№ 17: ДУША І МАШИНА

Мечтают ли андроиды об электроовцах: разговор с Евгением Пилецким

ПАВЛО НЕДАШКІВСЬКИЙ

6 ЛЮТОГО 2019

Как скоро мы получим электронных овец, неотличимых от живых? И как скоро мы встретим андроидов, мечтающих об электроовцах? Сегодня мы поговорим с Евгением Пилецким о роли бессознательного в теориях Искусственного Интеллекта, функционалистском подходе, и о том, как боты в компьютерных играх видят нас, а мы их.

Джерело: Нікола Самори, «Lingua greca»

Современные исследователи различают понятия «сильного» и «слабого» Искусственного Интеллекта (AI), в чем их отличие?

Согласно современным теоретическим представлениям, «сильный» Искусственный Интеллект должен обладать, по крайней мере, несколькими отличительными характеристиками, среди которых наиболее существенная – это деятельность интеллектуального агента от первого лица. Теоретически это должна быть машина самоцелеполагания.

Естественно, желательно, чтобы такая машина обладала сознанием и самосознанием (тем, что мы сейчас подразумеваем под этими словами). При этом отличительными признаками сознания и самосознания являются интенциональность и система глобального доступа к когнициям.

В свою очередь, «слабый» Искусственный Интеллект – это те системы Искусственного Интеллекта, которые сегодня практически повсеместно развиваются. Здесь возникает

вопрос: могут ли количественно различные элементы «слабого» AI дать качественный прыжок к «сильному» AI?

Так, согласно теории Джерри Алана Фодора, мы сами работаем на основании «модулей» (так называемого «modular mind»). Если будет создана конфигурация, которая будет отыгрывать то ли систему «синхронной осцилляции групп нейронов», то ли какую-то иную в рамках отдельных нейросетей (именно их мы сегодня чаще всего называем «слабым» Искусственным Интеллектом), то, возможно, мы и получим «сильный» Искусственный Интеллект.

По сути, «слабый» Искусственный Интеллект – это функциональные нейронные сети различных типов (сверточные и т.д.), которые представляют собой системы с множественными входами, аналитическими системами, и одним или n-количеством выходов. Широко известное их применение – это распознавание образов или речи, т.е. то, что называется «машинной перцепцией».

Здесь можно использовать нейронно-сетевую метафору «вероятностной машины» Алана Тьюринга, которая оценивает информацию на основании больших данных (big data). Например, я распознаю лицо в динамике, потому что у меня есть огромное количество входящих данных, которые интерпретируются так же, как это происходит в современных нейронных сетях. На выходе я имею определенный результат. На основании больших данных уже можно строить прогностические модели и т.д. Но для такой машины все равно нужен внешний интерпретатор, который, собственно, делает выводы.

Мы сами не уверены в том, чем будет Искусственный Интеллект. Вернее, мы не уверены в том, что то, что будет представляться нам Искусственным Интеллектом, на самом деле будет им.

С другой стороны, мы не знаем, как функционирует внешний интерпретатор. В частности, если мы уже сегодня используем внешние нейросети, которые способны осуществлять семантический анализ изображения, а значит описывать его естественным языком (пусть даже и формализированным), то вполне возможно, что и мы так работаем.

Есть некая система, которая интерпретирует (или нам, по крайней мере, кажется, что

интерпретирует) процессы, собирая их воедино. Наша эволюция привела к такой интегральности, назовем ее «ипостасной»: мы – это нечто единое, обладающее некоей природой. Хотя неизвестно, нужно ли воспроизведение такой системы для создания «сильного» Искусственного Интеллекта или нет.

В отличие от «сильного» Искусственного Интеллекта, «слабый» не является системой ни сознающей, ни самосознающей. Она не может выйти в третью позицию по отношению к самой себе и рефлексировать относительно собственных эмоций и мыслей.

Здесь возникает еще один очень хороший вопрос: являются ли эмоции и мысли по отношению к этому условному субъекту чем-то внешним? Хотя здесь мы немного отступаем от темы, однако, это применимо к агентской миссии Искусственного Интеллекта как разумной машины. По всей видимости, нам только кажется, что мы обладаем эмоциями, хотя эмоции – это скорее то, что с нами происходит, а не то, что мы делаем. Да, у нас есть некий интенциональный импульс, но что происходит дальше, нам совершенно не ясно.

Почему это важно для сильного искусственного интеллекта? Потому что значительная часть вещей, которые мы делаем, иррациональны. Вся наша мотивация и наша эмоционально-волевая сфера являются для нас закрытыми системами. Почему они чего-то хотят? Зачем им что-то делать? Какую они имеют цель? В частности, мы сейчас разговариваем. Почему? Потому что мне это интересно. Но почему мне интересно? Почему я хочу говорить? Почему мне интересно это, а не другое? Я не знаю. Я это не выбирал. Да, у меня есть какой-то опыт. Но почему мне понравился этот опыт? Я не знаю, мне это не доступно.

Для создания «сильного» Искусственного Интеллекта мы должны ответить на вопросы, которые уже несколько раз поднимались, например Рэймондом Курцвейлом. Речь идет о «бессознательном» Искусственного Интеллекта. Иными словами, что должно побуждать к действию? Сейчас мы этого не знаем. То ли это должна быть естественная эволюция модульных кусочков «слабого» Искусственного Интеллекта; то ли мы сами должны будем запрограммировать его мотивацию, в стиле Айзека Азимова (т.е. хотя бы отрицательную: не делать чего-то); то ли иным способом; то ли у «сильного» Искусственного Интеллекта «бессознательного» вообще не должно быть? Это сложный вопрос.

Еще один важный момент: мы сами не уверены в том, чем будет такой Искусственный Интеллект. Вернее, мы не уверены в том, что то, что будет представляться нам Искусственным Интеллектом или деятельностью агентского начала, на самом деле будет им.

Каким образом в нейронауках и теориях Искусственного Интеллекта объясняется переход

перспективы «от третьего лица» (ввод, анализ, вывод данных) к перспективе «от первого лица» (я мыслю, я ощущаю)?

Например, наша система боли может быть описана как система датчиков повреждения, только у нас она дополнительно окрашивается сильными интенциями: мне больно и одновременно мне плохо и это заставляет меня что-то делать (подать сигнал или что-то изменить). В свою очередь, при повреждении машина мне высвечивает индикатор: это ее боль или нет? Это еще один хороший вопрос. Современные философы сознания подходят к этому вопросу с различных и даже противоположных сторон - от Дэниела Деннета до Дэвида Чалмерса - хотя между ними есть еще и такие неофункционалисты, как Франсиско Варела.

Воспользуемся еще одной компьютерной метафорой. Предположим, я играю в компьютерную игру Battlefield. Что происходит на физическом уровне? Изменяется только магнитное состояние вещества. Есть поток электронов, которые, условно говоря, бегают по логическим вентилям процессора и общаются с памятью с определенной тактовой частотой. Там не происходит никакой игры, в нашем смысле, т.е. из перспективы от первого лица.

Возьмем еще один элементарный пример. Фотография на флешке. Пока я не подключил ее к компьютеру, есть там фото или его нет? Если мы исходим из перспективы от третьего лица, то там просто набор магнитного вещества. Когда я вставляю флешку, имеет место интерпретация – другой уровень обработки. Когда я смотрю (здесь можно справедливо сказать, ведь это я смотрю, не компьютер смотрит), я вижу фото.

Но представим себе, что у нас работает монитор и речь идет не о «философском зомби», а мы действительно можем наблюдать фотографию. Мы интерпретируем поток фотонов, которые никак прямо не связаны с состоянием вещества на флешке не подключенной к компьютеру. При этом даже если я выключу монитор и на нем ничего не будет отображаться, то тот же сигнал все равно будет идти от компьютера к монитору. Эта фотография все равно будет существовать неким «виртуальным» образом.

Предположим, я играю в компьютерную игру Battlefield. Что происходит на физическом уровне? Есть поток электронов, которые, условно говоря, бегают по логическим вентилям процессора и общаются с памятью с определенной тактовой частотой. Там не происходит

никакой игры, в нашем смысле, т.е. из перспективы от первого лица.

Поэтому сейчас очень хитро начинают подходить к вопросу об онтологии виртуальности. Потому что ответ на вопрос «что такое сознание?» с чисто философской (не математической, не информатической) точки зрения, будет одновременно и ответом на вопрос «что такое информация?».

Китайские иероглифы, на которые я смотрю, они что-то значат или нет? Когда появляется интерпретатор, то что-то возникает. Когда мы говорим на русском языке, очевидно, что мы кодируем и декодируем информацию моментально. Но для англоязычного человека, это была бы только некая звуковая мелодика, не более. Это одна часть вопроса.

Перейдем ко второй части вопроса. Я играю в незаскриптованную компьютерную игру. Например, в тот же Battlefield. Боты ведут себя абсолютно по-разному, не бывает двух одинаковых ситуаций. Понятно, что работа ботов предусмотрена определенными алгоритмами. Но когда я играю в компьютерную игру, я не вижу алгоритма. Я вижу, как бот стреляет в меня, прячется, ищет аптечку, патроны и т.д. При этом он это делает не по строгим прописанным рельсам, а на основании алгоритма. В какой форме представлен алгоритм? Написано «беги»? Нет. Он написан в виде импликаций «если…, то».

Кроме того, эти импликации вызывают другие скрипты, например: «бежать», «идти», «пригнутся» и т.д. Этих скриптов, на самом деле, гигантское количество. Но, играя, я этого не вижу - я играю с ботом, а не с алгоритмом. Вопрос: «для этого бота, когда он, условно, виртуально видит меня, что-то происходит?». Возможно, что это что-то вроде «философского зомби» или же у бота есть какие-то прото- или псевдоквалиа. В частности, он видит меня как персонажа. Но при этом ясно, что для него, это не более чем математические операции. Он оценивает мое состояние по определенным правилам и ведет себя соответственно. Для него я представлен в виду определенных запрограммированных «квалий», как, к примеру: «опасность», «направление» и т.д.

Перенося эту виртуальную модель на нас самих, и, дополнив ее некоторыми другими форматами (например, теми, о которых говорил Денетт: «исчезающая квалиа», «проявляющаяся квалиа» и т.д.), мы можем сказать, что такой же виртуальный характер носит и наша квалийность. Вопрос в том, что мы как будто постоянно исходим из вещественной онтологии.

У меня возникает ощущение, что виртуальность находится не совсем там. Иными словами, мы не можем сказать, что «ее нет» и не можем утверждать, что она «есть» в эссенциальном смысле.

Но как раз Денетт критикует такую позицию, озвученную, например, Чалмерсом. В метафизике Деннета - мир строго монистический и материалистический.

Если Деннет полагает что мир монистический, то что это значит? Монизм мира означает наше представление о нем, или что-то другое? Тогда это ответ на вопрос «что такое информация?». Смысл, например, является надстройкой над информацией. Точно так же как пучок фотонов является неким другим уровнем интерпретации того, что происходит на жестком диске.

Что тогда с математикой, ведь она обладает нормативностью и оперирует нематериальными объектами, при этом продуктивна при описании материальных объектов и их свойств?

Я думаю, что это такая же виртуальная структура. Я не могу говорить об онтологии математики. Я не могу ответить на этот вопрос как Кант.

Это серьезный вопрос, но дело в том, что любая информационная модель может быть настолько же продуктивной. «Денеттовость» нам здесь может смягчить тот факт, что у информации всегда есть какой-то субстрат. В частности, у песни на флешке есть конкретный субстрат. Для него нужен интерпретатор, но это уже другой вопрос. И если это так, то вопрос от первого лица - это вопрос конечной точки интерпретатора. Но здесь мы можем сказать, что это наше представление и ее может и не быть. В данном случае интерпретатор может быть одновременно и интерпретацией. Это одна сторона.

Мы просто привыкли к некой бинарности: «что-то есть, потому, что оно вещественно».

С другой стороны, имеем вопрос о той же математике. Математическая физика описывает нечто или субстратом математической физики являются аналоговые силы типа гравитации или чего-то еще? Математика является описанием природы или моделью природы?

Она также абсолютно виртуальна. В природе не существует вектора, траектории, импульса, гравитации и т.д. Тем не менее, математика каким-то имплицитным образом может

содержаться и, соответственно, она в виде формализации существует в нас. Но каким образом?

Я думаю таким же образом как наше сознание, квалиа и все остальное. Возможно (это только лишь рабочая гипотеза), виртуальность настолько же онтологична, как и все остальное. Мы просто привыкли к некой бинарности: «что-то есть, потому, что оно вещественно» (ведь об этом говорит эволюция, и мы привыкли работать с вещественными объектами). Поэтому мы точно также пытаемся овеществить онтологию квалиа. Хороший вопрос, можно ли вообще при таком подходе ответить на вышеперечисленные вопросы?

Более пятидесяти лет назад Тьюринг, так же как и современные нам функционалисты, исходил из метафизического принципа «agere sequitur esse» (способ действия проистекает из бытия). Иными словами, если не эксперт не сможет отличить ответы машины и человека, то машина и человек равно обладают интеллектом. Однако обычный человек, точно также, не всегда сможет отличить золото и пирит. Различают ли функционалисты познаваемую нами реальность и познавательные способности?

Я думаю, что это вопрос классической онтологии. В частности, вспоминается пример с уткой. Если что-то выглядит как утка, то это утка или «как-утка» или «будто-утка», но не утка? В данном случае, это снова вопрос об овеществлении. Это вопросы тождественности предметов, различимости предметов и т.д.

Если машина говорит, что у нее есть квалиа, что она что-то чувствует, что она сознательна и т.д., то можем ли мы в этом сомневаться?

Та же буддистская онтология говорит о том, что это бессмысленные вопросы. Имел место спор четырех школ о том, являются ли дхармы онтологически существующими или некими квази-объектами. У меня возникает ощущение, что вопрос «может ли что-то быть чем-то?» уже просто не актуален, по крайней мере, в этой области. Это скорее часть, условно говоря, технической философии. Нам удобно выделять объекты (мы так эволюционно развивались) и именно из-за подобной утилитарности мы и создаем такую онтологию. Нам важно знать, является или нет что-то чем-то, имеет положительный онтологический статус или не имеет такового.

Здесь возникает интересный вопрос. Если машина говорит, что у нее есть квалиа, что она что-то чувствует, что она сознательна и т.д., то можем ли мы в этом сомневаться? Это же классический вопрос. А могу ли я точно знать, что у моего собеседника есть квалиа?

Во время нашего разговора Вы несколько раз апеллировали к эволюции мозга. Однако в рамках строгого материализма, которого старается придерживаться тот же Деннет, эволюция – это череда случайных генетических мутаций. С этой же, материалистической, позиции наше мышление, включая математику и логику – это следствие, а не причина, процессов в мозге и никак не влияет на адаптивность нашего поведения. Исходя из такой позиции, насколько можно быть уверенным в истинности того или иного умозаключения?

Это вопрос разницы между случайностью как хаосом и упорядоченностью. Когда мы представляем себе не упорядоченные вещи, то мы исходим из ложного убеждения, что они хаотичны, в том смысле, что они являются абсолютным беспорядком. Это достаточно наивно, потому как для нас вопросы хаоса и порядка часто являются бытовыми. Мы не очень хорошо представляем их себе на метауровне. Например, когда я подбрасываю монетку два раза, какова вероятность выпадения орла или решки?

Иными словами, Вы рассматриваете эволюцию не как хаотичный механизм, а как нечто, имеющее целеполагание?

Даже если есть некие правила, которые, начиная с Большого взрыва, возникают определенным образом, то они могут быть эмергентными. То, что мы формализируем математику, может быть не более чем нашим представлением о реальности. Значит, и математика может быть другой. Но эволюция, описанная только с биологической точки зрения, не выглядит как абсолютный хаос. Подбрасывая монетку, невозможно предсказать, что выпадет фактически, но возможно математически. Получается, чем больше данных, тем точнее ответ.

Я не могу предсказать фактически, что произойдет, но могу узнать тенденции. Да, в эволюции много случайных вещей. Но это не случайность в смысле «бытового» хаоса. Интересно и то, в какой мере хаос, или абсолютный антидетерменизм предельного уровня, может быть фикцией нашего разума? Является ли это в полной мере хаосом?

Warning: count(): Parameter must be an array or an object that implements Countable in **/home/kairosbo/verbum.com.ua/www/wp-includes/post-template.php** on line **284**