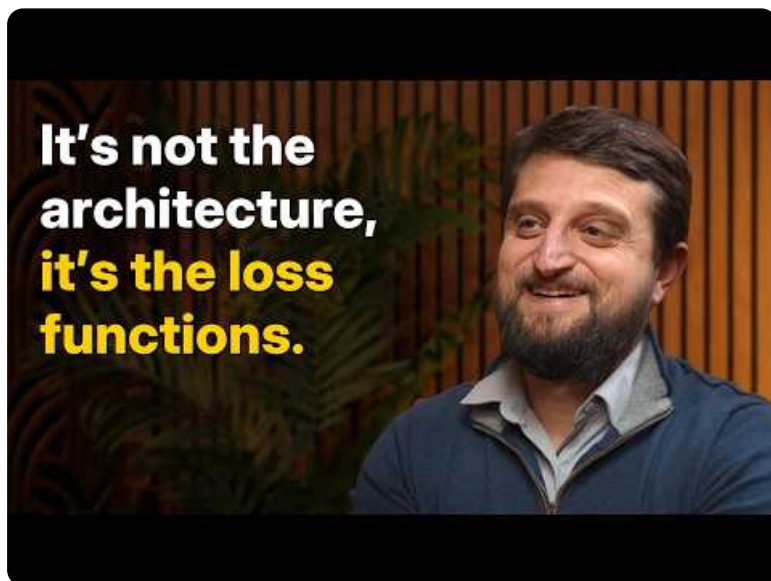


Adam Marblestone - AI is missing something fundamental about the brain

@DwarkeshPatel · 2025-12-30 · https://www.youtube.com/watch?v=_9V_Hbe-N1A



Выпуск посвящен фундаментальным различиям между современными нейросетями и биологическим мозгом. В беседе с Дваркешем Пателем Адам Марблстоун — физик, нейробиолог и генеральный директор организации Convergent Research — разбирает, почему большие языковые модели требуют колоссальных объемов данных, но все еще уступают человеку в гибкости обобщения и эффективности обучения.

Центральный тезис Марблстоуна: искусственный интеллект упирается в ограничения не столько из-за нехватки вычислительных мощностей, сколько из-за отсутствия ключевых биологических принципов. Мозг использует всенаправленный вероятностный вывод и опирается на разделение между универсальной обучающейся неокортикальной системой и древней подкорковой управляющей системой. Эволюция зашила в подкорку тысячи специализированных функций потерь, которые курируют процесс обучения в онтогенезе. Для воспроизведения этих принципов в кремнии необходим качественный технологический скачок в нейронауке — прежде всего масштабная оптическая коннектомика и новые форматы организации фундаментальных исследований.

Архитектура мозга против LLM: всенаправленный вывод и эволюционные функции потерь

Недооцененная роль эволюционно сложных функций потерь 00:00: В машинном обучении традиционно используются предельно простые целевые

функции — предсказание следующего токена или минимизация кросс-энтропии. В биологическом мозге эволюция создала сложнейший ансамбль динамических функций потерь, активирующихся в разных зонах коры на различных стадиях онтогенеза. Геном кодирует не жесткую разводку синапсов, а детальный алгоритмический учебный план (curriculum), определяющий, что именно и в какой последовательности должна усваивать каждая структура. Без понимания этих биологических функций потерь невозможно объяснить феноменальную sample efficiency (эффективность обучения на малом объеме данных) человеческого мозга по сравнению с современными LLM.

Фактчек: Мозг использует тысячи врождённых функций потерь, всенаправленный вывод и строгое разделение Learning/Steering — □: Это исследовательская рамка, а не установленная архитектура мозга; модель Стива Бырнса и «оценщики мыслей» не подтверждены как единая экспериментальная схема. Подкорковые структуры тоже обучаются, а генетические программы задают не только «учебный план», но и организацию связей.

Кора как универсальный движок всенаправленного вывода (omnidirectional inference) 02:39: Архитектура шестислойной коры кардинально отличается от однонаправленного прямого прохода трансформеров. В то время как LLM вычисляет конкретную условную вероятность следующего токена на основе предшествующего контекста, неокортекс функционально представляет собой вероятностную систему всенаправленного вывода. Любая зона коры оптимизирована под то, чтобы восстанавливать произвольное подмножество неизвестных переменных на основе любого другого наблюдаемого подмножества входных данных. Подобный механизм концептуально близок к энергетическим моделям (Energy-Based Models), продвигаемым Яном Лекуном, где система оценивает совместное распределение всех переменных и способна семплировать недостающие признаки при фиксации произвольных параметров.

Предсказание внутренних состояний подкорки как базис абстрактного мышления 05:25: Ассоциативные зоны коры занимаются не только предсказанием мультимодальных сенсорных сигналов (например, видео из аудио), но и прогнозированием состояний примитивного «рептильного» мозга и вегетативных реакций тела. Неокортекс непрерывно учится предсказывать, произойдет ли сокращение мышцы, скачок пульса, смех или рефлекторное отдергивание руки. Это позволяет связывать высокоуровневые семантические стимулы (например, услышанную фразу «у тебя на спине паук») с теми же базовыми защитными контурами, которые активируются при физическом зрительном контакте с реальной угрозой.

Модель Бырнса: разделение на обучающуюся (Learning) и управляющую (Steering) подсистемы

Парадокс кодирования высокоуровневых целей в эволюции 06:27:

Эволюция не могла предвидеть концепции вроде социального статуса в академической среде, критики со стороны авторитетных ученых или подкастов, однако мозг генерирует в ответ на социальные ошибки врожденные реакции стыда и желания спрятаться. Для согласования абстрактных моделей мира с древними инстинктами в архитектуре мозга существует строгое разделение: Обучающаяся подсистема (Learning Subsystem — кора, часть миндалины) строит модель мира, а Управляющая подсистема (Steering Subsystem — гипоталамус, ствол мозга) задает базовые цели, рефлексy и сигналы вознаграждения. Задача неокортекса — не просто моделировать окружающую среду, но и моделировать саму Управляющую подсистему, находя абстрактные паттерны, которые предсказывают активацию врожденных эвристик.

Механизм «оценщиков мыслей» (Thought Assessors) и генерализация подкорковых эвристик 09:17:

Управляющая подсистема обладает собственными рудиментарными сенсорными каналами — например, верхними двухолмиями среднего мозга (superior colliculus), способными автономно детектировать быстрые контрастные движения или базовые паттерны лиц без участия зрительной коры. Когда субкортикальный сенсор фиксирует опасность, гипоталамус генерирует негативное подкрепление и команду вздрагивания. Специализированные группы нейронов в миндалине или фронтальной коре (Thought Assessors по терминологии Стива Бырнса) выполняют роль предиктивных классификаторов: они за сотни миллисекунд предсказывают, сработает ли этот врожденный рефлекс. Поскольку классификатор получает на вход богатые репрезентации коры, он обобщает угрозу от конкретного темного пятна до абстрактного концепта или слова «паук», транслируя сигнал вознаграждения вниз по иерархии.

Ограниченность современных подходов к мультимодальности в ИИ 15:19:

Попытки обучить искусственные нейросети всенаправленному выводу через удаление масок или простое сопоставление пар видео, аудио и текста наталкиваются на негибкость зафиксированных латентных пространств. Если модель обучена заполнять пропуск в фиксированном месте, она не всегда способна вывести произвольную переменную в произвольной конфигурации на этапе инференса. Кроме того, современные архитектуры могут страдать от неверных индуктивных смещений в кодировщиках: биологические сенсоры производят колоссальную предварительную фильтрацию данных (например, детекция движения на уровне сетчатки или инварианты поверхностей и окклюзий в зрительной коре, исследованные Дорис Цао), задавая латентное пространство, оптимизированное под быстрое обучение.

Амортизированный вывод, байесовское сэмплирование и геномное бутылочное горлышко

Амортизированный вывод против прямого вероятностного семплирования

22:20: В классическом байесовском подходе восприятие требует итеративного перебора гипотез о скрытых причинах наблюдаемых данных через ресурсоемкие методы Монте-Карло. Глубокие нейросети заменяют этот процесс амортизированным выводом: в процессе обучения оптимизируется прямой проход от наблюдения сразу к наиболее вероятной причине. Однако сложные когнитивные задачи требуют вычислений на этапе инференса (test-time compute). В LLM это выражается в цепочках рассуждений (Chain of Thought), представляющих собой дискретное семплирование гипотез, результаты которого со временем дистиллируются обратно в веса базовой модели для выполнения в один прямой проход.

Разрешение парадокса информационной емкости генома **27:26:**

Человеческий геном содержит всего около 3 гигабайт данных, из которых лишь малая доля отведена под нервную систему, что исключает прямую запись синаптических весов. В рамках концепции Бырнса этот парадокс разрешается тем, что геном кодирует универсальный алгоритм обучения с гиперпараметрами слоев для обучающейся подсистемы и компактный набор алгоритмических правил (эквивалентных нескольким тысячам строк кода на Python) для управляющей подсистемы. Задавая лишь врожденные рефлексии и архитектуру Thought Assessors, эволюция перекладывает задачу наполнения модели мира фактами на сам онтогенетический опыт организма.

Фактчек: Человеческий геном содержит около 3 ГБ данных — Δ : Гаплоидный геном содержит около 3 млрд пар оснований — примерно 0,75–0,8 ГБ при двухбитном кодировании; около 3 ГБ получается при текстовом хранении по байту на основание. Рецензенты расходятся в строгости оценки — от неточности до ошибки, — но согласны с пересчётом; отсутствие прямого списка синаптических весов не доказывает конкретную модель «генома-учебного плана».

Фактчек: Управляющую подсистему мозга можно описать несколькими тысячами строк Python — $\square$: Это иллюстративная аналогия, не измерение сложности. Ни размер генома, ни клеточные атласы не позволяют вывести такой объём достаточной программы.

Фактчек: Мозг потребляет около 20 Вт, а нейроны работают на частоте около 200 Гц — Δ : Оценка энергопотребления разумна, но 200 Гц — высокая частота для многих корковых нейронов, а не типичный средний режим; средняя активность часто лежит примерно в диапазоне 0,1–10 Гц. Использование клеточного шума для вероятностного семплирования — гипотеза о вычислениях, а не доказанное объяснение работы мозга в целом.

Клеточное разнообразие как маркер жестко запрограммированных

контуров 30:50: Данные транскриптомных атласов отдельных клеток (single-cell RNA sequencing) программы BRAIN Initiative показывают резкую асимметрию между структурами мозга. В неокортексе наблюдается относительно небольшое разнообразие базовых типов нейронов, формирующих повторяющиеся модульные вычислительные колонки. Напротив, в управляющей подсистеме (гипоталамус, ствол мозга) обнаружены тысячи редких, уникальных и узкоспециализированных клеточных типов. Каждое уникальное поведение или функция потерь (детекция соленого вкуса, реакция на быстро приближающийся объект, сигналы насыщения) требует экспрессии уникальных молекулярных рецепторов для генетически детерминированной сборки синаптических цепей без предварительного обучения.

Фактчек: Гипоталамус и ствол мозга разнообразнее коры по типам нейронов —

□: Одноклеточные атласы поддерживают эту общую закономерность; в атласе мозга мыши выделено около 5300 транскриптомных кластеров, значительная часть разнообразия приходится на подкорковые структуры. Однако соответствие «один тип клеток — одно врождённое поведение или функция потерь» из этих данных не следует.

Эволюция мозга: масштабирование коры, приматы и социальное обучение

Взрывной рост мозга гоминид через повышение эффективности выборки

34:05: Резкое увеличение объема неокортекса у предков человека не требовало масштабной перестройки генома: фундаментальные схемы кортикальных колонок уже присутствовали у грызунов и ранних млекопитающих.

Эволюционный сдвиг произошел за счет небольших мутаций, изменивших параметры масштабирования количества нейронов (согласно работам Сузаны Эркулано-Эузель, мозг приматов пакует нейроны плотнее, чем мозг грызунов) и расширивших макроструктурные тракты между слуховыми, моторными и ассоциативными зонами (зоны Брока и Вернике). Главным драйвером стала обратная связь: появление социальных инстинктов и передача знаний от старейшин резко повысили sample efficiency отдельной особи, сделав энергетически выгодным содержание массивной коры.

Фактчек: Плотность нейронов и социальное обучение объясняют увеличение мозга человека — ▲: Различия в масштабировании подтверждены: при сопоставимой

большой массе мозг приматов содержит больше нейронов, чем предсказывает закономерность для грызунов. Но социальное обучение как главный двигатель увеличения коры и белая склера как специальная адаптация для считывания взгляда — эволюционные гипотезы, а не однозначно установленные причины.

Специализация сенсорных систем под социальные функции потерь 40:05:

В эволюции приматов кортикальные области, отвечающие за зрение и слух,

масштабировались непропорционально сильнее по сравнению с обонятельными зонами. Причина кроется не просто в физических законах масштабирования сигналов, а в природе социальных функций потерь управляющей подсистемы: обучение языку и перенимание культурных навыков критически зависят от визуального считывания мимики, жестов и совместного взгляда (joint attention). Морфология человеческого глаза с контрастной белой склерой развилась специфически для высокоточного трекинга направления взгляда сородичей, выступая врожденным интерфейсом калибровки обучения в первые два года жизни.

Обучение с подкреплением (RL): почему LLM используют примитивные алгоритмы по сравнению с мозгом

Примитивность методов RL в современных языковых моделях 42:42:

Обучение с подкреплением в современных фронтирных LLM концептуально упрощено: если сгенерированная траектория проходит модульный тест или решает задачу, вероятность каждого токена в этой цепочке просто увеличивается пропорционально. В системах отсутствуют явные функции ценности (value functions), которые успешно применялись еще 10 лет назад в DeepMind при игре в Atari через Q-learning. Игнорирование временного дисконтирования (temporal difference) делает оптимизацию контекстных генераций вычислительно неэффективной с точки зрения использования промежуточных шагов.

Фактчек: В RL для современных LLM отсутствуют функции ценности, а temporal difference означает временное дисконтирование — □: PPO использует обучаемого критика; GRPO и некоторые другие методы обходятся без отдельной модели ценности, поэтому обобщение на все LLM неверно. TD-обучение обновляет оценки через разность последовательных предсказаний, тогда как дисконтирование уменьшает вес отдалённых наград — это разные понятия.

Фактчек: AlphaProof достиг уровня серебряной медали IMO 2024 — △: Результат 28 из 42 баллов относится к совместной работе AlphaProof и AlphaGeometry 2, а не к AlphaProof в одиночку — здесь ревью Gemini требует поправки. Сам принцип обучения поиску доказательств с проверяемым вознаграждением в Lean подтверждён.

Многоуровневое обучение с подкреплением в нервной системе 44:58: Мозг реализует сразу несколько параллельных систем обучения с подкреплением. Базальные ганглии и полосатое тело (striatum) используют безмодельное обучение (model-free RL) с ограниченным пространством действий — от переключения моторных программ до когнитивного шлюзования (разрешение таламусу пропустить сигнал из одной зоны коры в другую). Дофаминовые нейроны транслируют не абсолютную награду, а ошибку предсказания награды (Reward Prediction Error, RPE), что математически эквивалентно обучению функциям ценности в TD-learning.

Фактчек: Дофамин кодирует ошибку предсказания награды, математически эквивалентную TD-learning — Δ : Связь активности многих дофаминовых нейронов с ошибкой предсказания награды хорошо подтверждена и составляет основу влиятельных TD-моделей. Но дофаминовые сигналы неоднородны, существуют альтернативные объяснения, поэтому полное отождествление биологии с одним алгоритмом слишком сильное.

RL как вероятностный вывод в неокортексе 46:37: Неокортекс реализует модельное обучение с подкреплением (model-based RL), выступая единой средой, где функции ценности и ожидаемые награды интегрированы в общее вероятностное представление мира. В рамках парадигмы «RL как вероятностный вывод» (RL as inference) задача планирования решается фиксацией переменной желаемой высокой награды в распределении и последующим сэмплированием цепочки гипотетических действий и промежуточных состояний, ведущих к этому результату.

[48:19-50:32] Аналогия культурной эволюции по Джозефу Генриху с безмодельным RL на уровне цивилизации и фиксация накопленных знаний в традициях.

Биологическое «железо» против кремния и клеточная сложность

Аппаратные компромиссы: медленные стохастические элементы против быстрого кремния 50:32: Биологический мозг ограничен энергопотреблением в ≈ 20 Вт и частотой срабатывания нейронов порядка 200 Гц при минимальных рабочих напряжениях. Платой за энергоэффективность является невозможность копирования весов и отсутствие произвольного доступа на чтение/запись параметров извне. Однако мозг выигрывает за счет полного совмещения памяти и вычислений (in-memory computing), неструктурированной разреженности активаций и естественной стохастичности мембранных каналов, позволяющей генерировать выборки вероятностных распределений на аппаратном уровне без программных генераторов случайных чисел.

Внутриклеточные вычисления: алгоритмическая необходимость или биологический костыль 53:44: Большинство сложных молекулярных каскадов внутри нейрона служат не отдельными вычислительными слоями памяти, а системами биологической поддержки синаптической пластичности при отсутствии центрального процессора. Например, перенос информации в ядро клетки необходим для регуляции транскрипции при нормализации синаптических весов. Исключением выступают специфические структуры, такие как клетки Пуркинье в мозжечке, где тело нейрона способно напрямую хранить временные задержки между условным и безусловным раздражителями на биохимическом уровне без создания громоздких рекуррентных синаптических петель. В целом синаптическая парадигма остается достаточным алгоритмическим базисом для моделирования интеллекта.

Выравнивание (Alignment), понятийный аппарат нейронаук и пределы интерпретируемости

Сложность управляющей подсистемы как ключевой барьер безопасности

ИИ 57:00: С точки зрения безопасности суперинтеллекта (AI alignment), минимальный набор требований к управляющей подсистеме для создания функционального агента существенно меньше набора, необходимого для воспроизведения человеческой морали и эмпатии. Чтобы агент научился строить космические корабли и постигать физику, ему достаточны базовые драйверы любопытства и исследования среды. Если эти драйверы не содержат тонко настроенных социальных функций потерь, формировавшихся у человека миллионы лет, результатом работы станет агент вроде «максимизатора скрепок», обладающий колоссальными способностями без этических ограничений.

Проблема теоретического языка в описании работы мозга 1:00:12:

Существует методологический раскол между исследователями, пытающимися интерпретировать мозг через призму концепций машинного обучения (сверточные сети, backprop, actor-critic), и сторонниками радикального нейробиологического подхода (такими как Дьёрдь Бужаки). Последние настаивают, что психологические термины и аналогии из ML некорректны, а истинными примитивами мозга являются эндогенные паттерны динамических осцилляций нейронных ансамблей. Тем не менее искусственные нейросети остаются наиболее точными предиктивными моделями для предсказания нейронных активаций (бенчмарк Brain-Score лаборатории Джима ДиКарло), что свидетельствует о конвергенции вычислительных принципов.

Границы механистической интерпретируемости: правила против весов

1:04:31: Попытки найти конкретные «цепи понятий» (например, контур «моста Золотые Ворота») внутри биологического коннектома или матрицы весов LLM обречены на частичный провал из-за колоссальной сложности выученных представлений. По аналогии с работой Конрада Кёрдинга и Тима Лилликрапа, истинное понимание системы заключается не в каталогизации значений миллиардов весов, а в описании фундаментальных принципов: топологии архитектуры, локальных правил синаптической пластичности, распределения инициализации весов и динамики функций потерь. Задача коннектомики — не прочесть статическую память мозга, а вскрыть жесткие ограничения на архитектуру и правила обучения.

Коннектомика нового поколения: оптические методы, проект E11 Bio и таймлайны AGI

Сроки создания AGI и практическая актуальность коннектомики 1:08:10:

Исследования структуры мозга теряют прямую применимость для индустрии ИИ

в сценарии коротких сроков достижения AGI (к 2027 году), поскольку фронтальные лаборатории ориентируются исключительно на масштабирование трансформеров на GPU. Однако на горизонте 10 лет текущая парадигма может упереться в непреодолимое плато по эффективности выборки и архитектурной гибкости. В этом случае наличие полных аппаратных карт мозга млекопитающих предоставит альтернативные инженерные чертежи для архитектур нового поколения.

Технологическая революция E11 Bio: переход от электронной к оптической микроскопии 1:12:02: Стандартная коннектомика на базе просвечивающей электронной микроскопии требует нарезки ткани на ультратонкие слои, разрушает органические молекулы и оценивается отчётами Wellcome Trust в несколько миллиардов долларов на один мозг мыши. Проект E11 Bio разрабатывает метод оптической коннектомики, снижающий стоимость картирования мозга мыши до 10–30 млн долларов. Использование фотонов вместо электронов позволяет применять флуоресцентную маркировку антителами, создавая «молекулярно аннотированный коннектом» — карту, фиксирующую не только геометрию контактов, но и типы синапсов, нейромедиаторов и рецепторов в каждом синаптическом соединении.

Фактчек: E11 Bio снижает стоимость коннектома мыши с миллиардов до 10–30 млн долларов — Δ : Миллиардные оценки электронной коннектомики и разработка более дешёвых оптических методов имеют основание, но десятки миллионов — целевой бюджет, а не подтверждённая стоимость уже завершённой карты. Gemini оценил тезис как подтверждённый; итоговая оценка осторожнее, поскольку технологическая цель не равна достигнутому результату.

Уроки Human Genome Project: падение стоимости через экспоненциальное распараллеливание 1:14:40: Первый геном человека стоил около 3 млрд долларов (≈ 1 доллар за пару оснований), но институциональная политика NHGRI и запуск технологической конкуренции снизили стоимость секвенирования в миллион раз за 10 лет. Ключом стал переход от макрохимического разделения к параллельному считыванию миллионов кластеров ДНК на оптических сенсорах. Проекты класса Focused Research Organizations (FRO) воспроизводят этот паттерн в коннектомике: приоритет отдаётся не лобовому картированию ткани на старом оборудовании, а созданию масштабируемой технологической платформы.

Фактчек: Human Genome Project стоил около 3 млрд, после чего секвенирование подешевело в миллион раз за десять лет — Δ : Около 3 млрд — бюджет всей программы, а не себестоимость секвенирования одного генома. Сопоставимый ряд NHGRI показывает падение примерно со 100 млн в 2001 году до 1000 в 2015-м: около 100 тысяч раз за 14 лет, а не миллион за десять.

Аугментация обучения ИИ сигналами активности живого мозга 1:19:42: Вместо попыток дистиллировать мозг целиком через сложные

нейрокомпьютерные интерфейсы перспективен подход с введением вспомогательной функции потерь: обучение модели предсказывать паттерны активности зрительной коры человека одновременно с классификацией изображений. Это навязывает искусственной сети геометрию представлений и индуктивные смещения биологического зрения, устраняя уязвимость к состязательным атакам (adversarial examples) и радикально снижая объем необходимых синтетических размеченных данных.

Фактчек: Обучение на активности зрительной коры устраняет adversarial-уязвимости и радикально сокращает потребность в данных — Δ : Биологически ориентированные ограничения улучшали устойчивость в отдельных экспериментах, но не устранили состязательные атаки. Масштаб выигрыша в данных требует конкретного исследования и условий сравнения; общего результата такой силы нет.

Формальная математика, язык Lean и верифицируемый ИИ

Lean и RLVR: алгоритмическая революция в доказательстве теорем

1:23:39: Язык интерактивного доказательства теорем Lean переводит математические рассуждения в верифицируемый код, истинность которого проверяется компилятором автоматически без участия человека. Это превращает формальную математику в идеальную среду для обучения с подкреплением с проверяемым вознаграждением (RLVR — Reinforcement Learning with Verifiable Rewards). Проекты вроде AlphaProof и стартапа Harmonic подтверждают, что в замкнутых системах с мгновенной верификацией поиск доказательств автоматизируется аналогично поиску победных стратегий в шахматах и Го.

Фактчек: Lean проверяет истинность, а LLM устраняют проблему спецификаций — Δ : Ядро Lean проверяет формальное доказательство относительно заданного утверждения и аксиом. Можно корректно доказать не то свойство, которое действительно нужно: LLM помогают формализовать требования, но не гарантируют соответствия спецификации человеческому намерению.

Проблема формализации гипотез и метрики содержательности

математики 1:26:23: Механический поиск доказательств оставляет нерешенной задачу генерации содержательных математических гипотез и понятийного структурирования теорий. Формализация хорошей гипотезы связана с колмогоровской сложностью: ценная теорема представляет собой компактное утверждение, служащее узлом сжатия для сотен следствий, радикально сокращая длину логического вывода к другим утверждениям. Без формулирования специализированной функции потерь для поиска глубоких сжатых объяснений ИИ рискует генерировать тривиальные или бессмысленные верифицируемые тавтологии.

Спецификация как фундаментальный барьер формальной верификации

софта 1:28:40: Инструменты Lean и верифицируемого программирования

позволяют математически доказывать надежность программного кода критической инфраструктуры (энергосети, финансовые протоколы, защита памяти от эксплойтов). Главное препятствие внедрения — проблема формулирования спецификаций (spec problem): инженеры способны написать код, но не обладают формальным математическим аппаратом для перевода требований безопасности на строгий язык доказательств. Развитие LLM устраняет этот разрыв, выступая транслятором между естественным языком технических требований и формальными логическими инвариантами.

Фактчек: Сон консолидирует память и защищает от забывания — □: Участие сна и воспроизведения нейронной активности в консолидации памяти хорошо подтверждено. Однако обучение происходит и во время бодрствования, replay не ограничен сном, а формулировки «кора обновляется только ночью» и «недосып заставляет перезаписывать старые знания» необоснованно категоричны.

Влияние автоматизации на абстрактное мышление: аналогия с переходом от ассемблера 1:31:56: Опасения, что передача доказательства рутинных лемм системам ИИ приведет к деградации математической интуиции, повторяют исторические дискуссии вокруг отказа от ручного написания ассемблерного кода в пользу высокоуровневых языков. Устранение механической рутины ускоряет циклы обратной связи, позволяя исследователям сфокусироваться на высокоуровневой концептуальной топологии задач. Подобно тому как Стив Бирнс (физик по образованию) смог синтезировать модели работы подкорковых структур без многолетней практики работы с микропипеткой в лаборатории, формальные системы ИИ откроют доступ к переднему краю теоретической физики и математики ученым со стороны.

Символический суперинтеллект и математически доказуемая безопасность агентов 1:35:39: В сценариях взаимодействия миллиардов автономных агентов AGI единственным надежным средством защиты от когнитивных атак, манипуляций и скрытого вредоносного кода станет требование строгой математической верификации каждого шага коммуникации. Программа Safeguarded AI (руководимая David D. G. в агентстве ARIA) разрабатывает модели мира, выраженные не в скрытых состояниях нейронов, а в явных верифицируемых дифференциальных уравнениях. Автоматизированные доказательные системы позволяют вернуться к интерпретируемым символическим методам на новом технологическом базисе.

Фактчек: Формальная верификация — единственная надёжная защита автономных агентов — □: Она даёт строгие гарантии только относительно спецификации, модели среды и принятых предположений. Наличие программы ARIA Safeguarded AI не доказывает ни исключительности подхода, ни возможности математически исключить любые манипуляции и ошибки описания реального мира.

Символические репрезентации, непрерывное обучение и инфраструктура науки (Gap Map)

Когнитивные карты против чисто символических регистров 1:38:25: В мозге отсутствуют классические аппаратные регистры с операциями копирования переменных (variable binding), характерные для цифровых компьютеров. Вместо символических систем мозг масштабирует геометрические принципы навигации: нейроны места (place cells) и решетчатые клетки (grid cells) гиппокампа грызунов формируют пространственные карты, которые у приматов абстрагируются в многомерные «когнитивные карты» для навигации по семантическим концептам, социальным иерархиям и логическим категориям.

Механизмы непрерывного обучения (continual learning) и системная консолидация 1:41:07: В современных нейросетях обучение замораживается после завершения тренировки во избежание катастрофического забывания (catastrophic forgetting). Мозг решает эту задачу через механизм системной консолидации: гиппокамп выполняет функцию быстрого адаптивного буфера, записывающего последовательности событий дня, а во время сна воспроизводит (replay) эти паттерны на высокой скорости, постепенно обучая неокортекс и встраивая новую информацию в долговременную память без разрушения старых представлений.

Фактчек: Без биологических принципов ИИ неизбежно упрётся в предел, а любопытство без социальных целей создаст «максимизатора скрепок» — □: Необходимость копирования мозга для общего интеллекта не установлена, как и сроки предполагаемого плато. Разрыв между способностями и безопасными целями — содержательный риск, но конкретный катастрофический исход не следует автоматически из отсутствия человеческих социальных механизмов.

Внимание в биологических сетях: кортико-таламические петли и быстрые веса 1:42:40: Архитектурный аналог разделения параметров и активаций (векторов ключей и значений) в трансформерах представлен в мозге взаимодействием коры и таламуса. Таламус выполняет роль адаптивного шлюза, регулирующего информационные потоки между различными зонами неокортекса на основе контекста задачи. Сочетание быстрых форм синаптической пластичности (изменение проводимости за миллисекунды) с медленными структурными изменениями синапсов обеспечивает функционал, эквивалентный механизмам внимания с динамически обновляемыми ключами.

Карта инфраструктурных разрывов (Gap Map) и новая организация науки 1:44:49: Анализ узких мест передовой науки показывает, что сотни фундаментальных направлений заблокированы не отсутствием теоретических идей, а нехваткой масштабируемой инженерной инфраструктуры — аналогов «миниатюрных космических телескопов Хаббл». Традиционные академические лаборатории с грантами на аспирантов не способны решать задачи

скоординированной системной инженерии, а венчурный капитал не готов финансировать проекты без быстрой коммерциализации. Создание целевых исследовательских организаций (FRO) с бюджетами в десятки миллионов долларов закрывает ключевые инфраструктурные разрывы (от верификаторов математики до оптических коннектов), разблокируя прогресс для мирового научного сообщества.

ЛИЧНО ДЛЯ ТЕБЯ

Проверяй требования раньше AI-кода: До генерации запиши разрешённые действия, запрещённые состояния и поведение при повторном запросе или частичном сбое. Проси модель искать контрпримеры к требованиям, а не только писать тесты: автоматически созданная проверка может повторить ошибку реализации, а тесты свойств не заменяют формального доказательства.

Не оплачивай обучение недосыпом: После интенсивного освоения нового или музыкальной практики сохраняй полноценный сон, вместо того чтобы регулярно добавлять занятия за его счёт. Это обоснованная поддержка памяти, но не универсальная формула «час сна всегда полезнее часа практики».

Оценивай AI по результатам, а не по прогнозам AGI: Для решений о подписках, автоматизации и собственном времени измеряй долю успешно завершённых задач, стоимость результата и частоту вмешательств. Ни «масштабирование решит всё», ни «без копирования мозга прогресса не будет» пока не годятся как надёжная основа финансового решения.

Не превращай нейроанalogии в личные протоколы: Из модели Learning/Steering не следует, что привычки закрепляются только через немедленное телесное подкрепление. Для тезисов вроде «дофамин — TD-ошибка» или «таламус — attention» полезно спрашивать: что измерено, какой эксперимент отличает объяснение от альтернатив и где аналогия перестает работать?

КОНСЕНСУС РЕЦЕНЗЕНТОВ

Все три рецензента сходятся в поправках к объёму генома и обобщению об отсутствии функций ценности; Claude и GPT также согласны, что нейробиологическая регуляризация не устраняет adversarial-уязвимости. Расхождения главным образом в строгости: Gemini чаще принимает биологические интерпретации и технологические цели как подтверждённые, GPT подчёркивает гипотетичность центральной модели мозга, а Claude добавляет ограничения по числам, срокам и физиологии.