Перспективы искусственного интеллекта и теорема Пенроуза

Часть 2. Теорема Пенроуза об ИИ и квантовая природа сознания

А. Д. Панов
НИИЯФ МГУ

2018.10.11 - Философское наследие Рерихов и наука

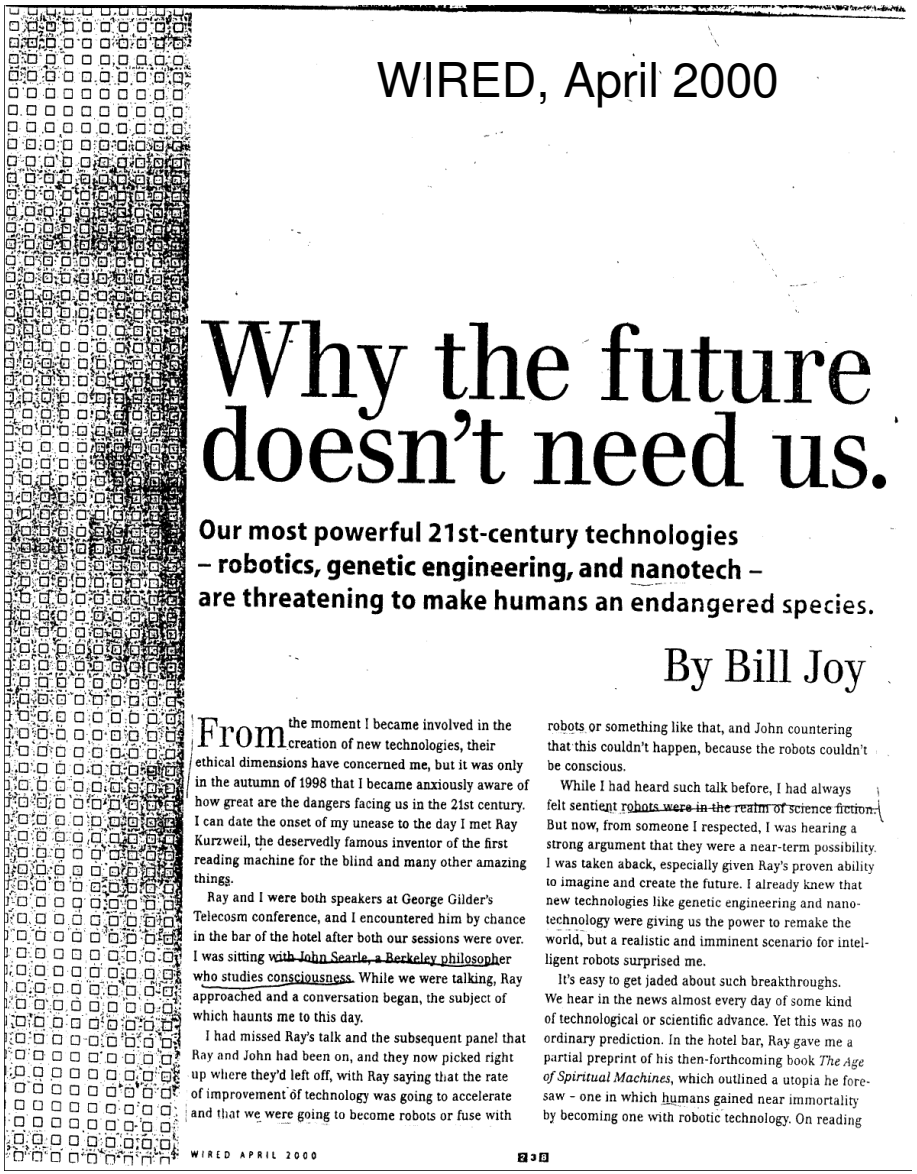
Сильный искусственный интеллект и технологическая сингулярность

Искусственный Интеллект (ИИ) превосходит человеческий

<u>во всех отношениях</u>

- сильный ИИ -

→ предсказание будущего становится невозможным



WIRED, April 2000

Why the future doesn't need us.

Our most powerful 21st-century technologies – robotics, genetic engineering, and nanotech – are threatening to make humans an endangered species.

By Bill Joy

From the moment I became involved in the creation of new technologies, their ethical dimensions have concerned me, but it was only in the autumn of 1998 that I became anxiously aware of how great are the dangers facing us in the 21st century. I can date the onset of my unease to the day I met Ray Kurzweil, the deservedly famous inventor of the first reading machine for the blind and many other amazing things.

Ray and I were both speakers at George Gilder's Telecosm conference, and I encountered him by chance in the bar of the hotel after both our sessions were over. I was sitting with John Searle, a Berkeley philosopher who studies consciousness. While we were talking, Ray approached and a conversation began, the subject of which haunts me to this day.

I had missed Ray's talk and the subsequent panel that Ray and John had been on, and they now picked right up where they'd left off, with Ray saying that the rate of improvement of technology was going to accelerate and that we were going to become robots or fuse with robots or something like that, and John countering that this couldn't happen, because the robots couldn't be conscious.

While I had heard such talk before, I had always felt sentient robots were in the realm of science fiction. But now, from someone I respected, I was hearing a strong argument that they were a near-term possibility. I was taken aback, especially given Ray's proven ability to imagine and create the future. I already knew that new technologies like genetic engineering and nanotechnology were giving us the power to remake the world, but a realistic and imminent scenario for intelligent robots surprised me.

It's easy to get jaded about such breakthroughs. We hear in the news almost every day of some kind of technological or scientific advance. Yet this was no ordinary prediction. In the hotel bar, Ray gave me a partial preprint of his then-forthcoming book *The Age of Spiritual Machines*, which outlined a utopia he foresaw – one in which humans gained near immortality by becoming one with robotic technology. On reading

Основные направления критики

Три плохо обоснованных предположения

1. Переоценка фактора мощности компьютера и недооценка фактора программного обеспечения в создании ИИ
2. Возможная недооценка скорости вычислений мозга
3. Аналогия мозг-классический компьютер

Одно полностью не понятое обстоятельство

No-go теорема Пенроуза об ИИ -

Компьютеры классической архитектуры не могут привести к созданию сильного ИИ в принципе

Все главные выводы Роджера Пенроуза вообще не поняты, поэтому некому их критиковать.

План:

1. **Представления Роджера Пенроуза «для пешеходов»**
2. **Критика**

No-go теорема Пенроуза об ИИ

Фундаментальные запреты в науке

Закон сохранения энергии (первое начало термодинамики)

Запрещает создание вечных двигателей первого рода

Второе начало термодинамики

Запрещает создание вечных двигателей второго рода

Теорема Пенроуза об искусственном ИИ

Запрещает создание сильного ИИ на базе конечного автомата

> Какой бы мощностью ни обладало устройство, имеющее архитектуру конечного автомата, человеческое мышление имеет некоторые возможности, недоступные такому устройству.

Р. Пенроуз. Новый ум короля. УРСС, Москва, 2003.

Р. Пенроуз. Тени разума: В поисках науки о сознании. Институт компьютерных исследований, Москва-Ижевск, 2005.

dec1.sinp.msu.ru/~panov/Penrose-Shadows.pdf

Shadows of the Mind

A Search for the Missing Science of Consciousness

ROGER PENROSE

Rouse Ball Professor of Mathematics
University of Oxford

OXFORD UNIVERSITY PRESS
New York Oxford

РОДЖЕР ПЕНРОУЗ

ТЕНИ РАЗУМА

В ПОИСКАХ НАУКИ О СОЗНАНИИ

Перевод с английского
А. Р. Логунова и Н. А. Зубченко

Москва ♦ Ижевск
2005

1-я теорема Гёделя о неполноте

Для любой аксиоматической системы, которая
1) Непротиворечива
2) Содержит в себе формальную арифметику

можно сформулировать осмысленное утверждение, которое нельзя ни доказать, ни опровергнуть средствами этой системы.

Доказательство конструктивно -
это утверждение строится в явном виде
(и является истинным по построению)

Используются два основных метода:
-Гёделева нумерация и
-Диагональный метод Кантора

Имеются аксиоматические системы, для которых теорема Гёделя неверна.
Пример: Арифметика Пресбургера (без умножения)

Теорема Гёделя-Тьюринга

Для любого конечного автомата, который
1) Реализует обоснованные процедуры
2) Достаточно силен, чтобы реализовывать алгоритмы, анализирующие другие алгоритмы на предмет их остановки

можно сформулировать осмысленное утверждение, истинность которого не может быть вычислена этим автоматом.

Доказательство конструктивно -
это утверждение строится в явном виде
если известна структура автомата
(и является истинным по построению)

Используются два основных метода:
-Гёделева нумерация и
-Диагональный метод Кантора

Рэй Курцвейл: Мозг – это тоже компьютер.
Для мозга тоже существуют Гёделевские утверждения.
Чем же мозг лучше компьютера?

No-go теорема Пенроуза об ИИ (полное доказательство)

1. Предположим, что некоторый компьютер, имеющий архитектуру конечного автомата, обладает всеми интеллектуальными способностями всего человечества (представляет собой сильный ИИ в узком смысле).

2. Тогда, любой математик, **используя свои математические способности**, на основе теоремы Гёделя-Тьюринга может построить истинное утверждение, истинность которого не может быть проверена **этим** компьютером, но которая ясна для математика (по построению). Построение всегда возможно, так как доказательство теоремы Гёделя-Тьюринга имеет конструктивный характер.

3. Следовательно, предполагая, что компьютер обладает всеми способностями людей, мы немедленно указываем способность человека, которой этот компьютер не обладает.

4. Это есть противоречие, и оно доказывает, что такой компьютер (сильный ИИ) не может существовать.

Сильный ИИ невозможен ни для каких компьютеров на основе архитектуры конечного автомата.

По мнению Пенроуза мозг – это не компьютер...

То, что мозг — это не компьютер,
НЕ является мнением Пенроуза -
это теорема, доказанная Пенроузом

Является ли теорема Пифагора
«мнением» Пифагора?

Алгоритмы компьютера должны быть познаваемыми (Роджер Пенроуз).

Структура компьютера может оказаться настолько сложной, что не может быть представлена в виде записи и изучена (непознаваема), поэтому и построить для него гёделевское утверждение будет невозможно.

- Конечный автомат по определению есть классическое устройство.
- Будучи классическим объектом он допускает исчерпывающее измерение своего состояния в любой момент времени.
- Информационная запись такого измерения может быть прочитана и изучена - алгоритмы познаны исчерпывающим образом.
- Конечный автомат сам является информационной записью собственных алгоритмов, поэтому алгоритмы конечного автомата в принципе всегда познаваемы.

Мозг математика, способного понять теорему Гёделя-Тьюринга, не является компьютером (альтернативная формулировка теоремы Пенроуза).

- Предположим, что мозг математика, способного понять теорему Гёделя-Тьюринга является конечным автоматом.

- Тогда алгоритмы этого конечного автомата могут быть конечным образом изучены, как и алгоритмы любого конечного автомата.

- По этим алгоритмам математик, пользуясь построением Гёделя-Тьюринга, может построить Гёделевское утверждение, которое является истинным по построению, но истинность которого математик не может установить, будучи именно тем конечным автоматом, для которого он сам это Гёделевское утверждение и строит.

- Тем самым математик истинность этого Гёделевского утверждения одновременно знает и не может установить. Противоречие доказывает, что мозг математика – не конечный автомат.

Ср. с утверждением Рэя Курцвейла.

Возражения (Пенроуз рассмотрел 20 возражений)

- Какова природа теоремы Пенроуза (все человечество не является математическим понятием...)? Это вообще не теорема?
  - Правильная теорема математической логики.

- Теорема Пенроуза адресует не реальные вычислительные машины, но идеальные конечные автоматы - машины Тьюринга.
  - Методическая ошибка, природа познаваема только на основе моделей

- Теорема Пенроуза основана на методе доказательства от противного, который не признается конструктивной математикой.
  - Откажитесь от математического анализа...

- Выпишем в ряд все правила рассуждений, которые использует математик, и запрограммируем их в компьютере. Тогда компьютер сможет проводить все рассуждения, которые может и математик.
  - Математик не конечный автомат, поэтому не следует ожидать конечного списка правил рассуждения.

Только ли в отношении некоторых математических способностей люди превосходят компьютеры?

Математические способности людей представляют только частный случай способностей, когда анализ удается довести до конца (в виде теоремы).

Поскольку в отношении математиков строго доказано превосходство человека над машиной, то, по аналогии, и не математики могут обладать способностями, недоступными конечному автомату.

Вероятно, мозг любого отдельного человека - не компьютер.

Совокупность мозгов всех людей - точно не компьютер.

Все люди обладают универсальной неформализуемой способностью к пониманию, и математические способности – лишь частный случай.

Стругацкие: «все фундаментальные идеи выдумываются..., они не висят на концах логических цепочек»

Роджер Пенроуз был не первый кто понял, что мозг — не компьютер.

1972, рус.1978

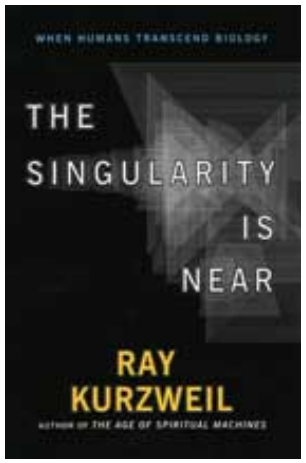
Хьюберт Дрейфус.

Почему мозг не компьютер?

Машины обрабатывают информацию, а человек работает со смыслами.

Вовсе не очевидно, что человеческие смыслы могут быть закодированы информацией.

Мысль — не вычисление



Рэй Курцвейл:

Для компьютера существуют Гёделевские утверждения.

Мозг — это тоже компьютер.

Для мозга тоже существуют Гёделевские утверждения.

Чем же мозг лучше компьютера?

Роджер Пенроуз:
Мозг реализует
невычислимую активность.

Проблемы математики:

1. Алгоритмически разрешимые (вычислимые).
   Решение систем линейных уравнений, …
2. Алгоритмически неразрешимые (невычислимые).
   Решение систем диофантовых
   уравнений общего вида, ...

Описание работы мозга не может быть представлено алгоритмически разрешимой задачей.

Если мозг - не компьютер, **ТО ЧТО ЭТО?**

В чем источник невычислимости активности мозга?

Природа невычислимой активности мозга по Роджеру Пенроузу

Нисходящее и восходящее программирование (нейронные сети).

Нейронная структура сама по себе никакой невычислмости породить не может, так как отображается на конечный автомат.

Аналоговые вычисления против цифровых вычислений.

Не заключается ли причина в открытости мозга как системы?

- Предположение 1: вне мозга невычислимой физики нет.

  Тогда никакое окружение невычислимости породить не может.

- Предположение 2: вне мозга невычислимая физика есть.

  Тогда она может быть и внутри мозга, а так как мозг - самая сложная система именно там ее нужно искать.

Причиной является невычислимая физика внутри мозга. Какая?

- невычислимая физика, в принципе, может существовать
- механика
- классические поля
- классическая статистическая физика (ансамбли)

  'Mersenne twister' random number generator: период $2^{19937}-1$

STOP ----------

КВАНТОВАЯ ФИЗИКА?!

- «обычная» квантовая физика — тоже вычислима!
Мозг — не квантовый компьютер!

Роджер Пенроуз:

Мозгу недостаточно быть квантовым компьютером, чтобы реализовывать невычислимую активность.

Но мы не знаем ничего сверх квантовой физики.

Что же остается? –

Еще не открытая, новая «невычислимая физика»
(квантовая гравитация, OR-процедура)

Тени разума, стр. 569:

Как бы то ни было, представляемые мною аргументы предполагают не только макроскопическую квантовую когерентность. Они предполагают, что биологическая система, называемая человеческим мозгом, каким-то образом ухитрилась воспользоваться в своих интересах физическими феноменами, человеческой же физике неизвестными! Эти феномены когда-нибудь опишет несуществующая пока теория **OR**, которая свяжет вместе классический и квантовый уровни и, я убежден, заменит временную **R**-процедуру иной, чрезвычайно тонкой и невычислимой (но все же, несомненно, математической) физической схемой.

В новую физику могут вести не только
1) суперколлайдеры (микрофизика) и
2) супертелескопы (космология ранней вселенной),
но и
3) **наука о сознании (!!!)**.

микрофизика	- самые мелкие масштабы
космология	- самые большие масштабы
мозг	- самые сложные сруктуры

Критика интерпретации Пенроуза его же теоремы

Вне мозга невычислимая активность есть – это невычислимая активность Вселенной в целом (риторическое замечание)

Невозможно в конечный автомат поместить модель всей Вселенной, в частности, из-за бесконечной рекурсии — невычислимость.

От Вселенной следует постоянно ожидать чего-то абсолютно непредсказуемого.

Вряд ли это имеет **прямое** отношение к Гёделевской нумерации и диагональному методу Кантора.

Хотя неточность в аргументации Пенроуза имеется, вряд ли это имеет отношение к невычислимой активности мозга в том смысле, который Пенроуз имеет в виду.

Вычислимость в квантовой физике.
I. «Наивный подход»

Алгоритмическая разрешимость задач квантовой теории:

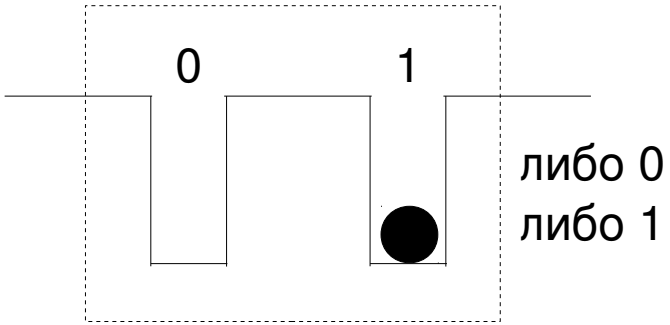
1. Состояния систем — векторы гильбертова пространства
2. Эволюция — унитарные преобразования или решение систем линейных дифференциальных уравнений
3. Измерения — скалярные произведения векторов, симуляция: генераторы случайных чисел.

Формально говоря:
Фрагменты квантовой реальности допускают исчерпывающее представление в классическом компьютере
В частности: квантовые компьютеры можно симулировать на классическом компьютере (и симуляторы существуют)

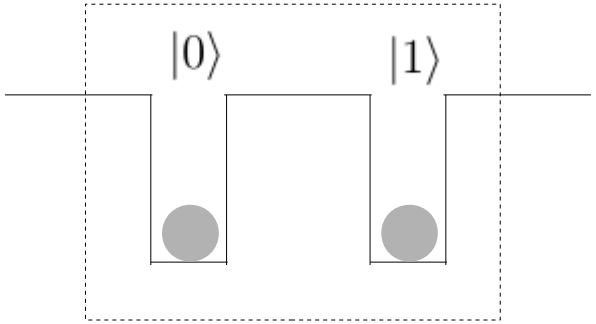
B. Julia-Dıaz, J.M. Burdis, and F. Tabakin.
Qdensity – a Mathematica quantum computer simulation.
arXiv:quant-ph/0508101.

Квантовый компьютер

Бит



Кубит



$$A_1|0\rangle + A_2|1\rangle$$

2 кубита

$$A_1|0\rangle|0\rangle + A_2|0\rangle|1\rangle + A_3|1\rangle|0\rangle + A_4|1\rangle|1\rangle$$

1000 кубитов

$$\begin{aligned} &A_1|0\rangle \ldots |0\rangle + \\ &A_1|0\rangle \ldots |1\rangle + \\ &\ldots \\ &\ldots \\ &A_{2^{1000}}|1\rangle \ldots |1\rangle \end{aligned}$$

Вычислимость в квантовой физике.
II. «Космологический горизонт» вычислимости.

1000-кубитный квантовый компьютер

$$A_1|0\rangle \ldots |0\rangle +$$
$$A_1|0\rangle \ldots |1\rangle +$$
$$\ldots$$
$$\ldots$$
$$A_{2^{1000}}|1\rangle \ldots |1\rangle$$

Для представления состояния требуется $2^{1000} \approx 10^{300}$ комплексных чисел.

Всего внутри космологического горизонта можно представить
не более, чем примерно 10^{90} бит информации (в обычной материи).
примерно 10^{175}, если бит разместить в каждой планковской ячейке.

Сложные квантовые системы строго невычислимы в смысле
космологического горизонта вычислимости.

Вычислимость сложных квантовых систем не имеет физического смысла

- Квантовая физика **мозга** заведомо невычислима классическим компьютером в смысле «космологического горизонта» вычислимости. Имеет место «физическая невычислимость».

- Нет уверенности, что аргумент Пенроуза от вычислимости квантовой теории в пользу еще неизвестной невычислимой (квантовогравитационной ?) физики неотразим.

- Возражение Пенроуза: мы не должны апеллировать к физике, когда проводим формальное математическое доказательство.

- Гипотеза: обычной квантовой физики может хватить для того, чтобы обеспечить «физическую невычислимость» процессов, происходящих в мозге.

- В принципе сохраняется возможность создать квантовый симулятор сознания – если удастся понять как это сделать.

Вычислимость в квантовой физике.
III. Неинформационная природа квантовых состояний

Хьюберт Дрейфус: Почему мозг не компьютер?

Машины обрабатывают информацию, а человек работает со смыслами.

Вовсе не очевидно, что человеческие смыслы могут быть закодированы информацией.

Теорема Пенроуза → в формировании смыслов **должны** играть роль квантовые процессы.

Квантовые состояния не являются носителями информации.

Возможность копирования — одно из основных свойств информации

Теорема о запрете клонирования состояний (no-cloning theorem)

Процесс $|A_\chi\rangle|B_0\rangle \longrightarrow |A_\chi\rangle|B_\chi\rangle$ запрещен

Квантовые состояния не допускают копирования, следовательно не обладают важнейшим свойством информации.

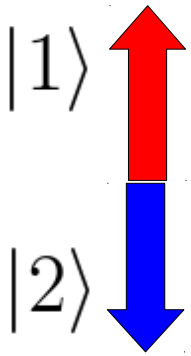
Носителем чего являются квантовые состояния?

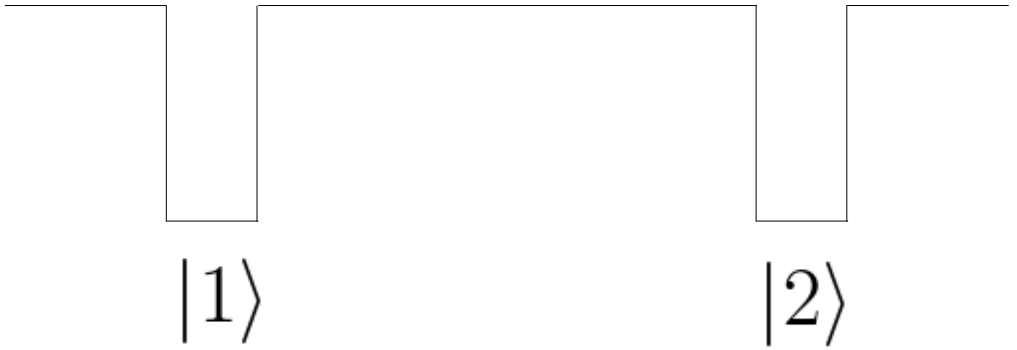
Квантовая информация

Одна и та же информация может быть записана на разные носители.
Информация — это *абстракция* от распределения физических неоднородностей, кодирующих эту информацию.

Квантовые состояния одной и той же структуры могут быть реализованы на разных носителях.

$$|\Psi\rangle = a_1|1\rangle + a_2|2\rangle$$





Абстракция от структуры квантового состояния, независимо от природы системы - это **квантовая информация**.

Квантовая информация ≠ Информация

Смыслы, с которыми работает мозг, могут на самом деле оказаться квантовой информацией

Телепортация квантовых состояний (квантовой информации)

Копирование квантовой информации запрещено:

$$|A_X\rangle|B_0\rangle \rightarrow |A_X\rangle|B_X\rangle$$

Телепортация квантовой информации разрешена:

$$|A_X\rangle|B_0\rangle \rightarrow |A_?\rangle|B_X\rangle$$

Upload сознания на искусственный носитель

Если сознание имеет квантовую природу, то:

- Можно скопировать структуры гильбертова пространства мозга, но не состояния мозга →
  Копируется мёртвая «оболочка сознания», не сознание.

- Скопировать сознание, сохранив оригинал, невозможно.

- В подготовленную на предыдущем этапе «оболочку» можно телепортировать сознание оригинала →
  Но в оригинале тогда сознание исчезает. $|A_X\rangle|B_0\rangle \to |A_?\rangle|B_X\rangle$

- Можно ли телепортировать куда-то сознание, не имея подготовленного физического субстрата?

- Сознание есть нечто, существующее принципиально в единственном экземпляре (душа)

- Телепортация сознания — довольно сложная, и «хрупкая» операция (осуществима ли принципально?)

Резюме

Пенроуз:

Теорема Пенроуза: мозг – не компьютер, мозг реализует невычислимую активность.

Невычислимая активность реализуется благодаря невычислимой физике, используемой мозгом.

Так как вся известная физика, включая квантовую физику, – вычислима, то работа мозга основана на неизвестной невычислимой физике.

Критика:

Квантовая физика таких сложных систем, как мозг, физически невычислима в нашей Вселенной в смысле космологического горизонта вычислимости.

Следовательно, чтобы эффективно проявлять невычислимую активность, мозгу достаточно использовать квантовые процессы.

Смыслы, которыми оперирует мозг, могут представляться не информацией, а квантовой информацией.