Выдержал ли аргумент китайской комнаты Сёрла проверку временем? «Круглый стол» в МГУ

И.В. СОЛНЦЕВ

Мысленный эксперимент Дж. Сёрла считается одной из самых плодотворных интуиций в современной философии, которая оказала значительное влияние не только на философов, но и на многих ученых, работающих в сфере искусственного интеллекта. Не удивительно, что через 30 лет после публикации статьи Сёрла «Сознание, мозг и программы»[1] на философском факультете МГУ было решено провести тематический «круглый стол», в котором приняли участие как именитые знатоки философии, так и обычные студенты.

Прежде чем приступить к изложению дискуссии, следует кратко изложить сам аргумент китайской комнаты. Американский философ пытался доказать, что даже если удастся создать компьютер, который будет обладать искусственным интеллектом (ИИ), то этот компьютер не будет обладать пониманием. Устройство обладает искусственным интеллектом, если оно может проходить классический тест Тьюринга[2]. Но подобное устройство, утверждает Сёрл, обладает ИИ в «слабом» смысле. То есть хотя его ответы неотличимы от человеческих, у него отсутствует понимание, что, по определению, равнозначно отсутствию у него искусственного интеллекта в «сильном» смысле.

Теперь доказательство, т.е. сам мысленный эксперимент Сёрла. Предположим, что Сёрл не знает китайского языка. Поместим его в так называемую китайскую комнату. В китайской комнате у Сёрла будет

[1] *Searle J.* Minds, brains, and programs // Behavioral and Brain Sciences. Vol. 3 (3). P. 417–445.

[2] *Turing A.* Computing machinery and intelligence // Mind. New Series. Vol. 59 (236). P. 433–460.

книга, где записаны различные правила сопоставления одних китайских значков другим. Сёрл утверждает, что в таком сопоставлении и заключается работа компьютера: он синтаксичен. Пусть правила в книге будут настолько замечательными, что если Сёрлу в комнату будут передаваться вопросы на китайском, то он с помощью книги получит новое словосочетание, которое и будет «человеческим» ответом на вопрос носителя китайского языка. Тогда Сёрл пройдет тест Тьюринга. Будет ли Сёрл понимать что-либо? Он утверждает, что нет. Получается, что машина, проходящая тест Тьюринга, не будет обладать пониманием.

Теперь вернемся к «круглому столу». Приведем мнения его участников по центральным вопросам: плодотворен ли мысленный эксперимент Сёрла? Нет ли в нем изъянов?

Итак, имеет ли он познавательную ценность? Утвердительно на эти вопросы ответили практически все участники «круглого стола», кроме **Д.И. Дубровского** (ИФ РАН). **Н.С. Юлина** (ИФ РАН) считает, что признаком плодотворности аргумента китайской комнаты является невероятно возросший интерес к философии сознания за последние 30 лет. **Н.А. Воронов** (МГУ), **Д.В. Волков** (Московский центр исследований сознания), **Н.Ю. Клюева** высказали мнение, что мысленный эксперимент Сёрла позволяет отграничить феноменальный и когнитивный аспекты сознания, что, по их словам, несколько «отрезвило» специалистов в области искусственного интеллекта[3]. Если и возможно создание робота, который будет имитировать наше поведение, то из этого еще не следует, что этот робот будет обладать ощущениями, субъективными переживаниями. **А.В. Потапов** (МГУ) даже считает, что чалмерсовский аргумент от представимости зомби сводится к аргументу китайской комнаты. Как мы увидим, скорее именно аргумент китайской комнаты, если он имеет какой-то смысл, сводится к доводу от представимости зомби.

В.В. Васильев (МГУ) видит плодотворность мысленного эксперимента в том, что он не только ставит вопросы, но и дает ответы на них. Будет ли компьютер, симулирующий человеческое поведение, с необходимостью обладать сознанием? В мысленном эксперименте дается ответ: нет, не будет. Но, как пытается показать Сёрл, ответы человека на подобные вопросы с необходимостью будут сопровождаться их пониманием. Эта черта китайской комнаты позволила Васильеву по плодотворности сравнить ее с зеноновской черепахой.

Но нам интереснее, конечно же, что думали участники «стола» по поводу того, нет ли в аргументации Сёрла изъянов? Действительно ли китайская комната доказывает то, что она была призвана доказать? **Д.В. Иванов** (МГУ) считает, что китайская комната убедительно показывает неверность машинного функционализма, т.е. понимания

[3] Хотя, по мнению Д.В. Волкова, эксперимент не очень актуален, так как до создания программы, эмитирующей человеческое речевое поведение, еще очень далеко.

сознания в терминах машины Тьюринга. Но он полагает, что доводы Сёрла не могут опровергнуть функционализм в более широком смысле. Ведь китайская комната синтаксична, у нее нет семантики, она не отсылает к предметам вовне. Если бы у нее появилась семантика, т.е. китайская комната превратилась бы в китайского робота, и у нее, возможно, появилось бы понимание. **Д.В. Волков** придерживается схожей позиции и добавляет, что если китайская комната будет как-то взаимодействовать с внешней средой, то, став китайским роботом, она получит семантику и понимание, даже если робот унаследует свою программу в неизменном виде от китайской комнаты.

Следует упомянуть так называемый **системный ответ** на аргумент Сёрла, с которым солидаризовались **М.Н. Белянин** (МГУ) и отчасти Д.В. Иванов. «Системщики» настаивают, что хотя Сёрл и не будет понимать китайского, китайская комната в целом будет понимать китайский. На это американский философ отвечает, что ничего не изменится, если вся книга с правилами будет им выучена и он таким образом «овнутрит» комнату. Тогда он сможет отвечать на вопросы где угодно, хоть на открытом воздухе, хоть на Марсе. Этот ответ Сёрла кажется вполне убедительным.

Е.В. Косилова (МГУ) подошла к проблемам, которые поднимаются китайской комнатой, с неожиданной стороны. Косилова считает, что создание даже слабой версии ИИ достаточно сомнительно, а это является центральной предпосылкой в доказательстве Сёрла. Ведь компьютер, машина Тьюринга, не может отличить интересные, плодотворные результаты от тривиальных. На наш взгляд, для того чтобы подобные доводы проходили, должны выполняться два условия: во-первых, нужно строго доказать, что в какой-то области компьютер ущербен; во-вторых, показать, что человек в отличие от компьютера справляется с задачами, которые не может решить компьютер. Так, сам Тьюринг доказал, что не существует общего решения всех математических задач, принадлежащих к некоторому, но вполне определенному классу (гильбертовская проблема алгоритмической разрешимости). Но ведь никто не доказал, что люди могут решить все эти задачи. В то же время даже если у нас есть серьезные основания утверждать, что человек всегда может отличить тривиальный результат от нетривиального, у нас нет строгого доказательства того, что то же самое не может сделать компьютер.

Д.И. Дубровский считает, что для того чтобы аргумент проходил, в книге «узника» китайской комнаты должны быть все семантические, прагматические и контекстуальные правила манипуляции с символами, тогда, возможно, у китайской комнаты и возникло бы понимание.

В целом позиция Дубровского, по его же словам, близка позиции Васильева. Васильев замечает, что китайская комната не может пройти тест Тьюринга в строгом смысле. Почему? Дело в том, что китайская комната в принципе не может отвечать на индексикальные вопросы, например на вопрос «сколько сейчас времени?». Со-

ставитель чудо-книги никак не мог узнать, когда узнику китайской комнаты зададут этот вопрос. Конечно, ответы на подобные вопросы могут определяться случайно. Но если машина, случайно определяющая ответы на подобные вопросы, будет проходить тест Тьюринга, то тест Тьюринга, очевидно, не является критерием наличия интеллекта.

Почему же тогда этот довод, достаточно очевидный, не высказывался ранее? Васильев видит две причины. Во-первых, Сёрл убедил философское сообщество в тождестве китайской комнаты, не имеющей «окон», и китайского робота, который взаимодействует с окружающей средой. А это различие принципиально, так как китайский робот в отличие от китайской комнаты будет проходить тест Тьюринга. Во-вторых, возражения Сёрлу исходили в основном от когнитивистов, которые были убеждены в возможности создания машины, которая проходила бы тест Тьюринга.

В то же время Васильев добавляет, что аргумент китайской комнаты показывает безальтернативность сильной версии искусственного интеллекта. Аргументы Васильева, доказывающие это положение, изложены в его книге «Трудная проблема сознания»[4], и мы не будем подробно на них останавливаться.

Я почти полностью разделяю точку зрения Васильева[5] и хотел бы привести дополнительные доводы в ее пользу[6].

Действительно, китайская комната не сможет правильно отвечать на вопросы, касающиеся внешних объектов, на которые нельзя ответить без наличия некоторой семантики. Значит, как показал Васильев, китайская комната демонстрирует ущербный интеллект, а не искусственный. Какого понимания тогда мы от нее ждем?

Однако на доводы Васильева о принципиальной семантичности индексикальных вопросов Волков в конце обсуждения привел контраргумент. По его мнению, китайская комната будет всегда ограниченна. Сёрл может не иметь часов, и тогда на вопрос о времени он просто ответит: «не знаю». Как и не сможет Сёрл сказать, что происходит вне китайской комнаты. На это автор данного обзора заметил, что возражение Волкова было бы верно, если бы китайская комната в принципе могла «овнутрить» все возможные явления.

Однако дело в том, что некоторые явления принципиально нельзя «овнутрить» с помощью чисто синтаксической системы, которая не отсылает к внешним явлениям. То есть наша комната никогда не сможет ответить на вопрос о текущем времени, не отсылая к каким-нибудь механическим часам колебаниям кварцевого осциллятора (на

[4] *Васильев В.В.* Трудная проблема сознания. М., 2009. Заметим, что они не были озвучены на «круглом столе».

[5] За исключением тезиса о безальтернативности «сильного ИИ».

[6] Часть из них была высказана мной Васильеву и Волкову.

нем основан механизм работы компьютерных часов) и т.п. А Сёрл, видя в своей чудо-книге, что необходимо брать сведения о состоянии механических часов или о любых других временных явлениях, которые не может описать никакая статичная синтаксическая система, получит возможность сопоставить их с определенными иероглифами и таким образом получить понимание. Ведь наличие семантики, отнесения к объектам вовне системы, и есть необходимое и достаточное условие наличия понимания по Сёрлу.

Но доводы Васильева можно заострить еще сильнее. Пусть действительно Сёрла поместят в темную китайскую бочку, где у него не будет ни часов, ни других измерителей периодических, изменяющихся процессов, которые не могут быть формализованы чисто синтаксически. В таком случае на все эти каверзные вопросы он будет отвечать просто: «Не знаю». Тогда пусть ему зададут следующий вопрос: «Какой предыдущий вопрос я задал?» Как Сёрл может дать правильный ответ? Существует только один способ: вопросы будут записываться в отдельную вопросительную книгу. Иероглифы в чудо-книге будут отсылаться к этой вопросительной книге. В таком случае Сёрлу без труда удастся понять, что определенные иероглифы имеют определенные значения: в нашем случае Сёрл может распознать китайские иероглифы, отвечающие за обозначение последовательности (порядковые числительные и т.п.). Поэтому китайская комната принципиально отличается от китайского робота и изначально ущербна.

Подведем итоги. Несмотря на то что участники «круглого стола» не пришли к единомыслию, я дерзну предположить, что загадка китайской комнаты была разгадана. Китайская комната из-за своей синтаксической ограниченности не может давать правильные ответы на элементарные вопросы: хотя бы на вопросы о том, о чем ее спрашивали ранее. Если же она выйдет за пределы синтаксиса, то у нас не останется оснований для того, чтобы отрицать наличие у нее понимания, на чем настаивает Сёрл. Здесь можно настаивать, что настоящее понимание такая система все равно не получит, так как сопоставления символов друг с другом и отсылки программы к внешней среде еще не достаточно для того, чтобы появилось субъективное ощущение понимания. В таком случае аргумент китайской комнаты сведется к одному из эпистемических доводов (так их предложил называть Д. Чалмерс), которые отсылают к нашему субъективному опыту, неподдающемуся описанию на физико-функциональном языке.