**Проблема языка в современных исследованиях по искусственному интеллекту**

Блинов А.К.

**1. Источник компьютерных аналогий**

В основании компьютерных аналогий ментального, широко используемых в когнитивистских моделях, лежит понятие машины Тьюринга. В работе 1950 года «Умеет ли машина мыслить?» Алан Матисон Тьюринг поставил задачу формулировки условий, при которых машина может быть описана как мыслящая. А поскольку понятие «мышления» весьма темное, то он полагает, что, прояснив, что значит мыслить для машины, мы сможем понять, что значит мыслить вообще. Тьюринг исходит из предпосылки, что если поведение машины неотличимо от поведения человека, то это – достаточный критерий считать такую машину мыслящей. Он предложил абстрактную модель машины, успешно имитирующей всю совокупность человеческого поведения, которая впоследствии получила название машины Тьюринга. В основе этой идеи лежат результаты, полученные в 30-х годах 20-го века и легшие в основу так называемой теории автоматов и обобщенные в теории алгоритмов. Тьюринг исходит из представления о разумном поведении как деятельности, направленной на решение задач. Задача полагается решаемой, если может быть обнаружен алгоритм — набор специфицируемых вычислительных процедур — ее решения. Понятие алгоритма было интуитивно ясным, но не существовало общей формулы алгоритма вообще. Тьюринг сформулировал следующий тезис: для всякого алгоритма можно смоделировать машину, отвечающую определенным характеристикам, которая будет реализацией этого алгоритма[20] . Тогда для всякой задачи (в широком, а не в узком, математическом, смысле), решаемой людьми, может найтись такая вычислительная машина, которая будет решать эту задачу так же хорошо. Машина Тьюринга – абстрактная универсальная вычислительная машина. Если такая машина может имитировать поведение любой другой машины, то она, в таком случае, сможет быть универсальным имитатором человеческого поведения. И, по мнению Тьюринга, нет никаких логических препятствий к допущению такой модели.

Машина Тьюринга отличается определенными свойствами. В основе ее лежит понятие автомата: самостоятельно действующего управляющего устройства. Для их описания используются три алфавита: алфавит входа, алфавит выхода и алфавит внутренних состояний автомата. Среди таких автоматов различают автоматы с конечной или бесконечной памятью, различаются они, разумеется, и количеством входов и выходов, а также могут быть вероятностными, если какая-нибудь из функций четко не задана, а предполагается случайно осуществляемой автоматом в каждый момент времени. Существенная характеристика машины Тьюринга – дискретность: в каждый момент дискретного времени она находится в совершенно определенном (одном и только одном состоянии), так что можно точно указать, что у машины «на входе» (иначе говоря, что «воспринимается» машиной), что «на выходе» (машинное «действие») и в каком состоянии она находится в данный момент времени. Таким образом, каждое дискретное состояние можно полностью описать в терминах входа-выхода и функции (иначе: алгоритма), которая(-ый) их связывает, т.е. используя только буквы соответствующих трех алфавитов плюс специальные термины, подобные логическим константам. Возможны и существуют машины с не дискретными состояниями: в этом случае никакому моменту дискретного времени нельзя сопоставить одно и только одно машинное состояние – данные продолжают поступать на вход постоянно и постоянно же происходит изменение состояния. В какой-либо момент времени могут быть «считаны» результаты неких измерений или действий, производимых с помощью такой машины, но внутри интервала между началом решения задачи этой машиной и получением результата различение дискретных состояний невозможно (или практически возможно только в виде какой-либо аппроксимации). Результаты, получаемые с помощью таких машин обладают большей погрешностью вследствие того, что считывание результатов – тоже процесс, требующий времени, а за это время показания изменяются. Машинам с дискретными состояниями соответствуют среди реальных машин, например, цифровые компьютеры, с не дискретными состояниями – аналоговые. В этом отношении на роль универсального «мыслящего» имитатора, построенного по модели Тьюринга, лучше подходят цифровые вычислительные машины. Трудно сказать, насколько уместно проводить аналогию между человеческим организмом и машиной Тьюринга в структурном отношении: для этого требуется доказать, что ментальные состояния подобны внутренним состояниям таких машин, а именно – дискретны. Трудности в проведении такой аналогии, однако, с точки зрения Тьюринга – не помеха компьютерному моделированию сознания: ведь его критерий основан на понятии имитации – если машина Тьюринга (с дискретными состояниями) способна имитировать поведение любой машины с не дискретными состояниями (а по Тьюрингу, это – так), т.е. решать всю совокупность задач, решаемых такими машинами, то нет разницы в том, насколько обоснованно полагать внутренние состояния человека дискретными. Вывод о способности машины Тьюринга мыслить как человек (т.е. решать весь комплекс релевантных задачи) не будет зависеть от успехов или неудач такого обоснования. Тогда, независимо от того, как решается метафизический вопрос (что такое ментальное), мышление может описываться в терминах машинной модели Тьюринга.

Не все, но многие версии машинного функционализма опираются на понятие машин Тьюринга, которые определяются по двум функциям: от входных данных и состояний к выходным данным и от входных данных и состояний к состояниям. В этой модели любое устройство, соответствующее понятию машины Тьюринга, можно описать с помощью так называемой машинной таблицы. Пример такой таблицы для примитивного автомата – выключатель:

|  |  |  |
| --- | --- | --- |
|  | Состояние S 1  | Состояние S 2  |
| Вход: нажатие  | Вкл. и переход в S 2 | Выкл. и переход в S 1 |
| Вход: нет нажатия  | Выкл. и остается в S 1 | Вкл. и остается в S 2 |

Любая система, имеющая набор входов и выходов, а также состояний, соотносящихся согласно машинной таблице, описывается машинной таблицей и является реализацией абстрактного конечного автомата. Простая версия машинного функционализма настаивает на том, что каждая система, имеющая ментальные состояния, описывается, по крайней мере, одной машиной Тьюринга определенного вида; она также утверждает, что каждый тип ментальных состояний системы тождественен одному из табличных машинных состояний, определенных машинной таблицей.

**2. Критика «машинного функционализма»**

Патнэм одно время разделял «машинную версию» функционализма; позднее, однако, он стал одним из энергичнейших критиков этой идеи. Функционализм (по крайней мере, в этом варианте) предлагает интерпретацию познавательного процесса как процесса обработки информации — формальной манипуляции репрезентациями или символами согласно некоторым правилам. Согласно этой теории, сознание аналогично вычислительной машине, и поэтому для изучения познания релевантны организация и структура, а не материальный субстрат сознания. Аргумент Патнэма здесь был основан на интуиции о том, что одно и то же ментальное состояние может быть реализовано на различных материальных носителях. Его позднейшие ревизии функционализма связаны с расширением требования о множественности реализаций ментального — в духе свойственного прагматистам плюрализма. Это расширение основано на инверсии исходной интуиции: если различные состояния вычислительной машины или машины Тьюринга могут реализовывать одно и то же ментальное состояние, то сознание не может быть идентифицировано ни с какой конкретной вычислительной машиной. Поддержка Патнэмом сильной формы множественной реализуемости ментального приводит его к пересмотру самого статуса психофизической проблемы как ключевой для философии сознания.

Главным образом критика компьютерных аналогий сосредоточилась на критике моделирования в терминах машин Тьюринга. Основной аргумент в этом случае состоит в том, что психологические состояния просто не тождественны машинным состояниям. Патнэм дает следующую этого экспозицию аргумента[21] . Состояние машины Тьюринга описывается таким образом, что эта машина может быть только в одном состоянии в один момент времени. Между тем, психологическое состояние человека допускает (и даже требует), чтобы субъект мог одновременно, например, испытывать боль, произносить число три и слышать, как скулит собака. Здесь мы имеем три вида входных данных, три табличных состояния и три диспозиции. Идея психологического состояния как машинного состояния требует, между тем, чтобы все эти три диспозиции детерминировались одной машинной таблицей. Более того, каждое теперешнее состояние субъекта должно быть таким, согласно машинной теории, чтобы оно детерминировало его следующее состояние. Но даже относительно аналогии с вероятностным автоматом субъект имеет слишком широкую вариацию диспозиций, каждая из которых может быть реализована в следующем психологическом состоянии (исключая детерминацию со стороны окружающей среды), чтобы можно было рассчитывать, что это следующее состояние полностью детерминировано предыдущим.

Другие аргументы против машинной версии функционализма таковы[22] .

Машинная версия не позволяет провести различий между диспозиционными состояниями организма – полаганиями, желаниями, склонностями и тому подобными – и его событийными состояниями. Можно сформулировать обеспечивающие такое различие условия тождества для событийных и диспозиционных состояний: два организма находятся в тождественных психологических событийных состояниях, если и только если они находятся в одном и том же машинном табличном состоянии; и организм находится в определенном диспозиционном состоянии, если и только если его машинная таблица содержит определенное машинное табличное состояние (для каждой диспозиции свое). Но каждое машинное табличное состояние организма таково, что организм может находится в нем в определенное время. Следовательно, для организма находиться в некотором диспозиционном состоянии значит быть способным для этого организма находиться в некотором событийном состоянии, которое должно соответствовать данному диспозиционному состоянию. Но какое событийное состояние соответствует диспозиции, например, «говорит по-французски»? А диспозиции «Говорит по-французски в данный момент»? А какое событийное состояние соответствует диспозиции «полагает, что р» или «думает сейчас, что р»? Эту проблему можно представить в более общем виде: любой событийный предикат, даже такой, как «думает сейчас…» или «испытывает (сейчас) боль», может быть интерпретирован как диспозиционный, если он может быть проанализирован в терминах других событий. Так, думать о летающих сосисках мы можем, не веря в их существование: в этом смысле затруднительно ставить в соответствие событийное состояние, выражаемое как «думаю (сейчас) о летающих сосисках» диспозиционному состоянию, выражаемому как «полагаю, что сосиски летают».

Поведение может быть продуктом серии психологических состояний, которые сами по себе не имеют «выхода» в форме внешнего поведения и машинная версия показывает, как это может быть. Но, с другой стороны, поведение может быть результатом взаимодействия одновременных психологических состояний: оно может быть функцией от того, что индивид ощущает и что он думает в данный момент. Машинная версия не в состоянии показать, как это возможно.

Машинная версия предполагает слишком сильные условия тождества психологических состояний. Так, если два человека различаются в отношении (функциональных характеристик) боли только тем, что наиболее вероятная реакция одного на боль от того, что ему наступили на ногу – крикнуть «ой!», а другого – «ай!», то из этого следует, что имеют место два различных типа боли, что мало правдоподобно. Этот аргумент допускает итерацию следующего вида: пусть имеет место такой автомат, что его машинные табличные состояния х и у различаются своими типами, если непосредственно следующие за ними табличные машинные состояния различаются своими типами. Но следующие состояния будут различаться лишь в случае, если непосредственно следующие за ними состояния различаются по типам, и т.д. В таком случае, если допускается, что каждое состояние связано с каждым другим компьютационным образом, любые два автомата, имеющие менее, чем все их состояния в качестве общих, не будут иметь ни одно из их состояний в качестве общего. Поэтому условия тождества для машинных табличных состояний не могут быть условиями тождества для психологических состояний.

Набор состояний, конституирующих машинную таблицу пробабилистского (как и детерминистского) автомата, представляет собой по определению список, тогда как набор ментальных состояний, по крайней мере, некоторых организмов, согласно эмпирическим фактам, заранее не дан. Абстрагируясь от физических ограничений (таких как ограниченность памяти и смерть), можно утверждать, что существует бесконечно много различающихся своими типами номологически возможных психологических состояний любой данной личности.

Даже если бы можно было достичь соответствия (типа один к одному) между машинными табличными состояниями и психологическими состояниями, эти соответствия с необходимостью не репрезентировали бы существенные свойства психологических состояний. Кажется достаточно очевидным, что существуют структурные сходства между, по крайней мере, некоторыми психологическими состояниями: например, между состоянием полагания, что р, и состоянием полагания, что р & q . Представление психологических состояний в виде списка просто не в состоянии репрезентировать такого рода структурные отношения.

Блок и Фодор, правда, предлагают различать между машинными табличными и компьютационными состояниями автоматов. Первые – это состояния, определяемые колонками машинной таблицы; последние – любые состояния машины, характеризуемые в терминах ее входных, выходных данных и/или машинных табличных состояний: предикаты «осуществил вычисление, включающее триста семьдесят два машинных табличных состояния» или «доказал последнюю теорему Ферма», или «напечатал энный символ своего словаря выходных данных» – все обозначают возможные компьютационные состояния машин. Относительно такого различия очевидно, что даже если аргумент показывает, что психологические состояния организма не могут быть поставлены в соответствие машинным табличным состояниям автомата, он еще совершенно не обязательно показывает, что они не могут быть поставлены в соответствие компьютационным состояниям автомата. Однако, условием возможности соответствий последнего рода является (по мнению Блока и Фодора) исчисляемость психологических состояний организма.

**3. «Машинное» решение проблемы языка мысли**

Как возможен частный язык мысли? Если допустить, что аргумент от бесконечного регресса, опирающийся на положение, что все языки, которыми мы владеем, приобретены нами путем изучения, устраняется принятием языка мысли как врожденного, то возможно другое возражение, снова апеллирующее к такому следствию, как бесконечный регресс. Если понимание предиката есть репрезентация его объема в некоем языке, который индивид уже понимает, то не должно ли понимание предикатов этого языка подразумевать репрезентацию условий его истинности в некоем мета-метаязыке, предварительно понятом, и т.д. (до бесконечности)? Стандартный ответ состоит в признании того факта, что изучение значения предиката включает репрезентацию его объема, и отказе признавать, что этого же требует понимание предиката. Достаточным условием последнего может быть просто то, что употребление этого предиката субъектом всегда в действительности соответствует правилу истинности для него. Не обязательно, с этой точки зрения, чтобы субъект следования правилу был также и субъектом знания этого правила. В качестве иллюстрации этого ответа предлагается реальный компьютер: он использует, по меньшей мере, два языка – язык входных и выходных данных, посредством которого он осуществляет коммуникацию с внешним миром, и машинный язык, на котором он осуществляет свои вычислительные операции. Реальный компьютер нуждается в «компиляторах», которые определяли бы (истинностные) значения формул входа-выхода в терминах машинного языка. Но такой компьютер, утверждают защитники этого взгляда, не нуждается в подобных компиляторах для машинного языка. «Тем фактом, что машина построена для использования машинного языка, избегается бесконечный регресс для компьютеров»[23] . Формулы машинного языка в этом случае понимаются как непосредственно соответствующие компьютационно релевантным физическим состояниям и операциям машины: физика машины, таким образом, гарантирует, что последовательности состояний и операций, которые она проходит, выполняя свои вычисления, выполняют семантические условия для формул его внутреннего языка. Инженерные принципы занимают в такой концепции место определений истинности для формул машинного языка и как такие они обеспечивают и соблюдение этих «определений».

Считается, что Виттгенштейн доказал, что не может быть такой вещи, как частный язык[24] . На это когнитивист может использовать аргумент, который воспроизводит Дж. Фодор, говоря, что «что бы ни доказал Виттгенштейн, не может быть, чтобы было невозможно, чтобы язык был частным в том смысле, в каком частным является машинный язык компьютера, поскольку существуют такие вещи, как компьютеры, а то, что существует в действительности, возможно»[25] . Это, несомненно, верно постольку, поскольку вообще верна машинная аналогия. С другой стороны, неадекватность машинной аналогии сознания в целом может еще не предполагать, что столь же неадекватна и аналогия машинного языка. Виттгенштейн характеризует индивидуальный язык двумя способами: или как язык, термины которого указывают на вещи, опыт которых может иметь только говорящий на этом языке, или как язык, для применения терминов которого не существует никаких публичных критериев, правил или конвенций. Внутренняя репрезентативная система является частным языком, по меньшей мере, во втором смысле: употребление его терминов не регулируется никакими публичными конвенциями, хотя вовсе не обязательно, чтобы референтами этих терминов были исключительно приватные события (ощущения). Ответ когнитивиста опять может состоять в отказе считать аргумент от невозможности частного языка (по крайней мере, в том виде, который приписывается Витгенштейну) возражением против теории, допускающей ментализ. Утверждается, что нет причин, почему защитник такой теории обязан полагать, что ментальные операции демонстрируют эпистемическую приватность в каком-либо строгом смысле этого понятия. Напротив, для него лучше не принимать этого, если он хочет, чтобы его психологические теории были совместимы с материалистической онтологией – ведь нейрофизиологические события публичны. Далее, утверждается, что самое большее, что этот аргумент показывает, это – что если нет публичных процедур для сообщения о том, когерентно ли употреблен термин, то нет способа знать, когерентно ли он употреблен. Но из этого не следует, что в таком случае в действительности не было бы различия между когерентным и произвольным применением термина; a fortiori из этого не следует, что нет никакого смысла утверждать, что в этом случае есть различие между когерентным и произвольным применением термина. Употребление языка для компьютерных операций не требует, чтобы употребляющий его был обязан иметь способность определять, что термины этого языка применены совместимым образом. Наконец, менталисту требуется показать, в каком смысле термины во внутренней репрезентативной системе употребляются когерентно, и что они в этом смысле разумно аналогичны терминам публичного языка (насколько те могут когерентно употребляться). Если мы не можем сделать первого, то, вероятно, само понятие языка мысли не когерентно; если не можем сделать второго, то бессмысленно называть язык мысли языком. В публичных языках когерентность употреблений терминов контролируется конвенциями, которые ставят термины в соответствие определенным парадигмальным публичным ситуациям (например, определенным перцептивным данным – таким, как появление коровы в поле зрения). Использовать термин когерентно в этом смысле значит применить его к ситуации, подпадающей под спецификацию парадигмальной, согласно конвенции, для данного термина (или, иначе, относящейся к классу ситуаций, относительно которых данный термин истинен). Но даже в случае публичных языков, — утверждает один из самых последовательных защитников машинной аналогии языка мысли Дж. Фодор, — когерентность (употребления терминов) не требует устойчивого отношения между способом употребления терминов и тем, каков мир (выражаемого в понятии «конвенции»). Что действительно требуется, так это – устойчивое отношение между тем, как термины употребляются и тем, каким мир полагается субъектом их употребления[26] . Как такое отношение может быть обусловлено чем-то, кроме публичных конвенций? Врожденная структура нервной системы выполняет, согласно этому подходу, данную функцию для внутреннего репрезентативного кода. Когда, говорит Фодор, мы размышляем об организме как о компьютере, мы пытаемся поставить в соответствие физическим состояниям организма (например, состояниям его нервной системы) формулы словаря психологической теории. В идеале такое соотнесение должно выполняться так, чтобы, по крайней мере, некоторые из последовательностей состояний, каузально имплицированных в производство поведения, могли интерпретироваться как компьютерные операции, имеющие соответствующие описания поведения в качестве своих «последних строк» (в записи – описании последовательности). В случае организмов, так же, как и в случае реальных компьютеров, если у нас есть правильный способ устанавливать соответствия между формулами и состояниями, то мы сможем интерпретировать последовательности событий, вызывающих поведение на выходе, как компьютационную деривацию событий «на выходе». Все, что требуется для, того, чтобы внутренней репрезентативной системе можно было приписывать пропозициональные установки, и, соответственно, чтобы относительно нее можно было утверждать, что она использует язык, – это чтобы было соответствие правильного вида между пропозициональными установками системы и ее отношениями к формулам данного языка[27] .

Остается, конечно, открытым вопросом: достаточно ли так истолкованная внутренняя репрезентация похожа на репрезентацию в естественном языке, чтобы обе можно было называть репрезентациями в одном и том же смысле? Но во всяком случае есть аналогия между двумя видами репрезентации. Поскольку публичные языки конвенциональны, а язык мысли нет, трудно ожидать, чтобы между ними было что-то большее, чем аналогия. Фодор пишет в этой связи: «Если вы впечатлены этой аналогией, то вы захотите сказать, что внутренний код является языком. Если вы не впечатлены аналогией, вы захотите сказать, что внутренний код является в некотором смысле репрезентативной системой, но не языком. Но ни один ответ не повлияет на то, что я полагаю серьезным вопросом: являются ли когерентными методологические допущения компьютационной психологии? Ничто … не говорит в пользу того, что они таковыми не являются»[28] .

[20] А. Тьюринг, Может ли машина мыслить, «Колледж», Саратов, 1999, 6 – 76.

[21] H. Putnam, ‘Philosophy and Our Mental Life’, Mind, Language, and Reality: Philosophical Papers, vol. 2, London, Cambridge University Press, 1975, 291 – 303.

[22] N. Block and J. Fodor, ‘What Psychological States Are Not?’, Philosophical Review 81, no. 2, 1972, 159 – 181.

[23] Там же , 66.

[24] См .: L. Wittgenstein, Philosophical Investigations, N. Y. Prentice-Hall Publishing Co., 1958, 258 и далее .

[25] J. Fodor, The Language of Thought, N. Y. Crowell, 1975, 68.

[26] Там же, 71.

[27] Там же, 74 – 77. Аналогия между индивидуальной и публичной репрезентациями, которую предлагает Фодор, такова: если ‘ a есть F ’ – формула в публичном языке, то ( S употребляет ‘ a есть F ’ для репрезентации обладания а свойством F ) только тогда, когда ( S полагает, что a есть F только тогда, когда S соглашается с ‘ a есть F ’). Так как обычно то, что ставит в соответствие полагание S , что a есть F его согласию с ‘ a есть F ’ в случае публичных языков – подчиненность S конвенциям такого языка, можно заместить сформулированное условие таким условием: ( S употребляет ‘ a есть F ’ для репрезентации обладания а свойством F ) только тогда, когда (( S полагает, что a есть F только тогда, когда S соглашается с ‘ a есть F ’) конвенционально). Если ‘ a есть F ’ – формула внутреннего кода, то «соглашается с» в последнем условии заменяется последовательностью, состоящей из одного и большего числа базисных отношений, из которых сконструированы компьютационные отношения к формулам внутреннего кода, и «конвенционально» заменяется в нем на «номологически необходимо».

[28] Там же , 78 – 79.