Может ли машина с искусственным интеллектом быть сознательной, Часть II?

12 июля 2019 года

David Mumford

После публикации "может ли машина с искусственным интеллектом быть сознательной " я прочитал две книги, в которых утверждается, что сознательные искусственные интеллекты находятся в процессе своего развития, и мы должны подготовиться. Я хочу обсудить две книги и оценить свои собственные аргументы в их свете. Первой я прочитал художественную литературу: последняя книга Иэна Макьюэна «Машины как я», в которой небольшая группа сознательных роботизированных мужчин и женщин производится и продается по всему миру. Его роман, возможно, является идеальным способом изобразить концепцию сознательных роботов как правдоподобную, так и пугающую. Два главных героя не сомневаются в том, что их Адам (как его называют) сознателен, как и его персонаж Тьюринг (версия Тьюринга, который живет долгой и удивительной научной жизнью). Но это плохо закончится! Вторым произведением, которое я изучил был «Туннель эго» немецкого философа Томаса Метцингера. Эта книга исчерпывающе исследует, что такое сознание с биологической, психологической, информационно-теоретической и философской точек зрения. В ней представлены очень важные данные из внетелесных переживаний, осознанных сновидений и многого другого. После анализа того, что происходит в человеческом мозге, он пишет раздел, озаглавленный "как построить искусственный сознательный субъект и почему мы не должны этого делать", в котором излагает, как это действительно может быть сделано. Позвольте мне рассмотреть обе книги более подробно.

Писатель

Предупреждение о спойлере: я не могу обсуждать «Машины как я», не раскрывая сюжет. Макьюэн погружается прямо в их робота Адама, влюбляется и спит с Мирандой, подругой покупателя Адама Чарли. Несмотря на то, что ему нужно регулярно подзаряжаться от розетки в пупке, он был наделен основными человеческими эмоциями, отчасти благодаря тому, что Чарли и Миранда щелкали по набору онлайн-вариантов. Затем он ломает запястье Чарли, когда тот неосторожно тянется к кнопке выключения на его шее. Но они продолжают сражаться, когда Адам извиняется перед Чарли, только чтобы обнаружить в развязке, что его идея морального поведения полностью не согласуется с шаткими моральными компромиссами человечества, с действиями, которые отправляют Миранду в тюрьму. Чарли из любви к Миранде разбивает Адаму череп, и Тьюринг клеймит его убийцей.

Макьюэн, безусловно, делает вывод со своей точной точки зрения: человеческие эмоции чрезвычайно сложны и запутанны, и поэтому приходится задаться вопросом, может ли робот когда- либо по-настоящему "понять" их. Тем не менее я утверждал, что существенной частью сознания является именно "чувство" эмоций. Я заключаю это в кавычки, поскольку чувство и понимание-это слова, которые касаются того, что такое сознание. Мне кажется, что Макьюэн делает слишком тонкую точку зрения, позволяя Адаму много сильных эмоций, но не давая ему более глубокого понимания того, как работают эмоции.

Его неудача подчеркивает модель поведения человека, выраженную словом "лояльный". Это слово относится к смешению эмоций и паттернов действий, как прошлых, так и будущих, и является типичным для сложного переплетения эмоций и действий в человеческих существах. Например, центральными принципами шотландской этики вполне могут быть бережливость, честность и верность, причем все три эти качества являются эмоционально нагруженными действиями. Адам бережлив и честен, но терпит неудачу в требованиях верности. С другой стороны, моя кузина Рут Силкок написала серию детских книг о коте по имени Альберт Джон. В своей первой книге она написала "Альберт Джон был верным котом", полагая, что эта концепция была совершенно ясна ее юным читателям. Но не для Адама. Таким образом, по большому счету, Макьюэн согласен с моим убеждением, что моделирование человеческих эмоций и их результирующей деятельности в роботе- это огромный барьер, хотя его персонажи действительно видят своего робота достаточно эмоциональным, чтобы считаться сознательным.

Философ

Книга Метцингера-это, без сомнения, самое глубокое исследование природы сознания, которое я когда-либо читал. Его основной тезис состоит в том, что наш мозг конструирует для нас феноменальную модель себя, под которой он понимает "сознательную модель организма в целом, которая активируется мозгом", и которую он также называет эго (стр. 4) Он уточняет это следующим образом (стр.7):

Сначала наш мозг создает симуляцию мира, настолько совершенную, что мы не воспринимаем ее как образ в нашем сознании. Затем они создают внутренний образ нас самих как целого. Этот образ включает в себя не только наше тело и наши психологические состояния, но и наши отношения с прошлым и будущим, а также с другими человеческими существами. Внутренний образ человека- как-целого - это феноменальное эго, "я" или "личность", как оно проявляется в сознательном опыте.

Он говорит, что мы чувствуем, что сознательно переживаем ощущения, с которыми наши тела сталкиваются в мире, потому что этот целостный внутренний образ нас самих прочно закреплен в наших чувствах и телесных ощущениях, и потому что мы неспособны признать наши модели самих себя просто моделями, потому что они прозрачны, как стеклянное окно, через которое мы видим мир. Таким образом, он вынужден описывать жизнь, которую мы ведем, как туннель эго. Наши умы заполнены моделью, которую мы принимаем за реальность, поэтому мы находимся в туннеле, по которому движемся с течением времени. Хотя он не упоминает Шопенгауэра, многое в этой теории похоже на идеи Шопенгауэра: Die Welt ist meine Vorstellung (мир-это мое представление) - это утверждение, с которым он открывает свой магнум-опус Die Welt als Wille und Vorstellung

Метцингер широко использует так называемую иллюзию резиновой руки. Здесь испытуемый сидит за столом, его левая рука находится за перегородкой, но резиновая левая рука помещена на стол перед ним. Затем резиновую руку щекочет перо, в то время как его настоящая левая рука тоже невидимо щекочется. Через некоторое время испытуемый начинает ощущать, что резиновая рука-его собственная, что невидимая рука соединяет ее с его телом и только щекоча ее, он чувствует, что его настоящая рука щекочется. Метцингер интерпретирует это как обман ума, заставляющий его изменить свою модель себя в нереальное представление, которое все еще кажется абсолютно реальным. Точно так же он подробно обсуждает такие феномены, как внетелесные переживания и осознанные сновидения (когда вы осознаете, что спите, но все еще чувствуете, что живете в ярком убедительном мире сновидений). Странно, но он не описывает некоторые другие эксперименты виртуальной реальности, такие как тот, где, надев очки, которые показывают, как вы идете по виртуальному утесу, вы падаете с неподдельным страхом (хотя на самом деле на ковер в пустой

комнате). Я был испытуемым и пережил это в Брауне. Он также не рассматривает огромный виртуальный мир в фильме "Матрица" и нынешнюю моду на очки виртуальной реальности и иммерсивные развлечения. Но, конечно, это только усиливает его аргумент, что мы живем в модели себя и можем слишком легко быть обмануты, принимая альтернативный мир за реальность.

Мудрец

Моя любимая история из богатого наследия индуистской мифологии-это история мудреца Нарады и его поисков понимания Майи Вишну. Это иллюстрирует, что феноменальная модель Метцингера имеет предшественников, которые восходят, по крайней мере, к первому тысячелетию до нашей эры. Она начинается с того, что Нарада совершает так много аскез, что он обретает духовную силу просить Вишну о милости. Он просит Майю понять (древнее санскритское слово, означающее "иллюзия"). Эта история продолжается в рассказе Генриха Циммера ("мифы и символы в индийском искусстве и цивилизации", стр. 32-34)

"Покажи мне магическую силу твоей Майи", - молился Нарада, и Бог ответил: "Я сделаю это. Пойдемте со мной, - с двусмысленной улыбкой на красивых изогнутых губах. Выйдя из приятной тени укрытой отшельнической рощи, Вишну повел Нараду через голый участок земли, сверкающий, как металл, под безжалостным сиянием палящего солнца. Вскоре им очень захотелось пить. На некотором расстоянии, в ярком свете, они увидели соломенные крыши маленькой деревушки. Вишну спросил: "не пойдешь ли ты туда и не принесешь ли мне воды?- Конечно, о Лорд, - ответил святой и направился к дальней группе хижин. Бог расслабился в тени скалы, ожидая его возвращения.

Добравшись до деревни, Нарада постучал в первую же дверь. Прекрасная девушка открыла ему, и святой человек испытал то, о чем до сих пор никогда не мечтал: очарование ее глаз. Они были похожи на его божественного повелителя и друга. Он стоял и смотрел. Он просто забыл, зачем пришел. Девушка, кроткая и искренняя, приветствовала его. Ее голос был золотой петлей на его шее. Двигаясь как в видении, он вошел в дверь. Обитатели дома были полны уважения к нему, но ни капельки не стеснялись. Его приняли с почетом, как святого человека, но почему-то не как чужестранца, а скорее, как старого и уважаемого знакомого, который долго отсутствовал. Нарада остался с ними, впечатленный веселой и благородной осанкой и чувствуя себя совершенно как дома. Никто не спрашивал его, зачем он пришел; казалось, он принадлежал к этой семье с незапамятных времен. И через некоторое время он попросил у отца разрешения жениться на девушке, чего все в доме ожидали. Он стал членом семьи и делил с ними вековые тяготы и простые радости крестьянского хозяйства.

Прошло двенадцать лет, у него было трое детей. Когда его тесть умер, он стал главой семьи, унаследовав поместье и управляя им, ухаживая за скотом и обрабатывая поля. На двенадцатый год сезон дождей был необычайно бурным; ручьи вздулись, потоки хлынули с холмов, и маленькая деревня была затоплена внезапным наводнением. Ночью соломенные хижины и скот были унесены, и все разбежались. Одной рукой поддерживая жену, другой ведя за собой двоих детей, а самого маленького взвалив на плечо, Нарада поспешно двинулся вперед. Пробираясь вперед сквозь кромешную тьму и хлещущий дождь, он шел по скользкой грязи, шатаясь, через бурлящие воды. Ноша была больше, чем он мог выдержать, когда течение тяжело тащило его ноги. Однажды, когда он споткнулся, ребенок соскользнул с его плеча и исчез в ревущей ночи. С отчаянным криком Нарада отпустил старших детей, чтобы поймать самого маленького, но было уже слишком поздно. Тем временем поток быстро унес двух других, и еще до того, как он успел осознать случившееся, он оторвал от себя жену, вырвал из-под себя собственные ноги и его швырнуло головой вперед в

поток, как бревно. Потеряв сознание, Нарада в конце концов застрял на небольшом утесе. Придя в себя, он открыл глаза и увидел огромную полосу мутной воды. Он мог только плакать.

- Дитя мое!- Он услышал знакомый голос, от которого у него чуть не остановилось сердце. -Где та вода, за которой ты ходил для меня? Я жду уже больше получаса. Нарада обернулся. Вместо воды он увидел сверкающую в лучах полуденного солнца пустыню. Он обнаружил, что Бог стоит у его плеча. Жестокие изгибы очаровательного рта, все еще улыбающегося, расступаются с нежным вопросом: "Постиг ли ты теперь тайну моей Майи?

Говоря языком Метцингера, я бы интерпретировал эту историю следующим образом: Вишну поместил вилку в туннель эго Нарады и повел его вниз по новой вилке, попросив воды. Новая развилка была длинной и в конечном счете привела к тому, что Нарада сам утонул. Но затем Вишну заставил новую вилку присоединиться к старой с легкой жестокой улыбкой на лице. Таким образом, Майю можно рассматривать как описание феноменального образа самого себя, убедительную реальность, но лишь маленькое окно в то, что находится снаружи, созданное нашим ограниченным сознанием.

Время

В более поздней главе "Путешествие по туннелю" Метцингер обсуждает различные проблемы, связанные с тем, как модель " Я " становится настолько реальной, что создает ощущение сознания. Именно здесь он разделяет со мной мнение, в разделе, озаглавленном "Проблема настоящего: возникает переживаемый момент" (стр. 34). Я не мог не согласиться с его предложением: "полное научное описание физической Вселенной не содержало бы информации о том, что такое время "сейчас". Как я уже писал, наука, почти по определению, ищет законы, которые действуют независимо от времени и места, и оставляет все, что связано с прошлым/настоящим/будущим, историкам и футуристам. Как уже упоминалось в моем предыдущем посте, сам Эйнштейн считал, что "переживание настоящего означает нечто особенное для человека, нечто существенно отличное от прошлого и будущего, но что это важное различие не имеет и не может иметь места в физике". Можно было бы предположить, что мышление Эйнштейна было также мотивировано его специальной теорией относительности, в которой было показано, что одновременность не имеет физического смысла, так что нельзя думать о Вселенной как о целом, имеющем прошлое и настоящее, постоянно разворачивающееся по мере того, как происходят будущие события. Из-за теории относительности у каждого человека есть свое настоящее, и как только космические путешествия станут реальными, будет совершенно невозможно поверить в объективное настоящее. В истории Нарады 12 лет для Нарады-это полчаса для Вишну, и космические путешествия могут сделать это и для двух человек.

И действительно, Метцингер продолжает утверждать: "моя идея состоит в том, что именно эта одновременность является причиной того, что нам нужно сознание Сейчас". Хорошо известно, что ум играет быстро и свободно с параллельностью, так что два сигнала могут быть восприняты сознательно как происходящие в порядке, противоположном их возникновению в физическом мире. Временная упорядоченность, по-видимому, в какой-то степени является конструкцией, которую разум создает как можно лучше. Но теперь Метцингер переворачивает логику. Исходя из того, что переживание "сейчас" подразумевает переживание параллельности, он хочет сказать, что переживание параллельности создает переживание "сейчас". Он утверждает, что создание общей временной системы отсчета для всех механизмов в мозге приводит к внутренней модели окружающего мира такого Сейчас (стр. 36). Здесь я не могу разделить его мнение: все компьютеры имеют часы и соответственно организуют свои вычисления и коммуникации. Но я не думаю, что у

них есть сознание. Нейроны передают сигналы гораздо медленнее, чем свет, и это, вероятно, как-то связано с ошибками временного порядка, которые делает мозг. Но прояснение большинства вещей во времени кажется мне недостаточным для того, чтобы вызвать сознательное ощущение настоящего момента. Вы, читатель, будете читать это в какой-то момент в будущем, и тогда я буду делать что-то еще (или буду мертв), таким образом, каждый из нас живет через нашу собственную уникальную последовательность Сейчас. Мы можем быть в некоторой степени синхронизированы, когда читаем один и тот же выпуск "Нью-Йорк Таймс", но тем не менее наши Сейчас различны, и наша синхронизация обречена на некоторую степень неопределенности.

Как уже упоминалось выше, Метцингер излагает применение этих идей к построению сознательных роботов в более позднем разделе "Как построить искусственный сознательный субъект и почему мы не должны этого делать" (стр. 190). На стр. 192 он описывает строительство в четыре этапа. Первый
- это наделить машину постоянно обновляющимся интегрированным внутренним образом мира. Второй - организовать свой внутренний информационный поток во времени, в результате чего возникает психологический момент, переживаемое Сейчас. Третье-убедиться, что эти внутренние структуры не могут быть распознаны искусственной сознательной системой как внутренне порожденные образы, поэтому они прозрачны. Четвертый шаг-это интеграция столь же прозрачного внутреннего образа самого себя в феноменальную реальность. Ни я, ни Метцингер не считаем ни то, ни другое невероятно трудным. Но это другой момент, где я чувствую, что он предполагает, что слишком много происходит в результате активности кремния. Сейчас мне кажется, что это поистине волшебный шаг, шаг, который создает то, что Поппер называет миром II, а другие называют духовным.

Тот факт, что наша жизнь протекает как путешествие вниз по реке времени, что мы всегда осознаем, что находимся в определенном месте, окруженном локальным кусочком трехмерного мира, который меняется по мере того, как "время идет", все это кажется очевидным и здравым, а вовсе не волшебным. Но это происходит потому, что это переживание времени является ядром сознания каждого, повседневной жизни каждого, а не потому, что в физике или в любой другой науке есть что-то подобное потоку времени с настоящим моментом, освещенному подобно маяку. В физике время статично, это просто один из способов поместить координаты на 4-мерный Панцирь всех событий, прошлых, настоящих и будущих, и в любом месте вообще. Мы можем искусственно построить математический "поток", одномерную группу гомеоморфизмов пространства-времени, но такой поток не задается физикой и не имеет определенного набора точек, называемых настоящим моментом. Жить в мире времени кажется мне чудесным даром, и я понятия не имею, как, подобно Богу и Адаму на потолке Сикстинской капеллы, можно передать этот дар роботу.

переведено: Andy Michael
Author of the best VPN research.