А не живем ли мы в «Матрице»? Доказательство методом моделирования

Ник Бостром

Перевод: Шиенков Е.В.

Почти каждый зритель «Матрицы» хотя бы в течение секунды или пары секунд допускает неприятную возможность того, что он может на самом деле жить в Матрице. Философ из Йельского университета Ник Бостром тоже рассматривает эту возможность и приходит к выводу, что она гораздо более вероятна, чем вы могли бы себе вообразить.

Фильм «Матрица» знакомит нас со странным, приводящим в ужас сценарием. Человечество лежит в коматозном состоянии в каких-то коконах, а каждая деталь реальности определяется и контролируется враждебными ему компьютерами.

Для большинства зрителей этот сценарий интересен как образец научной фантастики, невероятно далекой от всего, что существует сегодня или, скорее всего, появится в будущем. Однако после тщательного обдумывания подобный сценарий перестает казаться немыслимым. Он очень даже вероятен.

В одной из своих статей Рей Курцвейль обсуждает наблюдаемую тенденцию к развитию вычислительных мощностей с постоянно возрастающей скоростью. По прогнозам Курцвейля практически неограниченные вычислительные мощности станут доступными в течение следующих пятидесяти лет. Давайте предположим, что Курцвейль прав и рано или поздно человечество создаст практически безграничные вычислительные мощности. Для целей этой дискуссии не важно, когда это произойдет. На эти разработки может уйти сто, тысяча или миллион лет.

Как отмечается в статье Курцвейля, безграничные вычислительные возможности расширят способности человечества до невероятной степени. Эта цивилизация станет «постчеловеческой» и будет способна на необычайные технологические свершения.

Постчеловеческая цивилизация может принять различные формы. Она может оказаться во многом похожей на нашу современную цивилизацию или радикально от нее отличаться. Разумеется, почти невозможно предсказать, как будет развиваться подобная цивилизация. Но одно мы знаем точно: постчеловеческая цивилизация будет располагать доступом к практически бесконечным вычислительным мощностям.

Постчеловеческой цивилизации может оказаться по силам даже превращать планеты и другие астрономические объекты в сверхмощные компьютеры. Иными словами в данный момент сложно с уверенностью определить «потолок» тех вычислительных мощностей, которые могут оказаться доступными постчеловеческим цивилизациям.

В статье Ника Бострома «А не живешь ли ты в компьютерной симуляции?» путем логического построения представлены доказательства, согласно которым, по крайней мере одно из следующих утверждений верно:

1. Весьма вероятно, что как биологический вид человечество исчезнет с лица земли, не достигнув «постчеловеческой» стадии;
2. Очень маловероятно, что любая постчеловеческая цивилизация запустит большое количество симуляций (моделей), имитирующих ее эволюционную историю (или варианты этой истории);
3. Мы почти наверняка живем в компьютерной симуляции.

Давайте рассмотрим эти три утверждения поочередно. Первое утверждение сформулировано прямо: если мы уничтожим самих себя в результате ядерной войны, биологической катастрофы или нанотехнологического катаклизма, то остальные пункты этого доказательства к делу не относятся. Однако давайте предположим, что это утверждение неверно, и, следовательно, мы сумеем избежать самоуничтожения и вступим в постчеловеческую эпоху.

Сущность человеческой цивилизации в условиях постчеловеческой эпохи невозможно представить во всей полноте. Точно так же нельзя вообразить разнообразные способы использования практически неограниченных вычислительных мощностей. Но давайте рассмотрим один из них — создание сложных симуляций человеческой цивилизации.

Представим себе историков будущего, моделирующих различные сценарии исторического развития. Это будут не сегодняшние упрощенные модели. С учетом огромных вычислительных возможностей, которыми будут располагать эти историки, в их распоряжении могут оказаться очень подробные симуляции, в которых будет различимо каждое здание, каждая географическая деталь, каждая личность. И каждый из этих индивидуумов будет наделен тем же уровнем вычислительных мощностей, сложности и интеллекта, что и живой человек. Как и агент Смит, они будут созданы на основе программного обеспечения, но при этом будут обладать психическими характеристиками человека. Конечно, они могут так никогда и не осознать, что являются программой. Чтобы создать точную модель, нужно будет сделать восприятие смоделированных личностей неотличимым от восприятия людей, живущих в реальном мире.

Подобно жителям Матрицы, эти люди будут существовать в искусственном мире, считая его реальным. В отличие от сценария с Матрицей эти люди будут полностью состоять из компьютерных программ.

Однако будут ли эти искусственные личности настоящими «людьми»? Будут ли они разумными независимо от уровня их вычислительных мощностей? Будут ли они наделены сознанием?

Реальность — это то, с чем никто на самом деле не знаком. Однако философы, изучающие сознание, обычно делают допущение о его «независимости от субстрата». По существу это означает, что сознание может зависеть от многих вещей: от знания, интеллекта (вычислительных мощностей), психической организации, отдельных деталей логической структуры и т. д. Но одним из условий, выполнение которого для существования сознания не обязательно, является наличие биологической ткани. Воплощение сознания в основанных на углероде биологических нейронных сетях — это не единственно возможный вариант. В принципе, того же самого результата можно добиться от основанных на кремнии процессоров, встроенных в компьютер.

Многим людям, знакомым с современной компьютерной техникой, идея о программном обеспечении, наделенном сознанием, кажется невероятной. Однако это интуитивное недоверие является продуктом относительно жалких возможностей сегодняшних компьютеров. Благодаря продолжающемуся усовершенствованию железа и программного обеспечения, компьютеры будут становиться все в большей степени разумными и сознательными. На самом деле, с учетом склонности человека одушевлять все, что хотя бы отдаленно похоже на человека, люди могут начать наделять компьютеры сознанием задолго до того, как это станет реальностью.

Аргументы в пользу «независимости от субстрата» изложены в соответствующей философской литературе, и я не буду пытаться их воспроизводить в данной статье. Однако я укажу на то, что это допущение разумно. Клетка мозга — это физический объект, обладающий определенными характеристиками. Если мы придем к полному пониманию этих характеристик и научимся воспроизводить их электронным путем, тогда, без сомнения, наша электронная мозговая клетка сможет выполнять те же функции, что и клетка органического происхождения. А если это можно проделать с одной клеткой мозга, то почему нельзя повторить ту же самую операцию с целым

мозгом? А если так, то почему бы получившейся системе не обладать таким же сознанием, как у живого мозга?

Эти предположения очень любопытны. Располагая достаточными вычислительными мощностями, постлюди могут создать модели исторических личностей, у которых будет полноценное сознание и которые будут считать себя биологическими людьми, живущими в более раннем времени. Этот вывод подводит нас к утверждению под номером два.

Первое утверждение предполагает, что мы проживем достаточно долго, чтобы создать постчеловеческую цивилизацию. Эта постчеловеческая цивилизация получит возможность разрабатывать симуляции реальности подобные Матрице. Во втором утверждении отражена возможность того, что постлюди решат не разрабатывать эти модели.

Мы можем вообразить, что в постчеловеческую эпоху интерес к разработке исторических симуляций исчезнет. Это означает существенные изменения в мотивации людей постчеловеческой эпохи, ибо в наше время, разумеется, найдется немало людей, которым бы захотелось запустить модели предшествующих эпох, если бы они могли позволить себе это сделать. Однако, вероятно, многие из наших человеческих желаний покажутся глупыми любому постчеловеку. Может быть, симуляции прошлого будут представлять незначительную научную ценность для постчеловеческой цивилизации (что не так уж невероятно с учетом ее сложнооценимого интеллектуального превосходства), и, может быть, постлюди будут считать развлечения очень неэффективным способом получения удовольствия, которое можно получить куда проще: например, при помощи непосредственной стимуляции центров наслаждения головного мозга.

Этот вывод предполагает, что постчеловеческие общества будут весьма отличаться от человеческих: в них будут отсутствовать относительно обеспеченные и независимые субъекты, владеющие всей полнотой человеческих желаний и свободные действовать под их влиянием.

При другом раскладе возможно, что у некоторых постлюдей может появиться желание запустить симуляции прошлого, однако постчеловеческие законы помешают им сделать это. Что приведет к принятию подобных законов? Можно предположить, что развитые цивилизации пойдут по пути, который приведет их к признанию этического запрета на запуск моделей, имитирующих историческое прошлое, из-за страданий, которые выпадут на долю героев подобной модели. Однако с нашей сегодняшней точки зрения не очевидно, что создание человеческой расы есть безнравственное действие. Наоборот, мы склонны считать существование нашей расы процессом огромной этической ценности. Более того, одного существования этических воззрений об аморальности запуска симуляций прошлого недостаточно. К нему должно добавиться наличие такой социальной структуры в общецивилизационном масштабе, которая позволяет эффективно предотвращать деятельность, которая считается безнравственной.

Итак, второе утверждение верно только в том случае, если мотивации постлюдей либо будут разительно отличаться от мотиваций людей, либо постлюди будут должны наложить тотальный запрет на симуляции прошлого и эффективно контролировать действие этого запрета. Более того, этот вывод должен быть справедливым почти для всех постчеловеческих цивилизаций во Вселенной.

Следовательно, нам необходимо рассмотреть следующую вероятность: не исключено, что у цивилизаций человеческого уровня есть шанс стать постчеловеческими, далее, по крайней мере в некоторых постчеловеческих цивилизациях найдутся отдельные личности, которые запустят симуляции прошлого. Это подводит нас к нашему третьему утверждению: мы почти наверняка живем в компьютерной симуляции. К этому выводу мы приходим вполне естественно.

Если постлюди будут запускать симуляции прошлого, скорее всего, это будет происходить в очень широких масштабах. Не составляет труда представить миллионы индивидуумов, запускающих

тысячи вариантов симуляций на сотни различных тем, и в каждой такой симуляции будут задействованы миллиарды смоделированных личностей. Этих искусственных людей наберется многие триллионы. Все они будут считать, что они настоящие и живут в более раннем времени.

Сейчас, в 2003 году, на планете живет примерно шесть миллиардов биологических людей. Очень даже возможно, что в постчеловеческую эпоху триллионы созданных на основе компьютерных программ людей будут жить в смоделированном для них 2003 году, убежденные в том, что они биологического происхождения: точно такие же, как вы и я.

Математика здесь проста, как дважды два: подавляющее большинство этих людей ошибаются: они считают, что они из плоти и крови, но на самом деле они таковыми не являются. Нет причин исключать нашу цивилизацию из этих подсчетов. Таким образом, почти все шансы сводятся к тому, что мы живем в смоделированном 2003 году и что наши физические тела являются компьютерной иллюзией.

Стоит подчеркнуть, что рассуждения методом подобного логического построения не преследует цель доказать, что мы живем в компьютерной симуляции. Оно отражает лишь то, что, по крайней мере, одно из трех перечисленных выше утверждений верно.

Если кто-то не согласен с выводом о том, что мы находимся внутри симуляции, то вместо этого ему придется согласиться либо с тем, что практически все постчеловеческие цивилизации откажутся от запуска симуляций прошлого, либо с тем, что, мы начнем вымирать, не достигнув постчеловеческой эпохи.

Наше исчезновение может произойти в результате стагнации имеющегося сейчас прогресса в области вычислительной техники или стать следствием общего коллапса цивилизации. Либо вы должны признать, что научно-технический прогресс, по-видимому, будет набирать обороты, а не стагнироваться, и в этом случае вы могли бы предсказать, что ускорение прогресса и станет причиной нашего исчезновения. Подвести нас к этому печальному концу может, к примеру, молекулярная нанотехнология. Достигнув развитой стадии, она позволит создавать самовоспроизводящиеся наноботы, способные питаться пылью и органикой, эдакие механические бактерии. Такие наноботы, если они созданы с недобрыми намерениями, могут вызвать исчезновение всей жизни на нашей планете. В другой своей работе я попытался перечислить основные экзистенциальные опасности, угрожающие человечеству.

Если наша цивилизация действительно является симуляцией, отсюда не обязательно должна вытекать необходимость ограничивать наш прогресс. Не исключено, что смоделированные цивилизации могут стать постчеловеческими. Тогда они могут запустить свои собственные симуляции прошлого, используя мощные компьютеры, которые они создадут в своей искусственной Вселенной. Подобные компьютеры будут «виртуальными машинами». Этот термин знаком современной вычислительной технике. (К примеру, основанные на Java web-приложения используют виртуальную машину — смоделированный компьютер — внутри вашего «рабочего стола»).

Виртуальные машины можно объединять в последовательный кластер: смоделировать машину, моделирующую другую машину, и т. д., при этом итераций может быть произвольно много. Если мы действительно добьемся создания наших собственных моделей прошлого, это будет веским доказательством против второго и третьего утверждений, так что нам волей-неволей придется заключить, что мы живем в смоделированном мире. Более того, мы должны будем подозревать, что постлюди, управляющие моделью нашего мира, сами являются искусственно созданными существами, а их создатели, в свою очередь, могут тоже оказаться смоделированными.

Таким образом, реальность может оказаться многоуровневой (эта тема затрагивалась во многих научно-фантастических работах, особенно в фильме «Тринадцатый этаж»). Даже если

иерархической структуре на каком-то этапе необходимо замкнуться на саму себя — хотя метафизический статус этого утверждения не вполне ясен, — в ней может размещаться огромное количество уровней реальности, и с течением времени это количество может возрастать.

Один из доводов против мультиуровневой гипотезы состоит в том, что затраты на вычислительные ресурсы для базовых моделей будут очень велики. Моделирование даже одной постчеловеческой цивилизации может быть непомерно дорогостоящим мероприятием. Если так, то нам следует ожидать уничтожения нашей модели при приближении к постчеловеческой эпохе.

Несмотря на то, что все элементы описанной выше системы естественны и даже материальны, здесь можно провести некоторые вольные параллели с религиозными представлениями о мире. В каком-то смысле постлюди, запускающие симуляцию, похожи на богов по отношению к людям, населяющим эту симуляцию: постлюди создали окружающий нас мир; их уровень интеллекта намного превосходит наш; они «всемогущи» в том плане, что могут вмешиваться в жизнь нашего мира, даже способами, нарушающими его физические законы. К тому же они «всеведущи» в том смысле, что они могут наблюдать за всем, что у нас происходит. Однако все полубоги, за исключением тех, кто находится на базисном уровне реальности, подчиняются распоряжениям более могущественных богов, живущих на более глубоких уровнях.

Дальнейшие размышления на эту тему могут достичь своей кульминации в натуралистической теогонии, которая занималась бы изучением структуры этой иерархии и ограничений, наложенных на ее жителей, исходя из возможности того, что какие-то действия на их уровне могут повлечь за собой определенную реакцию со стороны обитателей более глубоких уровней. Например, если никто не может быть уверенным в том, что находится в основе иерархии, то любой должен учитывать возможность того, что за любые действия он может быть вознагражден либо наказан создателями модели. Возможно, последние будут при этом руководствоваться какими-то нравственными критериями. Жизнь после смерти станет реальной возможностью, как и реинкарнация. Из-за этой фундаментальной неуверенности, возможно, даже у основной цивилизации будут причины вести себя безупречно с точки зрения морали. Тот факт, что даже у этой цивилизации будет причина вести себя с соблюдением норм морали, разумеется, заставит еще в большей степени всех остальных стремиться вести себя точно так же, и так далее. Получится самый настоящий добродетельный круг. Возможно, каждый будет руководствоваться своего рода универсальным моральным императивом, повиноваться которому будет в интересах каждого, поскольку этот императив появился «ниоткуда».

В дополнение к моделям прошлого можно также рассмотреть возможность создания более избирательных симуляций, затрагивающих лишь небольшую группу людей или отдельного человека. В этом случае оставшаяся часть человечества превратится в зомбированных людей или в людей-теней — людей, смоделированных с минимальным уровнем детализации, достаточным для того, чтобы полностью смоделированные люди не замечали ничего подозрительного.

Не ясно, насколько моделирование людей-теней будет дешевле, чем моделирование полноценных людей. Далеко не очевидно, что какое-то существо может вести себя неотличимо от настоящего человека и в то же время быть лишенным сознательного опыта. Даже если такие отдельные симуляции и существуют, не следует предполагать, что вы находитесь в одной из них, пока вы не придете к выводу, что люди-тени более многочисленны, чем полные модели. Чтобы все условные личности попали в я-симуляцию (модель, имитирующая жизнь одного единственного разума), я-симуляций потребовалось бы в сто миллиардов раз больше, чем общих симуляций прошлого.

Также существует возможность того, что создатели симуляций уберут определенные моменты из жизней смоделированных существ и снабдят их ложной памятью о переживаниях, которые они должны были испытывать во время изъятых моментов. В этом случае можно рассмотреть следующее (притянутое за уши) решение проблемы зла: на самом деле страдания в мире не

существует, а все воспоминания о нем — это иллюзия. Разумеется, эту гипотезу можно серьезно воспринимать лишь тогда, когда вы не страдаете.

Если предположить, что мы живем в симуляции, то что этом может значить для нас, людей? Несмотря на высказанные выше соображения, последствия не так уж и радикальны. Самое обычное, привычное для нас эмпирическое исследование Вселенной, которую мы видим, лучше всего подскажет нам, как будут действовать (действовали) наши постчеловеческие создатели, устраивая этот мир. Пересмотр большей части наших убеждений, с учетом возможности нашего виртуального происхождении, приведет к довольно незначительным и даже едва заметным результатам — величина которых будет прямо пропорциональна нехватке уверенности в нашей способности понять логику постлюдей. Поэтому правильно понятая истина, содержащаяся в третьем утверждении, не должна «сводить с ума» или мешать нам продолжать заниматься своими делами, а также планировать и предсказывать завтрашний день.

Если мы узнаем больше о мотивациях постлюдей и об ограничениях на количество ресурсов, — а это может случиться в процессе нашего собственного движения к постчеловеческой цивилизации, — в этом случае гипотеза о том, что мы смоделированы, будет иметь гораздо более богатый набор эмпирических доказательств. Если печальная реальность все-таки заключается в том, что мы являемся симуляциями, созданными какой-то постчеловеческой цивилизацией, то можно считать, что нам выпала лучшая доля, чем обитателям Матрицы. Вместо того чтобы попасть в лапы враждебного ИИ и быть использованными в качестве источника энергии для его существования, нас создали на основе компьютерных программ как часть научно-исследовательского проекта. Или, может, нас создала какая-нибудь девочка-подросток из постчеловеческой цивилизации, выполняя домашнее задание. Тем не менее, нам все-таки лучше, чем жителям Матрицы. Разве нет?