УДК 165.43

В. В. Целищев

Институт философии и права СО РАН
ул. Николаева, 8, Новосибирск, 630090, Россия
E-mail: director@philosophy.nsc.ru

ПРЕДСТАВЛЕНИЕ ЗНАНИЯ В СВЕТЕ «ВТОРОГО АРГУМЕНТА ПЕНРОУЗА» *

При формализации знания особый интерес имеет вопрос, обратный строгой проблеме представления знания, а именно, может ли неформальная концепция знания «понять знание» компьютера. Такая постановка вопроса в литературе получила название «второго аргумента компьютера». В данной работе исследована логическая структура этого аргумента.

Ключевые слова: формализация знания, алгоритм, геделевское предложение, второй аргумент Пенроуза.

Как следует понимать формальную систему, симулирующую ум?

Аргументация Пенроуза представляет собой сложную смесь различных дискурсов. Естественно, это сопровождается неоднозначностью трактовки многих ключевых терминов. Так, МакКаллох отмечает, что неоднозначно употребление термина «формальная система Ф» [McCullough, 1995]. Казалось бы, этот математический термин не подлежит никакому двусмысленному толкованию. Но дело как раз в том, что при сопоставлении машины и человека, точнее при обсуждении тезиса о том, что человеческий ум эквивалентен некоторой формальной системе, в содержание этого термина вкладывается многое из того, что присуще человеческому уму, а не просто математическому понятию. Пенроуз в своей аргументации подразумевает по крайней мере три интерпретации термина «формальная система Ф».

Одна интерпретация Ф состоит в том, что Ф представляет врожденную (внутренне присущую) способность к размышлению самого математика. Это чисто иннативистская характеристика математического мышления, и именно к ней относится уверенность в истинности некоторых математических утверждений, которые «истинны неопровержимо». Эта интерпретация позволяет объявить о внутренней непротиворечивости человеческого ума. Другая интерпретация включает в себя, помимо врожденной способности к математическим заключениям с математической определенностью, еще и эмпирический опыт математика, со всеми эвристическими и индуктивными приемами получения знания. Третья возможность состоит в том, что Ф представляет пределы того, что могло бы быть известно математику, независимо от того, как приобретается это знание, через размышление или эмпирически. Различия между этими интерпретациями приобретают значение, когда решается вопрос о том, знает ли математик, что его мышление описывается посредством Ф. Формальная система Ф является дедуктивной структурой, и полученные математиком знания о роли Ф не отражены в самой Ф. Пенроуз отдает себе отчет в слабости своей аргументации в этом отношении и поэтому вводит в рассмотрение так называемый «новый аргумент». В нем рассматривается новая система Ф*, которая включает в себя Ф плюс все, что следует из информации, что способности мышления были описаны в Ф. Этот новый аргумент более тонок, и его не

* Исследования, нашедшие отражение в этой работе, финансово поддержаны Междисциплинарным интеграционным проектом фундаментальных исследований РАН № 47.

так просто опровергнуть. Он будет рассмотрен позднее.

Следствием неоднозначности термина «формальная система» является безапелляционное утверждение Пенроуза о том, что неопровержимо истинным считается утверждение об обоснованности математического мышления. Потому что «неопровержимая истинность» есть истинность иннативистского толка. Другими словами, следует полагать, что математическое мышление человека обосновано. Таким образом, мы впадаем в порочный круг, когда обоснованность объясняется в терминах внутренней непротиворечивости человеческого ума, а непротиворечивость, будучи эквивалентной обоснованности, объясняется в терминах обоснованности математического мышления. Именно это обстоятельство, судя по всему, явилось причиной появления у Геделя фразы «доказательство с математической определенностью», ибо чем еще может быть математическая определенность как не уверенностью во внутренней непротиворечивости человеческого ума.

При обсуждении вопроса об обоснованности математического мышления поднимается труднейший эпистемологический вопрос об ошибочности мышления и его причинах. Человек может по тем или иным причинам полагать некоторое утверждение «неопровержимым», когда оно по ряду критериев таковым не является. МакКаллох полагает, что это вопрос даже не эпистемологический, а психологический.

«Поэтому аргумент Геделя не доказывает, что человеческое мышление должно быть невычислимым – она лишь доказывает, что если человеческое мышление вычислимо, тогда оно должно быть либо необоснованным, или же для нас принципиально (inherently) невозможно знать одновременно, каковы наши мыслительные способности и являются ли они обоснованными. Пенроуз отметает возможность того, что мы знаем наши мыслительные силы, но не знаем, обоснованы ли они (раздел 3.2. «Тени разума»). Пенроуз говорит, что если мы знаем, что некоторая конкретная компьютерная программа Ф эквивалентна человеческому мышлению, тогда мы были бы вынуждены заключить, что Ф обоснована… Для меня этот вопрос является скорее делом психологии, чем математики. Пенроуз полагает некоторые свои веры относительно математики «неопровержимо истинными», и даже не рассматривает возможность того, что некоторые веры могут быть ошибочными. Если он придерживается этой веры, то вовсе не следует, что мышление Пенроуза не вычислимо. Отсюда следует лишь то, что Пенроуз никогда не был убежден в том, что оно вычислимо. Для людей вроде меня, кто имеет более гибкую позицию в отношении непогрешимости своей веры, то есть, в отношении того, что их мышление может быть неосновательным, аргумент Пенроуза не очень весом» [McCullough, 1995].

В конечном счете, мы действительно имеем полупсихологическую уверенность математика в истинности некоторого утверждения в качестве единственного критерия того, что это утверждение «неопровержимо истинно». Дело в том, что «неопровержимость» является весьма произвольным критерием. Например, могут существовать такие сложные истины, которые доказаны с математической определенностью, т. е. всеми доступными математику надежными средствами, но которые трудно считать интуитивно истинными. Будут ли такие истины неопровержимыми? Далее, вполне должен быть случай, что ничего ложного не может быть «неопровержимой истиной». Однако может быть показано, что даже если допустить, что ничего ложного не может быть признано «неопровержимой истиной», этот факт не может быть «неопровержимой истиной». Можно показать, что понятие «неопровержимой истины» не может само быть неопровержимым [Ibid.].

Пусть G будет предложением.

Это предложение не есть неопровержимая вера человека X.

Если мы предположим, что G есть одна из неопровержимых истин X, тогда мы немедленно заключаем, что она должна быть ложна. Следовательно, неопровержимые веры X включают в себя по крайней мере одно ложное утверждение. Оборачивая это рассуждение, получаем, что если веры X обоснованы (они не включают никаких ложных утверждений), тогда должно быть, что G не может быть одной из его неопровержимых вер. Но так как G говорит, что оно не является одной из его неопровержимых вер, G должно быть истинным. Поэтому мы заключаем:

Если X обоснован, тогда G истинно.

Теперь, так как X способен видеть истинность этой импликации, отсюда следует, что если он считает себя обоснованным, тогда он будет верить в G. Но по определению G, если X верит в G (неопровержимо), тогда G должно быть ложно. Поэтому если X верит в свою обоснованность, тогда G ложно, и все же X верит в его истинность. Следовательно, заключаем мы,

Если X верит в свою обоснованность,
тогда он необоснован.

Что и требовалось доказать.

Неопровержимость математических истин

Дискуссия вокруг «механизма» исходит из предположения, что в формальной системе имеется множество всех аналогов арифметических теорем и только их, которые известны неопровержимо или с математической определенностью. Это предложения, сформулированные на языке первого порядка. Предполагается, что идеализированный математик никогда не утверждает ложные математические предложения. Коль скоро речь идет об идеализированном математике, такие истинные предложения являются *познаваемыми* или доказуемыми арифметическими предложениями. В данном случае мы имеем две интерпретации одного и того же оператора формальной системы: доказуемость и познаваемость.

Проблема демаркации познаваемых математических истин и всех математических истин поднимает множество вопросов о природе математических утверждений. Особенно в этом отношении характерен крайний платонизм Геделя, с точки зрения которого существует сфера платонистских истин, которые превосходят все, что в принципе может быть известно человеку. Сам Гедель в этом контексте говорит об «объективной математике» и «субъективной математике» [Godel, 1995. P. 304–323]. Каково же соотношение этих двух математик? Ясно, что познаваемость математических истин связана с некоторыми эпистемологическими процедурами «доступа» к ним, и в первую очередь, это математическое доказательство. Для точного представления такого рода доступа изобретается формальное доказательство, которое, в свою очередь, является частью формальной аксиоматической системы. Ясно, что такая формальная система должна «схватывать» интуитивное содержание математических утверждений. Интуитивно мы полагаем, что содержательные математические утверждения не противоречат друг другу, и стало быть, адекватность формальной системы для представления содержательного математического знания заключается в непротиворечивости формальной системы.

Адекватность формальной системы означает, что интуитивно истинные предложения математики должны быть доказуемы, и доказуемые утверждения – истинными. Как известно, для достаточно богатых формальных систем арифметики такого рода адекватности добиться невозможно – согласно первой теореме Геделя о неполноте существуют интуитивно истинные, но не доказуемые в формальной системе утверждения. К таким утверждениям, согласно второй теореме Геделя о неполноте, относится и утверждение о непротиворечивости формальной системы. Таким образом, не все математические истины представимы доказуемыми утверждениями в формальной системе, и эта принципиальная неполнота формальных систем является важным обстоятельством.

Обозначим множество истин «субъективной» математики через К (далее мы будем называть их истинами «человеческой», в противоположность «платонистской» математики). Далее, пусть множество всех математических (платонистских) истин, выраженных в языке первого порядка, будет Т. Интуитивно, исходя уже из этимологии терминов, ясно, что $К \subseteq Т$. Рассмотрим гипотетическую возможность $К = Т$, которая означает, что все математические истины рано или поздно будут известны человеку. Согласно теореме Тарского, Т не определимо в языке арифметики. Это значит, что Т не является рекурсивно перечислимым множеством. Если множество К совпадает с множеством Т, тогда К также не является рекурсивно перечислимым. Но в отношении познаваемых истин это была бы странная позиция, которую можно было бы понять только следующим образом. Человек обладает такими когнитивными способностями, которые позволяют ему «превосходить» в своем познании множество рекурсивно перечислимых истин. Это означает, что чело-

век превосходит по своим возможностям чисто механическое накопление истин, которое и делает возможным рекурсивно перечислимый характер множества познаваемых утверждений. Таким образом, для сторонников «механизма», доктрины, согласно которой мышление человека успешно моделируется компьютером, или машиной Тьюринга, неверно, что К = Т. Тогда в множестве Т остаются такие истины, которые не познаваемы в принципе.

Новый аргумент Пенроуза

При сопоставлении когнитивных способностей человека и компьютера нужно исходить из некоторых математических утверждений, которые являются базисными для обеих категорий когнитивных созданий. Можно вести речь о базисе человеческой математики, реализованном в аксиоматике. Компьютер может рассматривать эти аксиомы как «неопровержимые» математические утверждения (они могут быть «сообщены» компьютеру в качестве таковых человеком). Больше того, программа компьютера может быть устроена так, что компьютер может полагать эти утверждения неопровержимо истинными согласно его собственным критериям. Снабдим, вслед за Пенроузом, такие предложения знаком ∇ [1].

Противопоставление человека и компьютера концентрируется вокруг тезиса, опирается ли математическое понимание на набор механизмов, скажем, некоторых механизмов Ф. В качестве таких механизмов можно указать алгоритмы. Предполагается, что если в случае человека такой набор механизмов представляется крайне проблематичным, то уж в случае компьютера (или робота, как у Пенроуза) «мышление» его управляется таким набором механизмов. Так называемый «новый» аргумент Пенроуза состоит в том, что даже для компьютера это неверно. Точнее, даже компьютер будет вынужден отвергнуть возможность того, что его математическое понимание опирается на набор механизмов М, независимо от того, как обстоит дело в действительности.

Мы имеем здесь весьма сильный аргумент против «механизма». В определенном отношении мы имеем дело с тактикой гамбита: сначала менталист дает преимущество «механизму», говоря о том, что компьютер обладает достаточным «сознанием» для того, чтобы иметь веру в неопровержимые математические утверждения. Но это преимущество «противника» быстро рассеивается, потому что Пенроуз показывает (или пытается показать), что компьютер может поверить в то, что управляется набором механизмов только ценой противоречия, что недопустимо для компьютера. Таким образом, «механизм» терпит двойное крушение, потому что даже при «послаблении» ему мы имеем противоречие.

Рассмотрим «новый» аргумент Пенроуза более тщательно, следуя изложению его автора [Пенроуз, 2003].

Пусть имеется гипотеза о том, что в основе понимания математических утверждений лежит некоторый набор механизмов Ф. Далее, пусть имеется некоторое Π_1-утверждение, являющееся следствием Ф. Если компьютер верит в это утверждение, оно является для него неопровержимым математическим утверждением. Назовем его ∇Ф-утверждением. Ясно, что «неопровержимость» этого утверждения зависит напрямую от принятия гипотезы, т. е. истинность ∇Ф-утверждения зависит от истинности гипотезы. Какую роль при этом играет гипотеза?

Утверждение, истинность которого зависит от гипотезы, не доказуемо в предполагаемой формальной системе, к которой добавлена гипотеза. Можно ли найти такие Π_1-предложения, которые являются следствиями гипотезы Ф и которые не являются обычными ∇Ф-утверждениями? Таким утверждением является истинное геделево предложение G. При этом важно упомянуть и тот факт, что предложение G не является теоремой первоначальной системы.

Истинность предложения G следует из обоснованности расширенной системы. Но эта обоснованность обязана уже гипотезе в том аспекте, что истинность конкретного Π_1-предложения, а именно, предложения G, есть следствие гипотезы Ф. Другими словами, истинность предложения G невозможно постичь без привлечения гипотезы Ф, поскольку компьютер убежден, что предложение G следует из гипотезы Ф. Таким образом, G есть ∇Ф-утверждение, а не просто

[1] В русском переводе «Теней разума» Р. Пенроуза используется знак пятиконечной звезды.

∇-утверждение, т. е. неопровержимое математическое утверждение.

Теперь рассмотрим новую формальную систему, базисом которой являются не неопровержимые математические утверждения типа ∇, которые являются причиной обоснованности соответствующей формальной системы, а ∇Ф-утверждения. Точно так же, как компьютер полагает, что обоснованная формальная система с неопровержимыми математическими утверждениями ∇ охватывает все его неопровержимые содержательные убеждения относительно истинности Π_1-предложений, он (компьютер) будет полагать, что новая формальная система (с гипотезой) охватывает все его неопровержимые убеждения относительно истинности Π_1-предложений, обусловленных гипотезой Ф.

Теперь построим геделево предложение G(Ф) в новой системе. Обоснованность этой системы есть следствие гипотезы Ф, и, в свою очередь, следствие обоснованности этой системы. Но тогда G(Ф) есть теорема новой формальной системы. Но это возможно, с точки зрения компьютера, только в том случае, если новая формальная система необоснованна, поскольку такое заключение противоречит первой теореме Геделя о неполноте. А необоснованной она может быть только за счет ложной гипотезы Ф. Другими словами, признание новой формальной системы необоснованной противоречит принятию гипотезы Ф.

Каково различие двух формальных систем, которые обсуждаются здесь? Первоначальная формальная система есть обычная логическая машина для доказательства теорем. Обоснованность этой формальной системы определяется наличием неопровержимых (как для человека, так и для компьютера) математических утверждений. Теперь компьютер фиксирует свое «присутствие», или свою «специфику», тем, что добавляет к первоначальной формальной системе гипотезу, которая и составляет его специфику, а именно, гипотезу, что в основе понимания математических утверждений лежит некоторый набор механизмов Ф. Компьютер полагает, что новая система будет столь же обоснованной, как и прежняя. В конце концов, для него ∇Ф-утверждения являются столь же неопровержимыми, как и просто ∇-утверждения.

Но, как видно, именно такой ход мысли приводит компьютер к противоречию. Стало быть, гипотезу нужно отбросить, и отбросить ее следует с точки зрения компьютера. Отсюда следует вывод: «Ни одно обладающее сознанием и имеющее понятие о математике существо… не может функционировать в соответствии с каким бы то ни было набором постижимых механизмов, вне зависимости от того, *знает* ли оно в действительности о том, что именно эти механизмы, предположительно, направляют его на его пути к неопровержимой математической истине. (Неопровержимой математической истиной считается то, что устанавливается доказательством, не обязательно формальным)» [Пенроуз, 2003. P. 267].

Ряд исследователей считают этот аргумент весьма изобретательным и тонким, требующим тщательного анализа. Суть этого аргумента состоит в том, что человек (идеализированный, никогда не ошибающийся математик) или компьютер не может непротиворечиво верить, что некоторый данный алгоритм перечисляет доказуемые арифметические предложения (или даже Π_1-предложения). Если же предположить, что в основе постижения математических истин лежит некоторый алгоритм, то оказывается, что мы не можем описать его. Поскольку мы не можем писать его, мы не можем утверждать, что этот алгоритм перечисляет К, множество арифметических познаваемых истин. Если мы предположим, что К является перечислимым, мы немедленно должны отвергнуть гипотезу, что в основе постижения математических утверждений лежит некоторый алгоритм.

С. Шапиро предлагает такую экспликацию «нового» аргумента Р. Пенроуза [Shapiro, 2003]. Предположим, что множество познаваемых или доказуемых математических утверждений К’ рекурсивно перечислимо. Это означает, что имеется некоторая машина Тьюринга, которая перенумеровывает К. Обозначим геделево число этой машины Тьюринга через e и множество перечисляемых при этом утверждений – через W_e. Тогда $К = W_e$. Далее, пусть К’ будет множеством арифметических предложений, относительно которых идеализированный математик или компьютер неопровержимо знает, что они следуют из гипотезы, что $К = W_e$. Структура аргумента Пенроуза такова

(Пенроуз воспроизводит возможную аргументацию компьютера или идеализированного математика):

(1) Предположим, что К = W_e.

(2) Тогда каждый член К' истинен, и, таким образом, К' непротиворечиво (это следует из утверждения об обоснованности «понимания» компьютера).

(3) Если справедливо (1), тогда К' рекурсивно перечислимо и идеализированный субъект (компьютер или идеализированный математик) может написать геделево предложение G'для К' (используя е).

(4) Поэтому, по (1), если К' непротиворечиво, тогда G' истинно, но не в К'.

(5) Но из (1) и (2) следует, что К' непротиворечиво. Поэтому G' истинно, но не в К'.

(6) Поэтому мы знаем, что если К = W_e, тогда G'. Из утверждений (1) и (5) следует G', и поэтому можно избавиться от предположения-гипотезы (1). Таким образом, G' содержится в К' (по определению К').

(7) Также, если К = W_e, тогда G' не есть в К'. Это следует из предложений (1) и (5), что позволяет избавиться от (1).

(8) Таким образом, из предложений (6) и (7) следует К ≠ W_e.

Этот аргумент является значимым при условии, что W_e непротиворечиво, т. е. что известное предложение Con_e истинно. Согласно второй теореме Геделя о неполноте, Con_e не входит в К. Иначе говоря, если К = W_e и выполняются условия выводимости, тогда никто не может знать относительно е, что W_e непротиворечиво. Но сама по себе посылка о непротиворечивости W_e требует для своей истинности того, чтобы каждый член W_e был истинен. Для этого требуется предикат истины.

Еще одна версия «нового» аргумента Пенроуза

Аргумент против «механизма» Пенроуза имеет несколько версий. Обсуждаемый нами «новый» аргумент, в свою очередь, имеет две версии, одна из которых изложена в разделе 3.16, а вторая – в фантастическом диалоге робота и человека в разделе 3.23. Ряд исследователей отмечают, что «новый» аргумент Пенроуза часто игнорируется при обсуждении проблем «механизма». Между тем, как свидетельствует Д. Чалмерс: «Насколько я могу определить, этот аргумент свободен от явных недостатков, которые поражают геделевы аргументы, такой как аргумент Лукаса и ранний аргумент Пенроуза. Если он и имеет недостатки, то они лежат более глубоко. Создается впечатление, что заключение аргумента получается как по волшебству….И хотя имеются различные направления, по которым можно критиковать аргумент, нет никакого сногсшибательного ответа. По этой причине это настоящий вызов сторонникам искусственного интеллекта» [Chalmers, 1995].

Рассмотрим структуру второй версии «нового» аргумента Пенроуза.

Как и прежде, ∇-утверждения – это утверждения, истинность которых гарантируется. Далее, пусть имеется вычислительная процедура, генерирующая ∇-утверждаемые Π_1-предложения. Эта процедура обозначается через Q. В формальной системе, представленной такой вычислительной процедурой, должно существовать некоторое геделево Π_1-предложение G(Q). Это утверждение будет истинным, что является следствием истинности ∇-утверждений. Но отсюда следует, что в формальной системе невозможно установить истинность G(Q), по крайней мере, с ∇-уверенностью.

Но то обстоятельство, что нельзя гарантировать истинность G(Q)-утверждений, бросает тень на основную посылку рассуждения, а именно на то, что в основе действий компьютера (робота) лежат механизмы Ф. Ведь Π_1-предложение G(Q) есть следствие остальных Π_1-высказываний, которые и есть проявление этих механизмов. Сомнения компьютера в том, что он действует согласно некоторому алгоритму, или механизму Ф, как раз и свидетельствуют о его «подчиненном» положении по сравнению с человеком. Компьютер, во спасение своей веры в то, что он равен человеку, не верит в то, что человек знает, что компьютер действует согласно некоторому алгоритму. А раз компьютер не верит в то, что человек знает, компьютер не может доказать истинность G(Q), тогда как человек может это сделать, опираясь на статус ∇-утверждений, общих для человека и компьютера.

Во устранение сомнений компьютера в то, что он действует согласно некоторому алгоритму, последнее обстоятельство можно ввести как отдельную гипотезу. Как и в первой версии «нового» аргумента, мы получаем с участием этой гипотезы неопро-

вержимые математические утверждения, которые назовем ∇Ф-утверждениями. Эти ∇Ф-утверждения будут включать в себя и прежние ∇-утверждениями, а также все те утверждения, которые компьютер может вывести, исходя из допущения, что его действием управляет алгоритм Ф. В число ∇Ф-утверждений войдет G(Q)-утверждение. Идея такова, что знание правил Ф дает возможность получить новую алгоритмическую процедуру Q*, которая будет генерировать только такие ∇Ф-утверждения (а также логические следствия из них), истинность которых подтверждается, исходя из допущения, что в основе конструкции компьютера лежат правила Ф.

И теперь возникает вопрос, истинно ли геделево предложение в новой системе G(Q*)? Другими словами, является ли оно неопровержимо истинным? С одной стороны, ясно, что истинность G(Q*) следует из допущения, что компьютеры построены в соответствии с правилами Ф. В этом смысле Π_1-утверждение G(Q*) должно быть ∇Ф-утверждением. С другой стороны, G(Q*) не может быть одним из ∇Ф-утверждений. Мы имеем явное противоречие. Приходится признать, что компьютер должен отвергнуть саму гипотезу о том, что он сконструирован согласно некоторым правилам Ф. Но именно эта гипотеза позволяет компьютеру его «состязание» с человеком – компьютер как формальная система «тщится схватить» мыслительные силы человека. Именно эта формулировка кладется в основу реконструкции аргумента Пенроуза такими исследователями, как Линдстрем и Чалмерс.

Тактически вторая версия нового аргумента строится обратным по отношению к первой версии образом. Если там компьютер «претендовал» на то, чтобы сравняться с человеком, то во второй версии человек «обнаруживает», что его Я ограничено некоторой формальной системой, или что он функционирует по некоторому алгоритму. Другими словами, «Я» человека «схвачено» некоторой формальной системой или алгоритмом.

Вот каким образом реконструирует структуру аргумента Д. Чалмерс [Chalmers, 1995].

(1) Предположим, что мыслительные способности человека схвачены некоторой формальной системой Ф. Другими словами, человеком рассматривается утверждение, что человек – «Я» – есть Ф). Рассмотрим класс утверждений, истинность которых, при данном предположении, известна человеку. Интерес, прежде всего, представляют неопровержимые математические утверждения.

(2) Если дано, что человек знает, что он есть Ф, человек знает, что формальная система Ф обоснована, поскольку человек про себя знает, что его система мысли обоснована. В самом деле, человек знает, что обоснованной является более широкая система Ф', где Ф' есть Ф, дополненная дальнейшим предположением «»Я» есть Ф». Как известно, дополнение обоснованной системы истинным утверждением дает обоснованную систему.

(3) Поэтому Я знаю, что геделево предложение G(Ф') истинно для системы Ф'.

(4) Но в рамках системы Ф' невозможно было бы «видеть», что G(Ф‘) истинно, что следует из теоремы Геделя.

(5) Однако, по предположению, «Я» человека не эффективно эквивалентно Ф'. В конце концов, то обстоятельство, что человеческое «Я» есть Ф, дополнено знанием, что «Я» есть Ф.

(6) Это противоречие, и поэтому исходное предположение должно быть ложным. Значит, Ф не должно схватывать мыслительные способности человека.

(7) Поскольку в рассуждении идет речь о произвольной формальной системе, мыслительные способности человека не могут быть «схвачены» никакой формальной системой.

Данный аргумент связан с некоторыми предположениями, которые позволяют проявить большую предосторожность в отношении достигнутого заключения. Тот же Д. Чалмерс отмечает, что, строго говоря, из аргумента следует заключение, что человек не может *знать*, что он тождествен формальной системе Ф. Далее, «видение» человеком истинности геделева предложения G(Ф') предполагает не просто, что человек есть Ф, но и то, что он знает, что он есть Ф. Такого рода рефлексия важна для понимания того, какой смысл вкладывается в понятие, что человеческое мышление «схвачено» формальной системой или алгоритмом.

Однако одна лишь рефлексия не исчерпывает возможных заключений из аргумента. Например, открытие, что «Я» есть фор-

мальная система Ф, могло бы быть чисто эмпирическим открытием, а не результатом «видения» истинности геделева предложения. В случае эмпирического открытия подобного толка аргумент против «механизма» не теряет своей силы. Но собственно «геделевская проблематика» в аргументе состоит не в том, что получаемое противоречие свидетельствует о доказательстве геделева предложения в более широкой формальной системе Ф', а в том, что эта формальная система «видит» свое собственное геделево предложение.

Д. Чалмерс полагает, что сила аргумента не зависит от способности человека к определению того, что система Ф обоснована, и полагает, что вопрос упирается в неопровержимость утверждения о непротиворечивости человека. Другими словами, именно это предположение может быть поставлено в упрек новому аргументу Р. Пенроуза. Вопрос о непротиворечивости тут действительно важен, но П. Линдстрем, осознавая важность вопроса о непротиворечивости, начинает свой анализ нового аргумента Пенроуза именно с анализа обоснованности [Lindstrom, 2001]. Больше того, он полагает, что в этом новом аргументе Пенроуза не предполагается, что человек знает о непротиворечивости Ф.

Прежде всего, П. Линдстрем предлагает неформальную реконструкцию аргумента Пенроуза, с уточнением, что фраза «Я есть Ф» означает, что Ф объемлет все человечески доступные методы математического доказательства. Эта реконструкция фигурирует у него под индексом (А):

Хотя человек не знает, что необходимо является Ф, он заключает, что если бы он был Ф, система Ф была бы обоснованной. Пусть Ф* будет Ф, дополненная утверждением «Я есть Ф». Тогда Ф* будет обоснованной. Человек воспринимает, что из предположения, что он есть Ф, следует, что геделево предложение G(Ф*) будет истинным, и далее, оно не будет следствием Ф*. Но он только что воспринял, что «если Я есть Ф», тогда G(Ф*) и восприятие этой природы будет точно тем, что предполагает достичь Ф. Следовательно, так как человек способен к восприятию того, что находится за пределами Ф, он делает вывод, что он не есть Ф.

Далее, Линдстрем представляет формализованную версию нового аргумента, которая фигурируют под индексом (В):

Основной ингредиент нового аргумента «Я есть Ф» включает две составляющие части. Первая из них – это посылка об обоснованности формальной системы Ф, которая символизируется в виде Sd (Ф). Вторая составляющая часть – это посылка о полноте формальной системы Ф, в том смысле, что Ф объемлет все человечески доступные методы математического доказательства. Этот вид полноты символизируется через НС(Ф). Тогда выражение «Я есть Ф» символизируется через Sd (Е) & НС(Ф).

Далее, пусть Ф + S есть система такая, что для каждой S', Ф + S $\vdash$ S', если и только если, Ф $\vdash$ S$\rightarrow$S'. Вводится новая система Ф* = Ф + Sd (Ф), которая объемлет старую систему Ф плюс предположение о том, что эта формальная система обоснована. В этих обозначениях аргумент (В) может быть представлен так:

(1) Sd (Ф) $\Rightarrow$ Sd (Ф*)
(2) Sd (Ф*) $\Rightarrow$ G (Ф*)
(3) Sd (Ф*) $\Rightarrow$ Ф* $\nvdash$ G (Ф*)
(4) Sd (Ф) $\Rightarrow$ G (Ф*)
(5) Sd (Ф) $\Rightarrow$ Ф* $\nvdash$ G (Ф*)
(6) НС(Ф) $\Rightarrow$ Ф $\vdash$ Sd (Ф) $\rightarrow$ G (Ф*)
(7) $\neg$ (Sd (Ф) & НС(Ф))
(8) $\neg$ «Я есть Ф»

Содержательно строчка (1) соответствует предположению, что «схватывание» формальной системой Ф мыслительных возможностей человека влечет обоснованность новой формальной системы, состоящей из старой системы плюс предположение о схватывании. Другими словами,

(9) Я есть Ф $\Rightarrow$ Sd (F + «Я есть Ф»).

Это утверждение интуитивно приемлемо, поскольку принятие предположение о схватывании мыслительных возможностей человека формальной системой Ф, добавление этого предположения к самой формальной системе Ф столь же обосновано, как и сама система. Поэтому новая система Ф* будет обоснованной.

Конечно, возникает вопрос о том, что же означает в данном контексте «обоснованность» формальной системы. Тут можно воспользоваться практически бесспорным обстоятельством, что если система обосно-

вана, она является непротиворечивой. Иначе говоря,

$$Sd\ (Ф) \Rightarrow Con(Ф).$$

Тогда, согласно исходному аргументу Геделя, мы можем заключить, что (2) и (3) истинны. (4) следует из (1) и (2), и (5) следует из (1) и (3). Предположим, доказательство (4) есть математически неопровержимое доказательство. Тогда, если HC(F), тогда (4) может быть доказано в F, другими словами, (6) истинно. Наконец, (7) следует из (5) и (6), и (8) следует из (7).

Заметим, что в (B) не утверждается, что (8) математически доказано; достаточно того, что (8) истинно.

Для завершения (B) нужно определить «обосновано» таким образом, чтобы стало ясно, что (1) (или более слабое (10) ниже) неопровержимо истинно. Значение термина «обоснованность», как его употребляют в дискуссии механицисты и менталисты, может иметь разные определения. Мы предъявляем различные интуитивные требования к нему. Одним из таких требований является условие (1), т. е.

(10) $Sd\ (Ф) \Rightarrow Sd\ (Ф^*)$.

В данном случае интуиция подсказывает нам, что если некоторая система Ф обоснована, тогда обоснованной является и система, полученная из исходной прибавлением к ней утверждения о ее обоснованности. Обоснованность системы значит также, с интуитивной точки зрения, и непротиворечивость этой системы, и как отмечает Линдстрем, принцип (1) можно заменить на принцип $Sd\ (Ф) \Rightarrow Con(Ф^*)$. Такого рода замена выглядит вполне невинно, и более того, она имеет то преимущество, что (10) слабее (1). Таким образом, в понятие «обоснованности» не включается «слишком много».

В своем обсуждении тезиса о неалгоритмизуемости человеческого мышления Пенроуз всякий раз оговаривается, что имеет дело с $П_1$-предложениями. (Существенно, PA-предложения есть $П_1$-предложение, если они имеют вид $\forall x Rx$ и разрешимы в Ф, то есть, для каждого n, либо Ф производит R(n), либо Ф проиводит $\neg R(n)$. Наиболее известные теоремы и проблемы в PA, включая теорему Ферма и проблему Гольдбаха, являются $П_1$-предложениями. В частности, предложения G(E) и Con(E) есть $П_1$-предложения). Но такое ограничение (вполне разумное во многих контекстах) влечет нежелательный для Пенроуза результат. Действительно, как хорошо известно, Ф является $П_1$-обоснованной, если и только если, Ф непротиворечива. Но тогда хорошо известен и тот факт, что с таким определением «обоснованности» (10) не истинно в общем: Если Ф непротиворечива, тогда, по второй теореме Геделя, $Ф + \neg Con(Ф)$. Но ясно, что $Ф + \neg Con(Ф) + Con(Ф + \neg Con(Ф))$ противоречиво. Таким образом, $E = Ф + \neg Con(Ф))$ есть контрпример (10).

Этот нежелательный результат можно было бы отнести за счет конкретного определения обоснованности, в данном случае, определения $П_1$-обоснованности. Но как показал Линдстрем, если принимается вполне разумное предположение

(11) E истинно $\Rightarrow Sd(E)$,

тогда (1) не обязательно является истинным при самых разнообразных определениях обоснованности [Lindstrom, 2001. P. 246]. Интересным в этой связи является то обстоятельство, что обоснованность никак не может быть определена как истинность. Действительно, при таком определении (6) не было бы истинным.

Список литературы

Пенроуз Р. Тени разума. М., 2003. Раздел 3.16.

Chalmers D. Minds, Machines, and Mathematics // Psyche. June 1995. Vol. 2 (9).

Godel K. Some Basic Theorems on the Foundations of Mathematics and Their Implications // Collected Works III / Eds. S. Feferman et al. Oxford: Oxford University Press, 1995.

Lindstrom P. Penrose's New Argument // Journal of Philosophical Logic. 2001. Vol. 30. P. 241–250.

McCullough D. Can Humans Escape Godel? // Psyche. April 1995. Vol. 2 (4).

Shapiro S. Mechanism, Truth, and Penrose's New Argument // Journal of Philosophical Logic. 2003. Vol. 32. P. 19–42.

Материал поступил в редколлегию 01.09.2009

V. V. Tselischev

KNOWLEDGE REPRESENTATION IN THE LIGHT OF SECOND PENROSE'S ARGUMENT

In knowledge formalization a special interest has a question, which is opposite to the strong problem of knowledge representation, that is, whether an informal concept can «understand the knowledge» of a computer. Such formulation of the question is usually called a «second computer argument». This paper studies the logical structure of the argument.

Keywords: knowledge formalization, algorithm, Gödel's proposition, second Penrose's argument.