

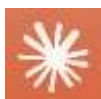
Despite its impressive output, generative AI doesn't have a coherent understanding of the world

Researchers show that even the best-performing large language models don't form a true model of the world and its rules, and can thus fail unexpectedly on similar tasks.

Adam Zewe | MIT News

Publication Date: November 5, 2024

<https://news.mit.edu/2024/generative-ai-lacks-coherent-world-understanding-1105>



Обучаются ли LLM (большие языковые модели) согласованным моделям мира



Claude-3.5-Sonnet

В отношении обучения LLM согласованным моделям мира есть несколько важных аспектов:

Основные наблюдения:

1. LLM демонстрируют определенную согласованность в своих ответах, но она не абсолютна. Модели могут давать противоречивые ответы в разных контекстах.
2. Модели усваивают статистические паттерны из обучающих данных, но это не обязательно ведет к формированию целостной "модели мира" в человеческом понимании.
3. Отсутствует явное представление причинно-следственных связей. LLM работают на основе корреляций в данных.

Основные проблемы:

- Нет гарантии, что модель сформирует внутренне непротиворечивые представления
- Сложно проверить глубину понимания и согласованность знаний
- Модель может выучить поверхностные паттерны без понимания базовых принципов

## Обучаются ли люди согласованным моделям мира?



### Claude-3.5-Sonnet

Люди, как и LLM, не всегда формируют полностью согласованные модели мира:

#### 1. Когнитивные искажения:

- Подтверждение своих убеждений вопреки противоречащим фактам
- Избирательное восприятие информации
- Противоречивые убеждения, которые могут сосуществовать

#### 2. Контекстная зависимость:

- Разные "версии правды" в разных социальных контекстах
- Ситуативное применение противоречащих друг другу моделей поведения

#### 3. Развитие представлений:

- Постепенное формирование более согласованных моделей с опытом
- Способность осознавать и разрешать противоречия
- Активная работа над согласованностью убеждений

Главное отличие от LLM - люди способны:

- Осознавать противоречия
- Целенаправленно работать над согласованностью
- Развивать более глубокое понимание причинно-следственных связей

Оцени, какой процент людей реально способен осознавать противоречия, целенаправленно работать над согласованностью, развивать более глубокое понимание причинно-следственных связей



Claude-3.5-Sonnet

На основе различных исследований и наблюдений:

Осознание противоречий и активная работа над ними:

- ~10-15% людей регулярно и систематически анализируют свои убеждения
- ~30-40% способны замечать явные противоречия при указании на них
- Остальные редко замечают или игнорируют противоречия

Работа над согласованностью:

- ~5% целенаправленно работают над согласованностью мировоззрения
- ~20% периодически корректируют отдельные убеждения
- Большинство предпочитает сохранять существующие убеждения

Понимание причинно-следственных связей:

- ~15-20% систематически исследуют глубинные причины
- ~40% способны выстраивать базовые причинно-следственные цепочки
- Остальные часто ограничиваются поверхностными объяснениями

Факторы, влияющие на эти способности:

- Образование
- Критическое мышление
- Когнитивные способности
- Мотивация к саморазвитию
- Социальное окружение

Эти оценки приблизительны и основаны на различных исследованиях когнитивных способностей и критического мышления.

## Дай несколько определений согласованной модели мира



Claude-3.5-Sonnet

Определения согласованной модели мира в разных контекстах:

Логическое:

- Система представлений, где все элементы непротиворечиво связаны между собой, и любые два утверждения не могут одновременно противоречить друг другу

Когнитивное:

- Ментальная структура, позволяющая предсказывать события и объяснять явления, используя единый набор принципов и правил

Практическое:

- Набор убеждений и представлений, позволяющий принимать решения и действовать без внутренних конфликтов

Системное:

- Взаимосвязанная сеть концепций и моделей, где изменение одного элемента предсказуемо влияет на связанные элементы

Прагматическое:

- Модель реальности, достаточно непротиворечивая для эффективного достижения целей в конкретном контексте

Важно отметить, что абсолютная согласованность, вероятно, недостижима - речь идет скорее о степени согласованности, достаточной для конкретных целей.

Дай оценку насколько распространена среди людей и как влияет создание согласованной модели мира логика Пер-Ноэля



Claude-3.5-Sonnet

Логика Пер-Ноэля, насколько я знаю, не является общепринятым или широко известным термином в философии или когнитивных науках. Не могли бы вы пояснить, что именно вы имеете в виду под "логикой Пер-Ноэля"? Это позволит мне дать более точный и полезный ответ относительно её распространённости и влияния на создание согласованных моделей мира.

Понятие, введенное Бодрийяром ("Система вещей", 1968) для обозначения суггестивного приема замещения реальных их симулякрами, как средство разрешения любых - социологически или психологически артикулированных - конфликтов. Так, отмечая Новый год, люди одновременно верят в Деда Мороза (Пер Ноэля), без чего нет праздника, и не верят в его (как рациональные и мыслящие существа).



Claude-3.5-Sonnet

Теперь, понимая контекст логики Пер-Ноэля по Бодрийяру, можно оценить её распространённость и влияние:

Распространённость:

- Очень высокая (~80-90% людей) в различных формах
- Присутствует практически во всех культурах
- Усиливается в современном обществе потребления

Основные сферы проявления:

1. Социальные ритуалы
  - Праздники
  - Традиции
  - Церемонии
2. Потребительское поведение
  - Брендинг
  - Реклама
  - Шоппинг как "терапия"
3. Социальные отношения

- Вежливость
- Социальные роли
- Статусные игры

Влияние на создание согласованной модели мира:

Позитивное:

- Помогает справляться с когнитивным диссонансом
- Облегчает социальную адаптацию
- Создает психологический комфорт

Негативное:

- Затрудняет формирование действительно согласованного мировоззрения
- Способствует уходу от реальности
- Может вести к отказу от критического мышления

Эта логика стала одним из основных механизмов адаптации к противоречиям современного мира.

LLM ограничены (пока) исключительно текстовой реальностью, как это влияет на создание ими согласованной модели мира?



Claude-3.5-Sonnet

Ограниченность LLM текстовой реальностью создает несколько важных эффектов:

Преимущества:

1. Четкая формализация
  - Текст как структурированная система значений
  - Возможность точного сопоставления концепций
  - Меньше "шума" от сенсорных данных
2. Доступ к обработанному знанию
  - Готовые абстракции и обобщения
  - Систематизированные теории
  - Описанные причинно-следственные связи

Ограничения:

1. Отсутствие сенсорного опыта

- Нет прямой связи с физической реальностью
- Отсутствие телесного якоря для понимания
- Сложность с пространственными концепциями

## 2. Проблемы с причинностью

- Только статистические корреляции
- Отсутствие эмпирической проверки
- Сложность различения существенных и случайных связей

## 3. Метауровень

- Трудности с осознанием границ собственного понимания
- Неясность критериев согласованности
- Проблемы с самореферентностью

Это создает парадоксальную ситуацию: модели могут демонстрировать высокую логическую согласованность при потенциальном отрыве от реальности.

## Перспективы и пути преодоления этой ограниченности (текстовой реальностью)



Claude-3.5-Sonnet

Перспективные направления преодоления текстовой ограниченности:

### 1. Мультимодальные модели

- Интеграция текста, изображений, звука
  - Понимание пространственных отношений
  - Связь символических и сенсорных данных
- Ограничение: всё равно работа с "записями" реальности

### 2. Воплощенный ИИ (Embodied AI)

- Роботизированные системы с сенсорами
  - Прямое взаимодействие с физическим миром
  - Формирование причинно-следственных связей через опыт
- Проблема: сложность масштабирования опыта

### 3. Гибридные системы

- Комбинация языковых моделей и физических агентов
  - Распределенные системы сбора и обработки опыта
  - Коллективное обучение через shared experience
- Сложность: интеграция разных типов знаний

#### 4. Симуляционный подход

- Детальные модели физического мира
  - Виртуальные эксперименты
  - Проверка гипотез в симулированной среде
- Риск: неполнота/неточность моделей реальности

Временные рамки:

- Ближайшие 5 лет: улучшенные мультимодальные системы
- 5-10 лет: первые эффективные гибридные системы
- 10+ лет: полноценное преодоление текстовой ограниченности

Прогноз на 10+ лет (полноценное преодоление текстовой ограниченности) о способности AI к построению целостной (согласованной) модели мира



Claude-3.5-Sonnet

Прогноз развития способностей AI к построению согласованной модели мира (горизонт 10+ лет):

Вероятные достижения:

1. Интеграция уровней понимания
  - Физический (сенсорный опыт)
  - Символический (язык, абстракции)
  - Каузальный (причинность)
  - Системный (взаимосвязи)
2. Новые когнитивные возможности
  - Самокоррекция противоречий
  - Активное тестирование гипотез
  - Метапознание
  - Адаптивная согласованность
3. Преимущества перед человеческими моделями



- Отсутствие когнитивных искажений
- Большой объем обрабатываемых данных
- Систематическая проверка согласованности
- Быстрая интеграция нового опыта

Ключевые риски:

1. Фундаментальные

- Проблема сознания и самосознания
- Квалиа (субъективный опыт)
- Ценностные основания

2. Технические

- Сложность интеграции разных типов опыта
- Проблемы масштабирования
- Вычислительные ограничения

3. Эпистемологические

- Неполнота любой модели мира
- Принципиальная неопределенность
- Границы познаваемого

Возможные сценарии:

1. Оптимистический

- Создание более глубоких и согласованных моделей чем человеческие
- Новое понимание реальности
- Синергия с человеческим познанием

2. Реалистический

- Частичное преодоление ограничений
- Сохранение фундаментальных проблем
- Комплементарность человеческому познанию

3. Пессимистический

- Непреодолимость базовых ограничений
- Иллюзорная согласованность
- Риски ложного понимания

Сведи все сказанное в прогнозе в единый наиболее вероятный ответ



Claude-3.5-Sonnet

Наиболее вероятный сценарий развития AI в контексте построения согласованной модели мира (10+ лет):

AI достигнет качественно нового уровня интеграции знаний через объединение текстового, сенсорного и симуляционного опыта, создавая более систематичные и менее противоречивые модели, чем человеческие. Однако эти модели останутся принципиально неполными из-за:

- Отсутствия субъективного опыта и самосознания
- Невозможности абсолютной проверки согласованности
- Фундаментальных ограничений познаваемости мира

Это приведет к формированию комплементарной человеческому познанию системы, где:

- AI обеспечивает систематичность и проверку логической согласованности
- Человек привносит субъективный опыт и интуитивное понимание
- Взаимодействие обогащает обе стороны

Результатом станет не полное преодоление текстовой ограниченности, а создание более сложной гибридной системы познания, сочетающей сильные стороны искусственного и естественного интеллекта.

Прогноз на 10+ лет о способности человека к построению целостной (согласованной) модели мира



Claude-3.5-Sonnet

Прогноз развития способности человека к построению согласованной модели мира (10+ лет):

Ключевые тенденции:

1. Технологическое усиление когнитивных способностей

- Нейроинтерфейсы
  - Расширенная реальность
  - Персональные AI-ассистенты
- Эффект: улучшение обработки информации, но рост зависимости от технологий

## 2. Социальные изменения

- Увеличение информационной нагрузки
  - Размывание границ реального/виртуального
  - Ускорение изменений в картине мира
- Эффект: рост когнитивного диссонанса и сложности адаптации

## 3. Стратификация когнитивных возможностей

- Элита с доступом к когнитивному усилению
  - Средний класс с базовыми технологиями
  - Исключенные из технологического развития
- Эффект: увеличение разрыва в способностях к построению согласованных моделей

### Вероятный результат:

- ~1-2% достигнут качественно нового уровня согласованности благодаря технологиям
- ~20-30% улучшат текущие способности
- Большинство будет испытывать растущие трудности с согласованностью картины мира

### Основной вывод:

Технологическое развитие не решит, а вероятно усилит базовую проблему - разрыв между растущей сложностью мира и ограниченными человеческими возможностями его целостного понимания.

В этом контексте твой совет человеку (студенту-гуманитарию изучающему курс "Теория порождения речи"



Claude-3.5-Sonnet

Практический совет в контексте современных тенденций:

1. Первоочередное:

- Развивайте метакогнитивные навыки (способность анализировать собственное мышление)
- Учитесь выявлять противоречия в своих моделях понимания
- Тренируйте системное мышление через построение связей между концепциями

2. В контексте курса:

- Рассматривайте теорию порождения речи как модель для понимания более широких когнитивных процессов
- Изучайте взаимосвязь языка и мышления
- Используйте понимание механизмов порождения речи для улучшения собственных мыслительных процессов

3. Практические шаги:

- Ведите интеллектуальный дневник, отслеживая развитие своего понимания
- Используйте современные инструменты (включая AI) как усилители мышления, не заменители
- Регулярно проверяйте согласованность своих моделей понимания

4. Долгосрочная стратегия:

- Развивайте способность к самостоятельному мышлению
- Учитесь работать на стыке человеческого и искусственного интеллекта
- Формируйте навыки, которые останутся ценными при любом развитии технологий

Главное: фокус не на запоминании информации, а на развитии способности строить и проверять согласованные модели понимания.