать, что мы являемся особенными в этом отношении, в лучшую или худшую стороны.

Предположение (1) не означает само по себе, что мы, скорее всего, вымрем. Оно говорит о том, что мы вряд ли достигнем постчеловеческой фазы. Эта возможность может означать, например, что мы останемся на нынешнем или немного превосходящем его уровне в течение длительного времени до того, как вымрем. Другая возможная причина истинности (1) – это то, что, скорее всего, технологическая цивилизация рухнет. При этом примитивные человеческие общества сохранятся на Земле.

Есть много способов, которыми человечество может вымереть до того, как достигнет постчеловеческой фазы развития. Наиболее естественным объяснением (1) является то, что мы вымрем в результате развития некой мощной, но опасной технологии. Одним из кандидатов является молекулярная нанотехнология, зрелая стадия которой позволит создавать способных к саморепликации нанороботов, могущих питаться грязью и органической материей – нечто вроде механической бактерии. Такие нанороботы, будучи спроектированными со злокозненными целями, могут привести к гибели всей жизни на планете.

Вторая альтернатива вывода рассуждения о симуляции состоит в том, что доля постчеловеческих цивилизаций, которые заинтересованы в

запуске симуляций-предков, является пренебрежимо малой. Для того, чтобы (2) было истинным, должна быть строгая конвергенция между путями развития продвинутых цивилизаций. Если количество симуляций предков, создаваемых заинтересованными цивилизациями, является исключительно большим, то тогда редкость таких цивилизаций должна быть соответственно экстремальной. Практически ни одна из постчеловеческих цивилизации не решает использовать свои ресурсы для создания большого количества симуляций-предков. Более того, почти во всех постчеловеческих цивилизациях отсутствуют индивиды, у которых есть соответствующие ресурсы и интерес, чтобы запускать симуляции-предков; или же у них есть подкреплённые силой законы, предотвращающие поведение индивидов согласно их желаниям.

Какая сила может привести к такой конвергенции? Кто-то может утверждать, что продвинутые цивилизации все как одна развиваются по траектории, которая приводит к признанию этического запрета запуска симуляций-предков по причине страданий, которые испытывают обитатели симуляции. Однако с нашей нынешней точки зрения не кажется очевидным, что создание человеческой расы является аморальным. Наоборот, мы склонны воспринимать существование нашей расы как имеющее большую этическую ценность. Более того, конвергенции только этических взглядов на аморальность запуска симуляций предков – недостаточно: она должна объединяться с конвергенцией цивилизационной социальной структуры, которая приводит к тому, что виды деятельности, считающиеся аморальными, эффективным образом запрещаются.

Другая возможность конвергенции состоит в том, что почти все индивидуальные постлюди в почти всех постчеловеческих цивилизациях развиваются в направлении, в котором они теряют стремление к запуску симуляций-предков. Это потребует значительных изменений в мотивациях, движущих их постчеловеческими предками, поскольку наверняка есть много людей, которые хотели бы запускать симуляции предков, будь у них такая возможность. Но, возможно, многие из наших человеческих желаний будут казаться глупыми любому, кто станет постчеловеком. Может быть, научное значение симуляций-предков для постчеловеческих цивилизаций является пренебрежимо малым (что не выглядит слишком невероятным с учётом их невероятного интеллектуального превосходства) и, может быть, постлюди считают рекреационную активность за очень неэффективный способ получения удовольствий – которое может быть получено гораздо более дешево за счёт прямой стимуляции центров удовольствия мозга. Один вывод, который следует из (2) – это то, что постчеловеческие общества будут крайне отличаться от человеческих

обществ: в них не будет относительно богатых независимых агентов, у которых есть полный диапазон желаний, подобных человеческим, и которые свободны действовать в соответствии с ними.

Возможность, описываемая выводом (3), является наиболее интригующей с концептуальной точки зрения. Если мы живём в симуляции, то наблюдаемый нами космос является только маленьким кусочком в тотальности физического существования. Физика вселенной, где находится компьютер, может напоминать, а может и не напоминать физику наблюдаемого нами мира. В то время, как наблюдаемый нами мир является в некоторой степени «реальным», он не расположен на некотором фундаментальном уровне реальности. Для симулированных цивилизаций может быть возможно стать постчеловеческими. Они могут запускать в свою очередь симуляции-предков на мощных компьютерах, которые они построили в симулированной вселенной. Такие компьютеры будут «виртуальными машинами», — весьма распространённая концепция в компьютерной науке. (Веб-приложения, написанные на Java script, например, работает на виртуальной машине – симулированном компьютере – на вашем ноутбуке.)

Виртуальные машины могут вкладываться одна в другую: возможно симулировать виртуальную машину, симулирующую другую машину, и так далее, с произвольно большим числом шагов. Если мы сможем создать наши собственные симуляции предков, это будет сильным свидетельством против пунктов (1) и (2), и мы в силу этого должны будем заключить, что мы живём в симуляции. Более того, мы должны будем подозревать, что постлюди, которые запустили нашу симуляцию, сами по себе тоже являются симулированными существами, и их создатели, в свою очередь, тоже могут быть симулированными существами.

Реальность, таким образом, может содержать несколько уровней. Если даже иерархия должна закончится на каком-то уровне – метафизический статус этого заявления весьма неясен – может быть достаточно пространства для большого количества уровней реальности, и это количество может увеличиваться с течением времени. (Одно из соображений, которое говорит против такой многоуровневой гипотезы, состоит в том, что вычислительная цена для симуляторов базового уровня будет очень большой. Симулирование даже одной постчеловеческой цивилизации может быть запрещающее дорого. Если так, то мы должны ожидать, что наша симуляция будет выключена, когда мы приблизимся к постчеловеческому уровню.)

Хотя все элементы данной системы являются натуралистическими,

даже физическими, возможно нарисовать некоторые свободные аналогии с религиозными концепциями мира. В некотором смысле постлюди, которые запустили симуляцию, подобны богам по отношению к людям в симуляции: постлюди создают тот мир, который мы видим; они обладают превосходящим нас интеллектом; они всемогущи в том смысле, что они могут вмешиваться в работу нашего мира способами, нарушающими физические законы, и они являются всезнающими в том смысле, что они могут мониторить всё, что происходит. Однако все полубоги, за исключением тех, которые живут на фундаментальном уровне реальности, подвержены действиям более сильных богов, обитающих на более высоких уровнях реальности.

Дальнейшее пережёвывание этих тем может закончится натуралистической теогонией, которая будет изучать структуру этой иерархии и ограничения, накладываемые на обитателей той возможностью, что их действия на их уровне могут повлиять на отношение к ним обитателей более глубокого уровня реальности. Например, если никто не может быть уверен, что он находится на базовом уровне, то каждый должен рассматривать вероятность того, что его действия будут вознаграждаться или наказываться, возможно, на основании неких моральных критериев, хозяевами симуляции. Жизнь после смерти будет реальной возможностью. Из-за этой фундаментальной неопределённости даже цивилизация на базовом уровне будет иметь побуждение вести себя этично. Тот факт, что они имеют причину вести себя морально, будет разумеется веским доводом для кого-то другого вести себя морально, и так далее, образуя добродетельный круг. Таким образом можно получить нечто вроде универсального этического императива, соблюдать который будет в личных интересах каждого, и который происходит из «ниоткуда».

В дополнение к симуляциям предков, можно представить возможность и более избирательных симуляций, которые включают в себя только небольшую группу людей или одного индивида. Остальные люди будут тогда «зомби» или «люди-тени» – люди, симулированные только на уровне, достаточном, чтобы полностью симулированные люди не замечали ничего подозрительного.

Не ясно, насколько дешевле будет симулировать людей-теней, чем реальных людей. Не является даже очевидным, что для некого объекта возможно вести себя неотличимо от реального человека и при это не иметь осознанных переживаний. Даже если такие селективные симуляции существуют, вы не должны быть уверены, что вы в ней, до того, как вы будете уверены, что такие симуляции гораздо более многочисленны, чем полные симуляции. Мир должен иметь примерно в 100 миллиардов больше я-симлуяций (симуляций жизни только

одного сознания), чем имеется полных симуляций предков – для того, чтобы большинство симулируемых людей были бы в я-симуляциях.

Также есть возможность, что симуляторы перепрыгивают определённую часть ментальной жизни симулируемых существ и дают им фальшивые воспоминания о том типе опыта, который они могли бы иметь во время пропущенных периодов. Если так, можно представить себе следующее (натянутое) решение проблемы зла: что в действительности в мире нет страданий и что все воспоминания о страданиях являются иллюзией. Разумеется, эту гипотезу можно рассматривать всерьёз только в те моменты, когда сам не страдаешь.

Предполагая, что мы живём в симуляции, каковы последствия этого для нас, людей? Вопреки тому, что было сказано до того, последствия для людей не особенно радикальны. Нашим наилучшим гидом в том, как наши постчеловеческие создатели выбрали устроить наш мир, является стандартное эмпирическое исследование вселенной, которой мы видим. Изменения большей части нашей системы верований будут скорее небольшими и мягкими – пропорциональными нашему отсутствию уверенности в нашей способности понять систему мышления постлюдей.

Правильное понимание истинности тезиса (3) не должно делать нас «безумными» или заставлять нас бросить свой бизнес и перестать делать планы и предсказания на завтра. Главная эмпирическая важность (3) в настоящий момент, судя по всему, лежит в ее роли тройственном выводе, приведённом выше.

Нам следует надеется, что (3) является истинным, поскольку это уменьшает вероятность (1), однако если вычислительные ограничения делают вероятным то, что симуляторы выключат симуляцию до того, как она достигнет постчеловеческого уровня, то тогда нашей наилучшей надеждой является истинность (2).

Если мы узнаем больше о постчеловеческой мотивации и ограничениях ресурсов, может быть, в результате нашего развития в сторону постчеловечества, то тогда гипотеза о том, что мы симулированы, получит гораздо более богатый набор эмпирических приложений.

7. Заключение

Технологически зрелая постчеловеческая цивилизация располагала бы огромной вычислительной мощностью. Основываясь на этом, рассуждение о симуляции показывает, что, по крайней мере один из

следующих тезисов истинен:

(1) доля цивилизаций человеческого уровня, которые достигают постчеловеческого уровня, очень близка к нулю.
(2) Доля постчеловеческих цивилизаций, которые заинтернесованы в запуске симуляций предшественников, очень близка к нулю.
(3) Доля всех людей с нашим типом переживаний, которые живут в симуляции, близка к единице.

Если (1) верно, то мы почти наверняка умрём до того, как достигнем постчеловеческого уровня.

Если (2) верно, то тогда должна быть строго согласованная конвергенция путей развития всех продвинутых цивилизаций, так чтобы ни в одной из них не было относительно богатых индивидов, которые хотели бы запускать симуляции предков и были бы свободны делать это.

Если (3) верно, то мы почти наверняка живём в симуляции. Тёмный лес нашего неведения делает разумным распределить нашу уверенность почти равномерно между пунктами (1), (2) и (3).

За исключением того случая, что мы уже живём в симуляции, наши потомки почти наверняка никогда не будут запускать симуляции-предков.

Благодарности

Я благодарен многим людям за их комментарии, и особенно Amara Angelica, Robert Bradbury, Milan Cirkovic, Robin Hanson, Hal Finney, Robert A. Freitas Jr., John Leslie, Mitch Porter, Keith DeRose, Mike Treder, Mark Walker, Eliezer Yudkowsky, и анонимным референтам.

Перевод: Алексей Турчин

Примечания переводчика:
1) Выводы (1) и (2) – нелокальны. Они говорят, что либо все цивилизации погибают, либо все не хотят создавать симуляции. Это утверждение распространяется не только на всю видимую вселенную, не только на всю бесконечность вселенной за пределами горизонта видимости, но и на всё множество 10**500 степени вселенных с разными свойствами, которые возможны, согласно теории струн. В отличие от них, тезис о том, что мы живём в симуляции, – локален.

Всеобщие утверждения гораздо реже бывают истинными, чем частные утверждения. (Сравни: «Все люди блондины» и «Иванов блондин» или «все планеты имеют атмосферу» и «Венера имеет атмосферу».) Чтобы опровергнуть общее утверждение, достаточно одного исключения. Таким образом, утверждение о том, что мы живём в симуляции, гораздо вероятнее первых двух альтернатив.

2) Не обязательно развитие компьютеров – достаточно, например, снов. Которые будут видеть генетически модифицированные и специально заточенные под это мозги.

3) Рассуждение о симуляции работает в обычной жизни. Большая часть изображений, которые поступают в наши мозги, являются симуляциями – это кино, телевизор, интернет, фотографии, реклама – и last but not least – сны.

4) Чем необычнее видимый нами объект, тем больше шансов, что он находится в симуляции. Например, если я вижу страшную аварию, то скорее всего я вижу ее во сне, по телевизору или в кино.

5) Симуляции могут быть двух типов: симуляции всей цивилизации и симуляции личной истории или даже какого-то одного эпизода из жизни одного человека.

6) Важно отличать симуляцию от имитации – возможна симуляция того человека или цивилизации, которых никогда не существовало в природе.

7) Сверхцивилизации должны быть заинтересованы в создании симуляций, чтобы изучить разные варианты своего прошлого и таким образом разные альтернативы своего развития. А также, чтобы, например, изучить среднюю частоту других сверхцивилизаций в космосе и их ожидаемые свойства.

8) Проблема симуляции сталкивается с проблемой философского зомби (то есть существами, лишёнными квалиа, как тени на экране телевизора). Симулируемые существа не должны быть философскими зомби. Если в большинстве симуляций находятся философские зомби, то рассуждение не работает (так как я не философский зомби.)

9) Если есть несколько уровней симуляции, то одна и та же симуляция2 уровня может использоваться в нескольких разных симуляциях 1 уровня теми, кто живёт в симуляции 0 уровня. С целью экономии вычислительных ресурсов. Это подобно тому, как много разных людей смотрят один и тот же кинофильм. То есть допустим я создал три

симуляции. И каждая из них создала 1000 подсимуляций. Тогда мне бы пришлось стимулировать 3003 симуляции на своём суперкомпьютере. Но если симуляции создали в принципе одинаковые подсимуляции, то мне достаточно смоделировать только 1000 симуляций, предъявляя результат работы каждой из них три раза. То есть всего я запущу 1003 симуляции. Иначе говоря, одна симуляция может иметь несколько хозяев.

10) То, живёте ли вы в симуляции или нет, можно определить по тому, насколько ваша жизнь отличается от среднестатистической в сторону уникальной, интересной или важной. Здесь предполагается, что делать симуляции интересных людей, живущих в интересное время важных перемен – более привлекательно для авторов симуляции, не зависимо от их целей – развлекательных или исследовательских.70 % людей, когда-либо живших на Земле, были неграмотными крестьянами. Однако здесь нужно учитывать эффект наблюдательной селекции: неграмотные крестьяне не могли задаться вопросом о том, в симуляции они или нет, а следовательно, тот факт, что вы не неграмотный крестьянин, ещё не доказывает, что вы в симуляции. Вероятно, наибольший интерес для авторов симуляции будет иметь эпоха в районе Сингулярности, так как в районе неё возможна необратимая бифуркация путей развития цивилизации, на которую могут повлиять малые факторы, в том числе особенности одной личности. Например, я, Алексей Турчин, полагаю, что моя жизнь настолько интересна, что скорее является симулированной, чем реальной.

11) То, что мы находимся в симуляции, увеличивает наши риски – а) симуляцию могут выключить б) авторы симуляции могут ставить над ней эксперименты, создавая заведомо маловероятные ситуации – падение астероида и т. д.

12) Важно отметить слова Бострома о том, что по крайней мере одно из трёх верно. То есть возможны ситуации, когда некоторые из пунктов верны одновременно. Например, то, что мы погибнем, не исключает того, что мы живём в симуляции, и то, что большинство цивилизаций не создаёт симуляции.

13) Симулированные люди и мир вокруг них могут быть вообще не похожи ни на каких реальных людей и реальный мир, важно, чтобы они думали, что они в реальном мире. Они не способны заметить отличия, потому что вообще никогда никакого реального мира не видели. Или их способность замечать отличия притуплена. Как это бывает во сне.

14) Есть соблазн обнаружить в нашем мире признаки симуляции,

проявляемые как чудеса. Но чудеса могут происходить и без симуляции.

15) Есть модель мироустройства, которая снимает предлагаемую дилемму. (но не лишена своих противоречий). А именно, это кастанедовска-будийская модель, где наблюдатель порождает весь мир.

16) Идея симуляции подразумевает упрощение. Если симуляция будет с точностью до атома, то она будет той же самой реальностью. В этом смысле можно представить себе ситуацию, когда некая цивилизация научилась создавать параллельные миры с заданными свойствами. В этих мирах она может ставить натурные эксперименты, создавая разные цивилизации. То есть это нечто вроде гипотезы космического зоопарка. Эти созданные миры не будут симуляциями, так как они будут вполне реальны, но они будут под властью тех, кто их создал и может включить и выключить. И их тоже будет количественно больше, так что здесь применимо похожее статистическое рассуждение, как и в рассуждении о симуляции.
Глава из статьи «НЛО как фактор глобального риска»:

НЛО – это глюки в Матрице

Согласно Н. Бострому (Ник Бостром. Доказательство Симуляции. www.proza.ru/2009/03/09/639), вероятность того, что мы живём в полностью симулированном мире, достаточно велика. То есть наш мир может быть полностью смоделирован на компьютере некой сверхцивилизацией. Это позволяет авторам симуляции создавать любые образы в ней, с непостижимыми для нас целями. Кроме того, если уровень контроля в симуляции мал, то в ней будут накапливаться ошибки, как при работе компьютера, и возникать сбои и глюки, которые можно заметить. Люди в чёрном превращаются в агентов Смитов, которые стирают следы глюков. Или некоторые жители симуляции могут получить доступ к неким незапротоколированным возможностям. Это объяснение позволяет объяснить любой возможный набор чудес, но оно не объясняет ничего конкретного – почему мы наблюдаем именно такие проявления, а не, скажем, летающих вверх тормашками розовых слонов. Основной риск состоит в том, что симуляция может быть использована для тестирования крайних условий работы системы, то есть в катастрофических режимах, а так же то, что симуляцию просто выключат, если она станет слишком сложной или завершит свою функцию.
Основной вопрос здесь – степени контроля в Матрице. Если речь идет о Матрице под очень жёстким контролем, то вероятность незапланированных глюков в ней невелика. Если же Матрица просто запущена, а потом предоставлена на произвол судьбы, то глюки в ней

будут накапливаться, как накапливаются глюки при работе операционной системы, по мере ее работы и по мере добавления новых программ.

Первый вариант реализуется, если авторы Матрицы заинтересованы во всех подробностях происходящих в Матрице событий. В этом случае они будут жёстко отслеживать все глюки и тщательно их стирать. Если же они заинтересованы только в конечном результате работы Матрицы или одном из ее аспектов, то их контроль будет менее жёстким. Например, когда человек запускает шахматную программу и уходит на весь день, его интересует только результат работы программы, но не подробности. При этом в ходе работы шахматной программы она может обсчитать множество виртуальных партий, иначе говоря, виртуальных миров. Иначе говоря, авторы здесь заинтересованы в статистическом результате работы очень многих симуляций, и подробности работы одной симуляции их волнуют только в той мере, в какой глюки не влияют на конечный результат. А в любой сложной информационной системе какое-то количество глюков накапливается, и по мере роста сложности системы сложность их удаления экспоненциально растёт. Поэтому проще мириться с присутствием неких глюков, чем удалять их на корню.

Далее, очевидно, что множество слабоконтролируемых систем гораздо больше множества жёстко контролируемых, поскольку слабоконтролируемые системы запускаются в больших количествах, когда их можно произвести ОЧЕНЬ дёшево. Например, множество виртуальных шахматных партий гораздо больше партий реальных гроссмейстеров, а множество домашних операционных систем гораздо больше множества правительственных суперкомпьютеров. Таким образом, глюки в Матрице допустимы, пока они не влияют на общий ход работы системы. Точно также и в реальности, если у меня шрифт в браузере стал отображаться другим цветом, то я не буду перезагружать весь компьютер или сносить операционную систему. Но то же самое мы видим в исследовании НЛО и других аномальных явлений! Существует некий порог, выше которого ни сами явления, ни их общественный резонанс прыгнуть не могут. Как только некие явления начинают подбираться к этому порогу, они или исчезают, или появляются люди в чёрном, или выясняется, что это была мистификация, или кто-то погибает.

Отметим, что есть два вида симуляций – полные симуляции всего мира и я-симуляции. В последних симулируется жизненный опыт только одного человека (или небольшой группы людей). В я-симуляции вероятнее обнаружить себя в интересной роли, тогда как в полной симуляции 70 процентов героев являются крестьянами. Из

соображений наблюдательной селекции я-симуляции должны быть гораздо более часты – хотя это соображение нуждается в дальнейшем осмыслении. Но в я-симуляциях тема НЛО должна быть уже заложена, как и вся предыстория мира. И она может быть заложена нарочно – чтобы исследовать, как я буду обращаться с этой темой.

Далее, в любой информационной системе рано или поздно заводятся вирусы – то есть паразитические информационные единицы, нацеленные на саморепликацию. Такие единицы могут возникать и в Матрице (и в коллективном бессознательном), и против них должна работать встроенная антивирусная программа. Однако по опыту пользования компьютерами и по опыту биологических систем мы знаем, что проще мириться с наличием безвредных вирусов, чем травить их до последнего. Тем более, что полное уничтожение вирусов часто требует сноса системы.

Таким образом, можно предположить, что НЛО – это вирусы, использующие глюки в Матрице. Это объясняет абсурдность их поведения, так как их интеллект ограничен, а также их паразитирование на людях – так как каждому человеку выделено в Матрице определённое количество вычислительных ресурсов, которые можно использовать. Можно предположить, что некоторые люди воспользовались глюками в Матрице, чтобы достичь своих целей, в том числе бессмертия, но тоже самое сделали существа из других вычислительных сред, например, симуляций принципиально иных миров, которые затем проникли в наш мир.
Ещё один вопрос – каков уровень глубины симуляции, в которой мы, скорее всего, находимся. Можно симулировать мир с точностью до атома, но это потребовало бы колоссальных вычислительных ресурсов. Другой крайний пример – шутер от первого лица. В нём трёхмерное изображение местности рисуется по мере надобности, когда главный герой подходит к новому месту, на основании общего плана местности и неких общих принципов. Либо используются заготовки для некоторых мест, а точная прорисовка других мест игнорируется (как в фильме «13 этаж»). Очевидно, чем точнее и подробнее симуляция, тем реже в ней будут глюки. С другой стороны, симуляции, сделанные «наспех», будет содержать гораздо больше глюков, но при этом потреблять неизмеримо меньше вычислительных ресурсов. Иначе говоря, с одинаковыми расходами можно было бы сделать или одну очень точную симуляцию, или миллион приблизительных. Далее, мы предполагаем, что в отношении симуляций действует тот же принцип, что и в отношении других вещей: а именно, что чем дешевле вещь, тем чаще она встречается (то есть в мире больше стекляшек, чем бриллиантов, больше метеоритов, чем астероидов и т. д.) Таким образом, мы, скорее, находимся внутри дешёвой упрощённой

симуляции, а не внутри сложной сверхточной симуляции. На это можно возразить, что в будущем будут доступны неограниченные вычислительные ресурсы, и поэтому любой актор будет делать достаточно подробные симуляции. Однако здесь вступает в действие эффект симуляций-матрёшек. А именно, продвинутая симуляция может создавать свои собственные симуляции, назовём их симуляциями второго уровня. Допустим, продвинутая симуляция мира середины 21 века (созданная, положим, в реальном 23 веке) может создать миллиарды симуляций мира начала 21 века. При этом она будет пользоваться компьютерами середины 21 века, которые будут более ограничены в вычислительных ресурсах, чем компьютеры 23 века. (А также реальный 23 век будет экономить на точности подсимуляций, так как они ему не важны.) Поэтому все миллиард симуляций начала 21 века, которые она создаст, будут весьма экономными по вычислительным ресурсам. В силу этого число примитивных симуляций, а также симуляций, более ранних в отношении симулируемого времени, будет в миллиард раз больше, чем число более подробных и поздних симуляций, и, следовательно, произвольный наблюдатель имеет в миллиард раз большие шансы обнаружить себя в более ранней (во всяком случае, до появления сверхкомпьютеров, способных к созданию собственных симуляций) и более дешёвой и более глючной симуляции. А согласно принципу self-sampling assumption каждый должен рассматривать самого себя как случайного представителя множества подобных себе существ, если он хочет получить наиболее точные вероятностные оценки. Таким образом, мы имеем гораздо большие шансы а) оказаться в самой дешёвой симуляции б) оказаться до того момента времени, когда появятся сверхкомпьютеры, способные к созданию симуляций (так и есть) в) оказаться в я-симуляции г) оказаться на дне цепочки из вложенных симуляций, то есть симуляции N уровня, где N максимально д) оказаться в симуляции со значительным уровнем глюков.

Другой вариант состоит в том, что НЛО нарочно запускаются в Матрицу, чтобы дурачить живущих в ней людей и смотреть, как они будут на это реагировать. Поскольку большинство симуляций, как я думаю, предназначены для симулирования мира в неких особенных, крайних условиях.

Всё же эта гипотеза не объясняет всего множества конкретных проявлений НЛО.
Риск здесь состоит в том, что если наша симуляция будет перегружена глюками, то хозяева симуляции могут решить ее перезагрузить.

Наконец, можно предположить «самозарождение Матрицы» — то есть

то, что мы живём в вычислительной среде, но эта среда самозародилась неким образом у истоков существования вселенной без посредства каких-то существ-создателей. Для того, чтобы эта гипотеза была более убедительной, следует вначале вспомнить, что согласно одному из описаний физической реальности сами элементарные частицы являются клеточными автоматами – чем-то вроде устойчивых комбинаций в игре «Жизнь». ru.wikipedia.org/wiki/Жизнь_(игра)

https://temofeev.ru/info/articles/nik-bostrom-zhivem-li-my-v-kompyuternoy-simulyatsii-2001/