БОРИС КРИГЕР

HOMO CREDENS

ВЕРУЮЩИЙ ВИД

ALTASPERA

Опубликовано издательством Altaspera Publishing .

Борис Кригер — междисциплинарный философ, занимающийся вопросом о том, как разрозненные области знаний могут быть объединены в целостное видение человеческого существования. В своих работах он стремится преодолеть разделение философии и науки, этики и политики, индивидуального опыта и коллективных структур. Объединяя идеи экзистенциализма, социальной теории, когнитивной науки и технологических исследований, он разрабатывает способ мышления, который не является ни редукционистским, ни утопическим, а открыт сложности современного мира.

Homo Credens : Верующий вид

Почему мы верим больше, чем можем доказать? Почему память нас обманывает? Почему самые сложные системы искусственного интеллекта «галлюцинируют» ложную информацию?

В книге «Homo Credens» Борис Кригер раскрывает глубокую истину о природе сложных умов: любая достаточно сложная система — будь то человеческий мозг, когнитивные способности животных или искусственный интеллект — должна верить гораздо больше, чем может подтвердить. Это не ограничение, которое нужно преодолеть, а архитектура, которую нужно понять.

Опираясь на три научные статьи, Кригер показывает, как компромисс между сложностью и достоверностью влияет на всё — от памяти и восприятия до рассуждений и искусственного интеллекта. По мере усложнения систем их способность проверять утверждения линейно растёт, в то время как пространство утверждений, которые они должны рассматривать, растёт экспоненциально. Неизбежный результат: сложные системы должны придерживаться определённых убеждений без проверки.

Последствия распространяются от философии до практических технологий. Полная рациональность не сложна, но невозможна. Даже квантовые компьютеры не могут избежать компромисса между сложностью и достоверностью. Галлюцинации, связанные с ИИ, невозможно устранить, не устранив генеративный потенциал, который делает ИИ полезным.

Мы долгое время называли себя Homo sapiens — знающим видом. В этой книге предлагается более точное название: Homo Credens — верящим видом. Мы не знатоки, которые лишь изредка доверяют, а доверяющие, которые лишь изредка проверяют. И, понимая это, мы делаем первый шаг к тому, чтобы верить по-настоящему.

Credens» — это глубокое переосмысление человеческого разума и его места во вселенной сложных систем, которое изменит ваше представление о знаниях, памяти, рациональности и о том, что значит верить.

Ключевые слова

Убеждение, сложность, память, проверка, познание, рациональность, доверие

Содержание

Предисловие

Мы привыкли считать веру противоположностью знания, а проверку — золотым стандартом познания. Эта книга показывает, что верно обратное: вера — это основа, на которой вообще становится возможным знание. Чем сложнее система, тем больше в ней должно быть веры. Это относится не только к людям, но и к любой сложной системе — животным , искусственному интеллекту и даже гипотетическим технологиям будущего.

Название *Homo Credens* — «верующий вид» — не призвано заменить *Homo sapiens* , а раскрыть истинное значение слова *sapiens* . Мы издавна прославляли человеческую мудрость, нашу способность познавать. Однако каждый акт познания основывается на более глубокой способности: способности доверять тому, что нельзя проверить. Мы доверяем своим чувствам, хотя они могут нас обмануть. Мы доверяем своей памяти, хотя она скорее восстанавливает, чем воспроизводит. Мы доверяем своей логике, хотя она основана на

аксиомах, которые сами по себе не могут быть доказаны. Мы доверяем другим, хотя они могут ввести нас в заблуждение. Без этого фундамента доверия — этой *уверенности* — никакое знание было бы невозможно вообще.

Эта книга не только о человеческой психологии. Изложенные здесь аргументы применимы к любой достаточно сложной системе, будь то эволюционировавшая или искусственно созданная. Бактерия может подтвердить почти всё, что имеет отношение к её выживанию; она существует в области прямых доказательств. Млекопитатель этого сделать не может; он должен подтвердить утверждения о хищниках, источниках пищи и социальных союзниках, которые выходят за рамки непосредственной проверки. Искусственный интеллект сталкивается с тем же ограничением: как только система становится достаточно сложной, чтобы моделировать свой мир, она переходит в область достоверности.

Представленные здесь аргументы основаны на трех взаимосвязанных исследовательских работах, которые вместе образуют единую теорию когнитивной архитектуры. *Эволюционная теория доверия* (Кригер, 2022) устанавливает, что сложные системы должны придерживаться утверждений, превышающих их возможности проверки, — не как недостаток, а как структурная необходимость. Гипотеза об *атемпоральности пространства ментальной памяти* (Кригер, 2025) демонстрирует, что временной порядок не является неотъемлемой частью хранения памяти, а возникает в процессе операций извлечения — объясняя, почему мы неправильно помним не как неудачу, а как следствие оптимального проектирования. А *эволюционный отбор на атемпоральное хранение памяти* (Кригер, 2019) показывает, почему естественный отбор сходится к таким архитектурам под действием трех независимых факторов: эффективности использования ресурсов, скорости извлечения и адаптивной гибкости.

Некоторые читатели могут задаться вопросом: смогут ли будущие технологии — возможно, квантовые компьютеры — наконец достичь полной верификации? В этой книге утверждается, что ответ — нет. Ограничение не технологическое, а математическое. Сложность растет комбинаторно; возможности верификации растут максимум полиномиально . Никакая архитектура, какой бы мощной она ни была, не сможет преодолеть этот разрыв для систем, достаточно сложных, чтобы представлять интерес. Квантовый компьютер, моделирующий достаточно сложный мир, столкнется с тем же компромиссом, что и любая сложная система: он будет верить больше, чем может проверить.

Книга приглашает читателя в путешествие по спектру сложности, от простейших организмов до самых совершенных машин, показывая, как доверие — вера в убежденность, не подлежащую проверке, — является ценой за вход в мир сложного адаптивного поведения. Это путешествие имеет практическое

значение: для нашего понимания памяти и ее искажений, для проектирования искусственных систем и интерпретации их «галлюцинаций», для нашего осмысления рациональности и ее пределов, и, в конечном итоге, для нашего понимания самих себя.

Здесь читатель не найдет формул или технического жаргона. Идеи, хотя и основаны на формальных исследованиях, представлены посредством объяснений и метафор, доступных любому вдумчивому читателю. Требуется не специализированные знания, а готовность переосмыслить предположения о знании и вере, о памяти и времени, о том, что значит быть разумом в мире, слишком сложном для проверки на практике.

Глава 1: Парадокс познающего разума

Задайте себе простой вопрос: насколько то, во что вы верите прямо сейчас, вы можете проверить на практике ? Не позже, не в принципе, а прямо сейчас, в этот момент, используя имеющиеся в вашем распоряжении ресурсы.

Вы верите, что стул под вами выдержит ваш вес. Вы верите, что пол простирается за пределы вашего поля зрения. Вы верите, что солнце взойдет завтра. Вы верите, что ваши воспоминания о вчерашнем дне приблизительно точны. Вы верите, что слова на этой странице означают то, что вы думаете. Вы верите, что автор существует, что эта книга была написана, а не сгенерирована случайным образом, что изложенные здесь идеи связаны с традицией человеческой мысли, уходящей корнями в тысячелетия.

Ни одно из этих убеждений нельзя проверить в данный момент. Чтобы убедиться, что стул выдержит, вам нужно понимать свойства его материала, физику

несущей конструкции, текущее распределение напряжений по его компонентам. У вас нет доступа к этой информации. Вы просто верите, что стул выдержит, потому что стулья, как правило, выдерживают. Чтобы проверить свои воспоминания о вчерашнем дне, вам понадобится независимая запись для сравнения — но любая запись, к которой вы можете обратиться, сама по себе является представлением, которому вы должны доверять.

В этом и заключается парадокс, лежащий в основе того, что мы называем знанием: существо, называющее себя « мудрым» — *Homo sapiens* — не может подтвердить большую часть того, что, как оно утверждает, знает. И это не случайное ограничение, которое можно преодолеть с помощью более совершенных технологий или более тщательного мышления. Это структурная особенность любого достаточно сложного разума.

Традиционная реакция на это наблюдение — скептицизм: поскольку мы не можем проверить свои убеждения, возможно, нам следует воздержаться от

суждений, воздерживаясь от принятия решений до тех пор, пока не будет достигнута уверенность. Но такая реакция неверно понимает ситуацию. Проблема не в том, что мы не смогли проверить свои убеждения из-за лени или небрежности. Проблема в том, что сама проверка требует ресурсов — времени, энергии , внимания — которые принципиально ограничены, в то время как пространство вещей, в которые нам, возможно, придется верить, расширяется безгранично.

Подумайте, что на самом деле требуется для проверки . Чтобы проверить утверждение — убедиться, что оно соответствует реальности — необходимо собрать доказательства, сравнить их с вашим утверждением, оценить, совпадают ли они. Каждый из этих шагов занимает время. Каждый потребляет когнитивные ресурсы. Каждый требует внимания, которое можно было бы посвятить чему-то другому. В мире бесконечного времени и неограниченных когнитивных возможностей проверка могла бы быть бесплатной. В мире, в

котором мы живем на самом деле , она всегда обходится дорого.

Между тем, количество вещей, в которые нам, возможно, приходится верить, множится без видимого предела. Рассмотрим решения, с которыми вы сталкиваетесь в обычный день: что съесть, что надеть, кому доверять, какой маршрут выбрать, какие задачи заслуживают приоритета, как интерпретировать выражения лиц других людей, был ли услышанный вами звук угрожающим или безобидным. Каждое решение основано на убеждениях о мире — убеждениях о питании, погоде, характере, дорожной ситуации, срочности, эмоциональном состоянии, опасности. Количество убеждений, влияющих даже на простые решения, огромно.

Если бы вам приходилось проверять каждое убеждение, прежде чем действовать, вы бы никогда ничего не сделали. К тому времени, как вы убедились бы, что стул выдержит ваш вес, разговор бы закончился. К тому времени, как вы вспомнили, где

припарковали машину, день бы уже прошел. К тому времени, как вы проверили бы все необходимое для принятия одного сложного решения, ваша жизнь бы закончилась.

Это не недостаток человеческого познания. Это математическая необходимость. Пространство возможных убеждений комбинаторно растет с увеличением сложности системы, в которую они верят. Каждая новая концепция умножает количество возможных утверждений. Каждое новое соотношение между концепциями умножает их еще больше. Разум, достаточно сложный для ориентации в сложном мире, должен рассматривать больше утверждений, чем он когда-либо сможет проверить.

Последствия этого глубоки: сложные умы должны функционировать прежде всего на основе того, что мы могли бы назвать *доверием* — приверженностью утверждениям, которые выходят за рамки непосредственной проверки. Это не вера в смысле религиозной веры или идеологического убеждения, хотя она имеет общие структурные черты

с обоими понятиями. Это основная когнитивная операция, заключающаяся в том, чтобы считать что-либо истинным для целей действия и вывода, не проверив предварительно, что это так.

Мы делаем это постоянно, незаметно, автоматически. Когда вы тянетесь за чашкой, вы верите в то, что чашка находится там, где кажется, что она будет вести себя так, как ведут себя чашки, что ваша рука подчинится вашему намерению. Когда вы разговариваете с другим человеком, вы верите в то, что он вас слышит, понимает вашу речь, что он примерно истолкует ваши слова так, как вы задумали. Когда вы планируете завтрашний день, вы верите в утверждения о непрерывности вашего «я», стабильности мира, надежности ваших собственных намерений.

Некоторые философы рассматривают эту повсеместную веру как скандал, который нужно преодолеть. Если бы мы были более осторожны, более строги, более привержены проверке, мы могли бы достичь подлинного знания. Но такой диагноз

совершенно неверно понимает ситуацию. Проблема не в недостаточной строгости. Проблема в том, что идеал проверки применим только к простым системам, работающим в простых средах. Для сложных систем в сложных средах вера не является отступлением от надлежащего познания. Она и есть надлежащее познание.

В этой книге разрабатывается теория, объясняющая, почему это должно быть так — не только для людей, но и для любой сложной системы. Теория опирается на три взаимосвязанные исследовательские программы. *Эволюционная теория доверия* (Кригер, 2022) закладывает ресурсно-теоретические основы: почему сложность неизбежно превышает возможности проверки. Анализ *вневременности пространства ментальной памяти* (Кригер, 2025) показывает одно важное следствие: сама память — это система, основанная на доверии, которая конструирует, а не извлекает прошлое. А *эволюционный отбор в пользу вневременного хранения памяти* (Кригер, 2019) демонстрирует,

почему естественный отбор сходится к таким архитектурам — не вопреки их «ошибкам», а благодаря их преимуществам.

В совокупности эти анализы предполагают фундаментальное переосмысление того, что такое разум и что он делает. Мы не машины для проверки, которые лишь изредка прибегают к вере, когда проверка не удаётся. Мы — машины для веры, которые лишь изредка достигают проверки, когда позволяют условия. Мы не *Homo sapiens* , которые иногда доверяют. Мы — *Homo Credens* — вид, верящий, — которые иногда проверяют.

Понимание этого не является поводом для отчаяния. Это начало мудрости. Скептик, требующий проверки всего, не понимает ни проверки, ни веры. Догматик, полностью отказывающийся от проверки, утратил способность к калибровке , которая делает веру полезной. Мудрый верующий — а в этой книге утверждается, что мудрость означает стать вдумчивым верующим — понимает, когда проверка возможна и полезна, когда доверие необходимо и

уместно, и как соотносить свои обязательства с доказательствами и последствиями.

Наше исследование начинается с самого основного вопроса: что значит что-либо проверять, и почему проверка требует ресурсов, которые всегда конечны?

Глава 2: Проблема верификации

Что значит проверить утверждение? На первый взгляд ответ кажется очевидным: проверить — подтвердить, установить, что что-то действительно истинно . Но за этим простым ответом скрывается сложная реальность. Проверка — это не одна операция, а целый комплекс процессов, каждый из которых имеет свои требования и ограничения.

Рассмотрим утверждение «на столе стоит чашка». Чтобы проверить это, вы можете посмотреть на стол. Если вы видите чашку, вы можете объявить утверждение верным. Но что вы сделали на самом деле ? Вы собрали сенсорную информацию — свет, отраженный от стола и предметов на нем, обработанный вашей зрительной системой в представление — и сравнили это представление с вашим понятием «чашка» и вашим понятием «на столе». Совпадение между представлением и понятиями дает вам уверенность в истинности утверждения.

Но обратите внимание, на скольких непроверенных предположениях основана эта проверка. Вы предполагаете, что ваша зрительная система функционирует правильно. Вы предполагаете, что свет, достигающий ваших глаз, точно отражает предметы в комнате. Вы предполагаете, что ваши понятия «чашка» и «стол» соответствуют реальным категориям в мире. Вы предполагаете, что момент, когда вы посмотрели, является репрезентативным — что чашка не исчезла в тот же миг, как вы моргнули. Каждое из этих предположений в принципе можно проверить. Но проверка требует дополнительных предположений, которые, в свою очередь, потребуют дополнительных проверок в регрессионном анализе, не имеющем естественной точки остановки.

Это первый урок о проверке: она всегда относительна и основана на множестве непроверенных предположений. Чтобы проверить утверждение A, необходимо предположить утверждения B, C и D. Эти утверждения сами по себе

могли бы быть проверены, но только путем предположения других. Проверка никогда не достигает фундамента. Она всегда плывет по морю доверия.

Некоторые философы ответили на этот регресс фундаментализмом: утверждением, что должны существовать некие базовые убеждения, которые самоподтверждаются и не требуют дальнейшего подтверждения. В качестве кандидатов часто предлагаются перцептивные убеждения — «Мне кажется, я вижу чашку» не может быть ошибочным так же, как «там есть чашка». Но этот ответ смешивает несомненность с проверкой. Даже если в некоторых убеждениях нельзя сомневаться, это не означает, что они были проверены на соответствие реальности. Это означает лишь то, что сам механизм проверки их одобрил. Этот механизм может быть надежным или ненадежным; мы не можем выйти за его пределы, чтобы проверить, какой именно.

Второй урок, касающийся проверки, заключается в том, что она требует ресурсов. Каждый

акт проверки требует времени: времени на сбор доказательств, времени на их обработку, времени на сравнение с проверяемым утверждением. Каждый акт проверки требует энергии: метаболических затрат на восприятие, внимание и познание. Каждый акт проверки требует пропускной способности: ограниченной пропускной способности каналов сенсорных систем и рабочей памяти.

Эти затраты могут показаться незначительными для простых утверждений, таких как «на столе стоит чашка». Одного взгляда достаточно, чтобы всё осмыслить. Но рассмотрим более сложные утверждения: «этот человек заслуживает доверия», «эта инвестиция надёжна», «эта память точна». Проверка таких утверждений — если это вообще возможно — требует длительного наблюдения, множества источников доказательств и тщательного анализа альтернативных вариантов. Затраты возрастают пропорционально сложности утверждения.

Хуже того, затраты масштабируются особенно проблематичным образом. Простые утверждения часто можно проверить с помощью простых наблюдений. Сложные утверждения требуют сложных процедур проверки — процедур, которые сами по себе включают множество шагов, каждый из которых может оказаться неудачным или ввести в заблуждение. Проверка сложных утверждений не просто дороже, чем проверка простых; она непропорционально дороже.

Рассмотрим проверку утверждения «экономика улучшится в следующем году». Проверить это заранее невозможно; нужно дождаться следующего года. Для проверки задним числом необходимо определить, что значит «улучшится», собрать экономические данные, сравнить данные за разные периоды времени и оценить, подтверждает ли сравнение утверждение об улучшении. Каждый шаг включает в себя выбор, влияющий на результат. Различные определения, различные источники данных, различные методы сравнения могут привести

к различным выводам. Полная проверка — проверка , которая однозначно решает вопрос, — может быть недостижима даже в принципе.

Третий урок, касающийся проверки, заключается в том, что ресурсы, затраченные на неё, недоступны для других целей. Время, потраченное на проверку утверждения, — это время, не затраченное на его реализацию. Энергия, затраченная на сбор доказательств, — это энергия, недоступная для использования открывающихся возможностей. Внимание, сосредоточенное на проверке, отвлекается от всего остального. Проверка сопряжена с альтернативными издержками, и эти издержки могут быть существенными.

В условиях конкуренции альтернативные издержки могут иметь решающее значение. Организм, который слишком долго проверяет, не является ли тень хищником, становится добычей хищника. Агент, который настаивает на проверке каждого предположения перед действием, теряет возможности в пользу агентов, которые действуют,

основываясь на разумной степени доверия. Существует оптимальное количество проверок, и это не «максимально возможное количество». Это «достаточное количество, чтобы откалибровать степень доверия в соответствии с поставленными задачами».

Оптимальное количество проверок варьируется в зависимости от обстоятельств. В стабильных условиях, где одни и те же ситуации повторяются предсказуемо, проверка прошлых данных может уменьшить необходимость в проверке текущих данных — можно полагаться на то, что было изучено ранее. В новых условиях может потребоваться больше проверок. Когда ставки высоки, дополнительная проверка может быть оправдана, несмотря на её стоимость. В условиях сильного временного давления даже ценная проверка может оказаться невозможной.

Четвертый урок о верификации заключается в том, что она требует инфраструктуры. Для проверки утверждений о мире необходимы датчики для сбора

информации, процессоры для ее анализа, память для хранения промежуточных результатов и механизмы сравнения для оценки совпадений. Сама по себе эта инфраструктура дорогостояща в создании и поддержании. Биологические организмы развили сложные сенсорные и когнитивные системы; искусственные системы необходимо проектировать и производить. Ни один из этих процессов не является бесплатным.

Кроме того, инфраструктура верификации ограничена. У вас есть лишь ограниченное количество сенсорных рецепторов, ограниченная вычислительная мощность, ограниченный объем рабочей памяти. Эти ограничения определяют, что можно проверить и как быстро. Даже при неограниченном времени и мотивации ваши возможности верификации имеют предел.

Теперь мы можем увидеть проблему верификации во всей ее полноте. Верификация (1) всегда предполагает непроверенные фоновые предположения, (2) требует времени, энергии и

внимания, (3) конкурирует с другими способами использования этих ресурсов и (4) ограничена возможностями инфраструктуры верификации. Для любой конечной системы верификация является дефицитным ресурсом. Вопрос не в том, следует ли экономить на верификации — все конечные системы должны экономить — а в том, как экономить разумно.

В следующей главе показано, почему эта экономия становится все более серьезной по мере усложнения систем. Проблема верификации не статична; она обостряется по мере расширения пространства утверждений, которые могут нуждаться в проверке. А для сложных систем это пространство расширяется взрывным образом.

Глава 3: Взрыв сложности

Представьте себе простой организм — например, бактерию, — ориентирующуюся в окружающей среде. Ее мир — это химический суп, и ее задача — двигаться к питательным веществам и удаляться от токсинов. Вопросы, имеющие отношение к этой задаче, ограничены: выше ли концентрация питательного вещества A здесь, чем мгновение назад? Увеличивается или уменьшается концентрация токсина B? Сенсорный аппарат бактерии может ответить на эти вопросы напрямую через химические рецепторы. Для этой простой системы пропускная способность системы проверки примерно соответствует пространству соответствующих вопросов.

Теперь рассмотрим более сложный организм — скажем, мышь. Мышь должна ориентироваться в пространственной среде, распознавать объекты, отслеживать других существ, запоминать места, где находятся пища и опасность, предвидеть будущие события. Вопросов, имеющих отношение к

выживанию мыши, гораздо больше, чем вопросов, имеющих отношение к выживанию бактерий. Где находится пища? Где находится хищник? Безопасно ли это место? Это существо — друг или враг? Куда ведет домой? Что здесь происходило раньше? Каждый из этих вопросов разветвляется на множество подвопросов, и ответы на них взаимодействуют сложным образом.

Разница между бактерией и мышью иллюстрирует общий принцип: по мере усложнения систем пространство утверждений, которые они должны рассматривать, расширяется не линейно, а комбинаторно. Это взрыв сложности, и понимание его имеет решающее значение для понимания того, почему сложные системы должны полагаться на достоверность.

Что движет комбинаторным разложением? Рассмотрим три фактора: концепции, взаимосвязи и время.

Во-первых, понятия. Система с n понятиями может формировать утверждения, включающие

любое из этих понятий. Число возможных простых утверждений (приписывающих объекту свойство) линейно возрастает с n. Но многие важные утверждения включают отношения между понятиями: «кот на коврике», «ключ открывает дверь», «Джон любит Мэри». Число возможных бинарных отношений растет пропорционально квадрату n. Тройные отношения (A дал B C) растут пропорционально кубу. Отношения более высокого порядка — а человеческое мышление, безусловно, включает их — растут еще быстрее.

Во-вторых, отношения между утверждениями. Сложное мышление включает в себя не только отдельные утверждения , но и структуры, построенные на их основе: условные утверждения (если A, то B), дизъюнкции (A или B), конъюнкции (A и B), отрицания (не A). При наличии n базовых утверждений количество возможных составных утверждений экспоненциально возрастает с допустимой глубиной вложенности. Система, способная рассуждать о принципе «если не A, то (B и

C)», может представлять экспоненциально больше возможностей, чем система, ограниченная простыми утверждениями.

В-третьих, время. Система, рассматривающая только настоящий момент, сталкивается с управляемым набором утверждений о текущих состояниях. Система, рассуждающая о будущем, должна рассматривать возможные последовательности событий. Если на каждом временном шаге существует k возможных событий, и система рассматривает n временных шагов вперед, то количество возможных последовательностей равно k^n — экспоненциально зависящему от глубины планирования. Шахматист, просчитывающий пять ходов вперед, сталкивается с гораздо большим количеством возможностей, чем тот, кто просчитывает три хода вперед, даже несмотря на то, что разница в глубине составляет всего два.

Взаимодействие этих факторов приводит к поистине взрывному росту сложности . Система, обладающая множеством концепций, способная

формировать сложные взаимосвязи и рассуждать в течение длительных временных горизонтов, сталкивается с пространством утверждений, которое растет быстрее любой полиномиальной функции от ее основных параметров. Это математическое проявление истинной сложности.

Теперь рассмотрим, что происходит с проверкой в этом контексте. Пропускная способность проверки — скорость, с которой утверждения могут быть проверены на соответствие доказательствам, — ограничена пропускной способностью сенсорных систем, скоростью обработки и доступным временем. Эти величины растут в лучшем случае линейно с выделенными на них ресурсами. Можно добавить больше сенсоров, более быстрые процессоры, больше времени. Каждое добавление обеспечивает линейное увеличение пропускной способности проверки.

Однако пространство утверждений растет комбинаторно. Это несоответствие имеет решающее значение. Независимо от того, сколько возможностей для проверки вы добавите, если пространство

утверждений растет комбинаторно, а возможности для проверки растут линейно, то отношение проверенных утверждений к непроверенным приближается к нулю по мере увеличения сложности. Сложные системы неизбежно имеют охват проверки — долю релевантных утверждений, которые они могут фактически проверить , — который пренебрежимо мал.

Давайте нагляднее рассмотрим пример. Предположим, вы можете проверять одно утверждение в секунду. За день вы можете проверить 86 400 утверждений. За год — около 31 миллиона. За всю жизнь — возможно, 2-3 миллиарда утверждений, что является существенным числом. Но теперь представьте себе пространство утверждений взрослого человека, способного к сложным концепциям, комплексному рассуждению и длительному планированию. Количество утверждений, которые могут иметь отношение к вашим жизненным решениям — утверждения о людях, местах, возможностях, истории, будущем,

абстрактных отношениях — вероятно, превышает количество атомов в видимой Вселенной. Ваша способность к проверке за всю жизнь не сможет существенно повлиять на это.

Это не просто спор о больших числах. Он имеет качественное следствие. Когда охват проверки ничтожно мал, система не может полагаться на проверку как на основной способ работы. Ей необходим другой способ — тот , который позволяет принимать утверждения без проверки. Этот другой способ — доверие.

Эволюционная *теория доверия* (Кригер, 2022) формализует эту взаимосвязь. Покрытие проверки можно аппроксимировать как отношение пропускной способности проверки (скорости проверки) к доступному времени, деленное на размер пространства утверждений. По мере усложнения систем пространство утверждений растет комбинаторно, в то время как пропускная способность проверки растет максимум линейно. Отношение стремится к нулю. Это не случайный факт

биологической эволюции или современных технологий. Это математическая необходимость.

Некоторые могут возразить, что этот анализ преувеличивает проблему. Безусловно, сложным системам не нужно рассматривать все возможные утверждения — только релевантные. А интеллектуальные системы могут сосредоточиться на наиболее важных утверждениях, проверяя их выборочно, а не исчерпывающе.

Это возражение указывает на важную истину, но не позволяет избежать взрыва сложности. Да, сложные системы могут сосредоточить свои ресурсы на проверке наиболее важных утверждений. Но определение того, какие утверждения являются «наиболее важными», само по себе является суждением, которое необходимо вынести. На каком основании? Если основанием является проверка, мы имеем дело с регрессом: необходимо проверить, какие утверждения следует проверять. Если основанием является что-то иное, помимо проверки — интуиция, эвристика, априорные убеждения, —

тогда мы уже признали, что доверие играет фундаментальную роль. Система использует непроверенные суждения о важности, чтобы направлять свою проверенную проверку конкретных утверждений .

Более того, даже «релевантные» утверждения могут быть чрезвычайно многочисленными для сложных систем. Что имеет значение при принятии решения о том, доверять ли кому-то? Его прошлое поведение, его мотивы, его отношения с вами, его отношения с другими вашими знакомыми, его репутация, контекст текущего взаимодействия, альтернативы, доступные вам обоим — и этот список едва затрагивает поверхность. Каждый фактор разлагается на подфакторы. Пространство релевантных утверждений, хотя и меньше, чем общее пространство утверждений, остается комбинаторно большим.

Взрывной рост сложности — это не проблема, которую нужно решить, а ограничение, которое нужно преодолеть. Сложные системы не могут

избежать его; они могут лишь разрабатывать стратегии для работы в его рамках. Фундаментальная стратегия — это доверие: принятие утверждений, основанных на чем-то ином, чем прямая проверка. В следующей главе рассматривается, как доверие возникает как структурная необходимость для любой системы, достаточно сложной, чтобы столкнуться с проблемой проверки.

Глава 4: Рождение Кредо

Что происходит, когда проверка не удаётся — не время от времени, а систематически, не как исключение, а как правило? Система всё равно должна действовать. Организм всё равно должен ориентироваться в окружающей среде, находить пищу, избегать хищников, размножаться. Разум всё равно должен принимать решения, строить планы, координировать свои действия с другими. Действие не может ждать проверки, которая никогда не произойдёт.

Это рождение доверия: момент, когда система принимает решения не потому, что они были проверены, а потому, что это необходимо, а проверка невозможна. Доверие — это не провал проверки, а реакция на невозможность достаточной проверки. Именно так поступают сложные системы, когда им приходится действовать в мире, который они не могут полностью проверить.

Представьте себе газель на африканской саванне. Движение в высокой траве привлекает её внимание. Хищник ли это? Безобидное животное? Игра ветра? Газель не может это подтвердить. Для подтверждения потребовалось бы приблизиться, исследовать, собрать больше доказательств — а если это хищник, то подтверждение приходит слишком поздно. Газель должна принять решение: либо воспринять движение как опасное и убежать, либо воспринять его как безобидное и продолжить пастись. Это решение, принятое без подтверждения, является доверием.

Газель, которая всегда убегает, тратит энергию на ложные тревоги. Газель, которая никогда не убегает, становится добычей. Естественный отбор устанавливает порог доверия: сколько доказательств достаточно, чтобы оправдать побег? Ответ не «достаточно доказательств для проверки», потому что проверка невозможна в отведенное время. Ответ — «достаточно доказательств, чтобы оправдать

приверженность, учитывая издержки и выгоды от ошибки в каждом направлении».

Этот пример иллюстрирует несколько особенностей доверия как когнитивной стратегии. Во-первых, доверие формируется обстоятельствами, а не выбирается из альтернатив. Газель не выбирает между проверкой и доверием; она не может проверить, поэтому должна доверять. Во-вторых, доверие калибруется последствиями, а не только доказательствами. Порог принятия решения зависит от того, что произойдет, если вы ошибетесь. В-третьих, доверие может быть лучше или хуже откалибровано. Газель, которая убегает при каждой тени, плохо откалибрована; так же плохо откалибрована и та, которая игнорирует очевидные угрозы.

Человеческая система доверия более сложна, но структурно схожа. Когда вы доверяете другу, вы делаете предположения о его характере, намерениях и надежности, не имея возможности проверить их напрямую. Когда вы принимаете научное открытие,

вы делаете предположения о методологии, данных и интерпретации, не повторяя эксперименты самостоятельно. Когда вы верите своим воспоминаниям о вчерашнем разговоре, вы делаете предположения о том, что было сказано, без возможности проверить это на основе независимых данных.

В каждом случае необходимы обязательства, а проверка невозможна или нецелесообразна. Вы не можете проверить сокровенные черты характера своего друга; вы можете лишь наблюдать за его поведением и делать выводы. Вы не можете повторить каждый научный эксперимент; вы можете лишь оценить экспертность и послужной список. Вы не можете проверить воспоминания на соответствие реальности; вы можете лишь проверить их внутреннюю согласованность и целостность с другими воспоминаниями.

Эволюционная *теория доверия* (Кригер, 2022) определяет это как универсальную особенность сложных систем. Как указано в этой работе, «для

любой системы, где размер пространства утверждений значительно превышает пропускную способность проверки, умноженную на доступное время, существуют различия, которые имеют отношение к адаптивному функционированию системы и не могут быть проверены этой системой в момент принятия решения». Это утверждение не относится конкретно к психологии человека; это утверждение относится к структуре самой сложности.

В этом понимании доверие — это не вера в противоположность знанию. Это фундаментальная познавательная операция, которая делает возможными как веру, так и знание. Без способности принимать непроверенные утверждения система не смогла бы формировать гипотезы для проверки, не смогла бы накапливать доказательства для выводов, не смогла бы функционировать в мире достаточно долго, чтобы что-либо проверить. Доверие — это фундамент; проверка — это случайные достижения, построенные на нём.

Это переворачивает традиционную эпистемологическую иерархию. Философия часто рассматривала знание как норму, а веру — как недостаточную замену: «то, на чём вы соглашаетесь». Когда знания недоступны. Но анализ сложности предполагает обратное: вера (уверенность) является нормой, режимом работы по умолчанию для любой сложной системы. Знание — подтвержденная вера — это редкое достижение, возможное только в ограниченных областях, где можно сосредоточить ресурсы для проверки.

Подумайте о бесчисленном множестве утверждений, на которые вы полагаетесь в повседневной жизни. Вы верите, что пол выдержит ваш вес, что еда, которую вы едите, не отравлена, что окружающие вас люди реальны, а не являются сложными симуляциями, что законы физики будут продолжать действовать так же, как и раньше. Ни одно из этих убеждений не подтверждено в сколько-нибудь значимом смысле; все они — это догмы , без которых вы не смогли бы функционировать. Ваши

«знания» конкретных фактов — год вашего рождения, название вашего города, содержание вчерашних новостей — плавают в этом море неподтвержденных догм.

Что делает доверие рациональным, а не произвольным? Если мы не можем проверить свои обязательства, как отличить разумное доверие от неразумного? Ответ связан с калибровкой, согласованностью и историей. Хорошо откалиброванное доверие присваивает уровни уверенности, соответствующие фактической надежности: вы должны быть более уверены в утверждениях, которые чаще оказываются верными . Согласованное доверие формирует последовательную сеть обязательств: утверждения не должны противоречить друг другу или своим последствиям. Доверие, чувствительное к истории, учится на опыте: если определенные типы доверия оказываются ненадежными, следует внести соответствующие корректировки.

Обратите внимание, что ни один из этих критериев не требует проверки отдельных утверждений. Калибровку, согласованность и чувствительность к истории работы можно оценить без проверки конкретного содержания убеждений. Система может обладать хорошо откалиброванной степенью доверия — присваивая соответствующие уровни уверенности — даже если большинство её конкретных убеждений остаются непроверенными. Именно так становится возможным рациональное убеждение в мире, где проверка является редкостью.

Таким образом, рождение доверия — это не падение из какого-то рая верификации. Это появление когнитивных способностей, которые делают возможным сложное адаптивное поведение. Каждая сложная система — биологическая или искусственная — должна решить проблему верификации. Решение — доверие: обязательство, структурированное не верификацией, а калибровкой , согласованностью и усвоенной надежностью. Это решение не является необязательным; оно архитектурно необходимо.

Глава 5: Прогнозирующее решение

Как сложные системы на самом деле реализуют доверие? Какова вычислительная архитектура, позволяющая принимать непроверенные утверждения, сохраняя при этом адекватную калибровку? Ответ, который был найден в нейробиологии и когнитивной науке за последние десятилетия, — это предсказание. Сложные мозги — это машины предсказания, постоянно генерирующие ожидания относительно того, что произойдет дальше, и обновляющие эти ожидания, когда реальность их удивляет.

Эта предсказательная архитектура — механизм доверия. Когда вы предсказываете, что чашка окажется там, где вы за ней потянетесь, вы подтверждаете своё утверждение, не проверяя его заранее. Когда вы предсказываете, что друг будет вести себя в соответствии со своим характером, вы подтверждаете свои утверждения о его психологии. Когда вы предсказываете, что за прочитанным вами словом последует слово, имеющее грамматический

смысл, вы подтверждаете свои утверждения о языке. Каждое предсказание — это акт доверия — подтверждение до проверки.

Мозг, обладающий способностью предсказывать будущее, не ждет сенсорных данных, чтобы сформировать картину мира. Вместо этого он начинает с модели — набора ожиданий о том, каким является мир, — а затем использует сенсорные данные для уточнения и корректировки этой модели. То, что мы называем «восприятием», в значительной степени является наилучшим предсказанием мозга о причинах получаемой сенсорной информации. То, что мы называем «действием», — это попытка мозга воплотить свои предсказания в жизнь.

Данная архитектура элегантно решает проблему верификации. Чистая система верификации должна была бы проверять каждое утверждение перед принятием решения — что невозможно при ограниченных ресурсах. Система прогнозирования сначала подтверждает результат, а затем проверяет его, сопоставляя прогнозы с результатами и

корректируя их при расхождении. Верификация, если таковая имеется, осуществляется посредством ошибок прогнозирования: несоответствий между ожидаемым и фактическим опытом, которые сигнализируют о необходимости пересмотра модели.

Рассмотрим зрительное восприятие. Наивные теории рассматривают зрение как процесс построения изображения из сенсорных данных: свет попадает на сетчатку, обнаруживаются узоры, распознаются объекты. Но этот восходящий процесс сам по себе слишком медленный и подверженный влиянию шума, чтобы объяснить скорость и точность зрительного восприятия. Вместо этого мозг предсказывает, что он ожидает увидеть, основываясь на контексте, предыдущем опыте и текущих целях. Визуальная обработка затем сводится к проверке этих предсказаний на соответствие поступающим сигналам — уделяя внимание ошибкам предсказания, игнорируя при этом ожидаемый входной сигнал.

Это объясняет, почему вы можете читать текст с чем угодно. сообщение vwls : ваш мозг

предсказывает, какие слова должны появиться, исходя из контекста, и проверяет, соответствуют ли доступные буквы этим предсказаниям. Это объясняет, почему вы можете не заметить опечатку в знакомом тексте: предсказание преобладает над сенсорными данными. Это объясняет оптические иллюзии: когда предсказание и ощущения вступают в конфликт, предсказание иногда побеждает.

Предсказательная архитектура выходит далеко за рамки восприятия. Управление движениями основано на прогнозах последствий движений. Социальное познание основано на прогнозах поведения и психического состояния других людей. Понимание языка основано на прогнозах будущих слов и значений. Планирование основано на прогнозах результатов возможных действий. В каждой области мозг принимает решения, основывается на прогнозах — проявляет доверие — а затем использует обратную связь для калибровки.

Эволюционная *теория доверия* (Кригер, Б. (2022). Эволюционная теория доверия:

концептуальная основа с формальными аналогиями для понимания генеративного моделирования как ресурсо-теоретического следствия сложности. Zenodo . https://doi.org/10.5281/zenodo.18379476) связывает эту предсказательную архитектуру с проблемой верификации. В статье вводится так называемый «принцип предсказательной жизнеспособности»: для адаптивных систем с сенсомоторной задержкой, работающих в изменяющейся среде, внутренние состояния должны кодировать предсказательную информацию о будущих состояниях, выходящую за рамки того, что предоставляют текущие наблюдения. Проще говоря: если существует задержка между восприятием и действием, и мир меняется в течение этой задержки, чистая реакция на текущие данные всегда будет слишком запоздалой. Требуется предсказание.

Это имеет глубокие последствия. Нейронные сигналы распространяются с конечной скоростью — примерно 100 метров в секунду максимум. Для организма значительных размеров сенсорная

информация о настоящем уже относится к прошлому к тому моменту, когда она достигает центров принятия решений. Животное, реагирующее только на текущие данные, всегда будет реагировать на мир, которого больше не существует. Предсказание — это не роскошь, а необходимость для любого организма, который должен координировать свое поведение во времени и пространстве.

Предсказательное решение показывает, как доверие реализуется на практике в реальных когнитивных системах. Вместо хранения проверенных фактов и их извлечения по мере необходимости, мозг поддерживает предсказательные модели, которые генерируют ожидания по требованию. Эти модели кодируют то, во что система «верит» в функциональном смысле: утверждения, которые она будет воспринимать как истинные, ожидания, которые направляют поведение, обязательства, которые структурируют взаимодействие с миром.

Важно отметить, что прогностические модели являются генеративными. Они не просто хранят закономерности из прошлого опыта ; они генерируют новые закономерности, которые никогда ранее не встречались. Это источник воображения, творчества и формирования гипотез, но также и ошибок, галлюцинаций и ложных убеждений. Генеративная система, способная создавать новые прогнозы, выходящие за рамки обучающих данных, иногда будет выдавать ложные прогнозы. Это не ошибка; это цена генеративности.

Таким образом, прогностическое решение выявляет глубокую связь между доверием, прогнозированием и генерацией. Система, которая должна подтверждать свои данные после проверки, разрабатывает прогностические модели. Прогностические модели являются генеративными. Генеративные модели создают как полезные экстраполяции, так и случайные ошибки. Эти ошибки не являются сбоями в и без того надежной системе;

это неизбежная цена системы, разработанной для гибкости, а не просто для точности.

Глава 6: Вера важнее доказательств

В предыдущих главах утверждалось , что сложные системы должны опираться прежде всего на доверие, а не на проверку. В этой главе аргумент углубляется, показывая, что даже сама проверка — даже доказательство и рассуждение — основывается на доверии. От веры нет выхода к чистому знанию; вера распространяется на все уровни.

Рассмотрим, что требуется для доказательства . Чтобы доказать утверждение, необходимо исходить из посылок, применять правила вывода и выводить заключения. Но что обосновывает посылки? Либо они сами доказаны — что требует наличия более ранних посылок, влекущих за собой бесконечный регресс, — либо они принимаются без доказательства. Регресс должен где-то остановиться, и где бы он ни остановился, мы обнаружим утверждения, принятые на веру, а не на доказательство.

Философы по-разному называли эти точки остановки: аксиомы, первые принципы, основные

убеждения, фундаментальные предположения. Названия различаются, но структура остается той же. Каждая система доказательства основывается на утверждениях, которые сами по себе не доказываются в рамках этой системы. Это не случайное ограничение, которое могли бы преодолеть более совершенные системы; это структурная особенность того, что представляет собой доказательство.

Та же закономерность наблюдается и с правилами вывода. Для корректного рассуждения необходимо следовать правилам, сохраняющим истинность от посылок к выводам. Но что обосновывает эти правила? Можно попытаться доказать, что правила сохраняют истинность, — но любое такое доказательство само по себе будет использовать правила вывода, порождая замкнутый круг. Нельзя обосновать логику с помощью логики, не предполагая того, что вы пытаетесь доказать.

Это не скептический парадокс, который нужно разрешить или обойти. Это простое наблюдение о структуре обоснования. Доказательство действует в

рамках принятых предпосылок и правил. Эта структура сама по себе не доказывается ; ей доверяют. Вера предшествует доказательству в логическом порядке обоснования.

Эволюционная *теория доверия* выделяет несколько категорий такого фундаментального доверия. Во-первых, это доверие к самой логике: предположение, что обоснованный вывод сохраняет истину, что противоречия невозможны, что правила рассуждения отражают нечто реальное в мире. Во-вторых, это доверие к восприятию: предположение, что чувственный опыт предоставляет информацию о внешней реальности, что мир существует за пределами нашего восприятия его. В-третьих, это доверие к памяти: предположение, что текущие состояния памяти отражают прошлый опыт, что содержание памяти соответствует реальной истории.

Каждое из этих оснований доверия уязвимо для скептического оспаривания. Возможно, логика неприменима к реальности. Возможно, наши чувства систематически нас обманывают. Возможно, память

— это конфабуляция. Эти скептические сценарии нельзя исключить с помощью аргументации, поскольку аргументация предполагает именно те способности, которые подвергаются сомнению. Чтобы доказать, что логика применима к реальности, необходимо использовать логику. Чтобы доказать, что восприятие надежно, необходимо полагаться на восприятие. Защита основополагающего доверия является замкнутой, то есть она вообще не является защитой. Доверие просто дано, это предварительное условие для любого рассуждения.

Некоторые философы были глубоко обеспокоены этим наблюдением. Если доказательство основывается на недоказанных предположениях, не подрывает ли это всю идею рационального исследования? Но эта обеспокоенность основана на неправильном понимании того, что такое рациональное исследование. Рациональность не требует выхода за рамки всех предположений в некую независимую от предположений уверенность. Она требует

вдумчивого действия в рамках предположений, которые делают мышление возможным.

Философ Людвиг Витгенштейн выразил эту мысль с помощью понятия «опорных положений» — положений, которые должны оставаться неизменными для того, чтобы исследование могло продолжаться, подобно двери, вращающейся на петлях. Нельзя одновременно сомневаться во всем и рассуждать о чем угодно. Некоторые положения должны оставаться неизменными, в то время как другие должны исследоваться. Но то, какие положения являются опорными, а какие исследуются, может меняться со временем. То, что было основополагающим в одном исследовании, может стать предметом следующего.

теории ресурсов . С философской точки зрения мы видим, что доказательство требует недоказанных предположений. С точки зрения теории сложности мы видим, что верификация требует непроверенных обязательств. Выводы взаимно усиливают друг друга: невозможно достичь состояния чистого знания,

свободного от всякой веры. Вера является первостепенной — логически, структурно и эволюционно.

Это имеет последствия для нашего понимания рациональности. Традиционный идеал — рассуждающий, принимающий только то, что можно доказать, и воздерживающийся от суждений обо всем остальном — не просто непрактичен, но и непоследователен. Такой рассуждающий не смог бы ничего доказать, поскольку доказательство требует исходных предположений. Традиционный идеал ошибочно принимает локальный метод (доказывать утверждения относительно принятых посылок) за глобальную архитектуру (принимать только доказанные утверждения). Локальный метод ценен; глобальная архитектура невозможна.

Лучшим идеалом является вдумчивый верующий: тот, кто понимает неизбежность доверия, кто соизмеряет доверие с доказательствами и последствиями, кто остается открытым для пересмотра своих убеждений, когда они оказываются

ненадежными, кто различает то, что можно проверить, и то, чему просто нужно доверять. Это не более низкий стандарт, чем традиционный идеал; это более точное описание того, что на самом деле подразумевает хорошее рассуждение .

Мы все уже верующие. Вопрос не в том, верить ли, а в том, как верить правильно. А для того, чтобы верить правильно, необходимо понимать структуру веры , включая её глубочайшую особенность: вера предшествует доказательству и делает доказательство возможным.

Глава 7: Архитектура памяти

Если доверие — это основа сложного познания, то память — это его архив. Когда мы доверяем тому, что события вчерашнего дня произошли так, как мы их помним, когда мы полагаемся на усвоенные факты и отработанные навыки, когда мы узнаём знакомые лица и места — мы обращаемся к памяти. Но что же это за система — память? Ответ оказывается неожиданным и важным.

Наивное представление о памяти носит архивный характер: переживания записываются, хранятся и впоследствии извлекаются, подобно файлам в картотеке или видеозаписям на жестком диске. Согласно этой точке зрения, искажение памяти — это своего рода порча: исходная запись ухудшается со временем, из-за помех или дефектов. Более качественная память должна быть более точной, подобно тому как более качественный архив сохраняет свое содержимое более точно.

Однако десятилетия исследований опровергли это наивное представление. Память не архивируется, а реконструируется. Когда вы вспоминаете событие, вы не воспроизводите запись. Вместо этого вы реконструируете событие из фрагментов, закономерностей и схем, заполняя пробелы правдоподобными деталями, подстраивая реконструкцию под текущий контекст и цель. «Воспоминания», которые вы испытываете, генерируются в момент извлечения информации, а не хранятся заранее.

Этот реконструктивный характер объясняет характерные ошибки памяти. Мы неправильно помним детали, путаем похожие события, сжимаем время, вырываем события из их первоначального контекста, включаем информацию, полученную после события, и иногда «вспоминаем» события, которых никогда не было. Это не сбои в работе системы записи; это естественные последствия системы реконструкции.

Ключевое открытие современной науки о памяти заключается в том, что память должна быть реконструктивной. Дело не в том, что эволюция не смогла создать точную архивную память и вместо этого создала реконструктивную. Дело в том, что точная архивная память невозможна, учитывая ограничения, с которыми сталкиваются системы памяти. Реконструкция — это не недостаток; это единственная жизнеспособная архитектура.

Почему память должна быть реконструктивной? *В* работе Кригера (2025). «Вневременность пространства ментальной памяти: структурная гипотеза, основанная на ограничениях ресурсов, циклическом замыкании и реконструктивном извлечении информации» (Zenodo , https://doi.org/10.5281/zenodo.18381912) выделены три условия, которые в совокупности обусловливают необходимость реконструктивной архитектуры: ограничения ресурсов, циклическое замыкание и разрыв в проверке, который мы уже обсуждали.

Рассмотрим сначала ограничения ресурсов. Архивная система, хранящая полные записи опыта, столкнется с огромными затратами на хранение. Каждый момент сознательного опыта включает в себя обработку огромного количества сенсорной информации, эмоциональную окраску, контекстную структуру и ассоциативные связи. Для точного хранения всего этого потребуется емкость хранилища, линейно возрастающая с опытом, что в конечном итоге перегрузит любую систему с ограниченными ресурсами.

Реконструктивная система позволяет избежать этих затрат, храня не полные записи, а шаблоны, схемы и силу ассоциаций. Из этих сжатых представлений можно по запросу восстанавливать конкретные воспоминания. Реконструкция никогда не бывает идеальной, но она достаточно хороша для большинства целей, при этом потребляя гораздо меньше памяти. Это классический инженерный компромисс: смириться с некоторой потерей

точности в обмен на существенное снижение требований к ресурсам.

Далее рассмотрим циклическое замыкание — наблюдение, что воспоминания не существуют изолированно, а образуют взаимосвязанные сети ассоциаций. Воспоминание A вызывает воспоминание B, которое вызывает B, которое может привести обратно к A. Эта циклическая структура создает проблему для архивных систем, которые полагаются на фиксированные адреса или временные метки. Как индексировать воспоминание, к которому можно получить доступ по нескольким ассоциативным путям? Как поддерживать временной порядок в сети межвременных ассоциаций?

Реконструктивная система естественным образом обрабатывает циклическое замыкание. Вместо фиксированных мест хранения, воспоминания существуют в виде паттернов активации в ассоциативных сетях. Одни и те же базовые паттерны могут активироваться по разным путям, что приводит к реконструкциям, которые

различаются в зависимости от пути доступа. Это объясняет, почему одно и то же событие может запоминаться по-разному в зависимости от того, что послужило поводом для его воспроизведения.

Наконец, рассмотрим проблему проверки. Когда вы восстанавливаете воспоминание, вы не можете проверить, соответствует ли восстановленное содержание исходному опыту. Нет независимой записи, с которой можно было бы сравнить данные. Вам остается лишь поверить, что реконструкция отражает что-то реальное из вашего прошлого — акт доверия, ничем не отличающийся от доверия к предсказанию о будущем.

Возможно, это наиболее глубокое следствие реконструктивной памяти. Извлечение воспоминаний — это не операция проверки, а операция генерации. Когнитивная архитектура, порождающая воспоминания, — это та же самая архитектура, которая порождает предсказания и фантазии — генеративная модель, создающая правдоподобное содержание на основе усвоенных паттернов.

Воспоминания, предсказания и фантазии — все это результаты одного и того же генеративного процесса, различающиеся прежде всего рамками, в которых они помечены (прошлое, будущее, гипотетическое).

Нейробиология подтверждает это единство. Области мозга, участвующие в запоминании, в значительной степени совпадают с областями, участвующими в представлении будущих событий и моделировании гипотетических сценариев. Память, воображение и проекция имеют общие нейронные субстраты, поскольку они имеют общую когнитивную архитектуру: все это реконструктивные процессы, которые генерируют содержание из сохраненных паттернов.

Таким образом, реконструктивная архитектура памяти — это не курьез человеческой психологии, а структурное следствие ограничений, с которыми сталкивается любая сложная система памяти. Те же самые ограничения — ограниченность ресурсов, ассоциативная структура, невозможность проверки — подтолкнули бы любую достаточно сложную

систему к реконструктивной памяти, независимо от того, эволюционировала ли эта система естественным путем или была создана искусственно.

Глава 8: Время как содержание, а не как структура.

Мы, естественно, говорим о воспоминаниях как о событиях, занимающих определённые позиции во времени: это произошло раньше, вчерашний день наступил после прошлой недели, детство — это более далёкая прошлое, чем взрослая жизнь. Такая временная структура предполагает, что память организована хронологически — что время является фундаментальным измерением хранения памяти, подобно временной шкале в видеоредакторе или отметкам дат в фотоархиве.

Однако такая естественная трактовка вводит в заблуждение. Имеющиеся данные свидетельствуют о совершенно ином: время — это не структурное измерение хранения памяти , а характеристика содержания, восстанавливаемая при извлечении. Состояния памяти не занимают временные позиции по своей природе; временная упорядоченность возникает в процессе запоминания.

Это тезис об атемпоральном хранении памяти, разработанный в работе *«Атемпоральность пространства ментальной памяти»* (Кригер, 2025). Основная идея может быть сформулирована просто: в архитектуре хранения памяти отсутствует временная ось. Состояния памяти обладают идентичностью независимо от того, когда они были закодированы или когда были извлечены. Временная информация, если она существует, кодируется как содержание — как характеристики памяти, такие как «лето», «детство» или «до переезда », — а не как координаты во временной структуре.

Различие между временем как структурой и временем как содержанием тонкое, но крайне важное. Рассмотрим разницу между базой данных с временными координатами (каждая запись имеет временную позицию как часть своего адреса) и базой данных, где временная информация является просто еще одним полем (каждая запись может содержать временное содержание, но ее адрес не зависит от времени). В первой архитектуре ответ на вопрос

«когда это произошло» дается путем поиска позиции записи. Во второй — путем чтения содержимого из самой записи.

Похоже, что память имеет вторую архитектуру. Когда вы пытаетесь определить, когда произошло событие, вы не обращаетесь к его временным координатам напрямую. Вместо этого вы восстанавливаете временную информацию из контекстных подсказок внутри памяти: какое это было время года, на каком этапе жизни вы находились, какие другие события происходили примерно в то же время. Временная последовательность выводится косвенно, а не считывается.

Это объясняет многие загадочные особенности временной памяти. Почему летнее воспоминание из детства может казаться ближе, чем прошлый месяц? Потому что «близость» восстанавливается на основе яркости, эмоциональной значимости и доступных деталей — характеристик, которые не соответствуют хронологической дистанции. Почему мы сжимаем

отдаленные события (воспринимая их как более близкие к настоящему, чем они были на самом деле) и расширяем недавние? Потому что оценка времени основана на эвристических методах, которые могут ввести в заблуждение. Почему мы иногда путаем порядок событий? Потому что порядок должен быть выведен из контекстных подсказок, а эти подсказки могут быть неоднозначными.

Научная литература подтверждает эту вневременную интерпретацию. Исследования показывают, что точность запоминания времени коррелирует с богатством контекстной информации, а не с течением времени. События с выраженным контекстом — праздники, переходные периоды , драматические происшествия — записываются во времени точнее, чем обыденные события, независимо от того, как давно они произошли. Это имеет смысл, если временное позиционирование зависит от восстанавливаемого контекста, а не от внутренних временных координат.

Модель временного контекста, ведущая теория извлечения информации из памяти, кодирует временные отношения посредством постепенно изменяющегося контекста, а не посредством явных временных меток. Элементы, закодированные близко друг к другу во времени, имеют схожий контекст; элементы, разделенные во времени, имеют расходящиеся контексты. На вопрос «Когда это произошло?» дается ответ путем оценки сходства контекста с другими воспоминаниями, время которых известно. Это временная информация как содержание, а не временное положение как структура.

А что насчет временных клеток — нейронов в гиппокампе, которые активируются через определенные интервалы во время задержек, по-видимому, кодируя временную информацию? Эти специализированные механизмы на самом деле подтверждают , а не противоречат гипотезе об отсутствии временной структуры. Если бы хранение памяти было по своей природе временным,

специализированные механизмы синхронизации были бы не нужны. Их существование предполагает, что архитектура по умолчанию лишена временной структуры, требуя специализированных дополнений для задач, требующих точного временного кодирования.

Аналогия с цветовым зрением весьма поучительна. Восприятие цвета требует специализированных фоторецепторных систем; это не означает, что зрительная кора по своей природе окрашена. Аналогично, временное кодирование требует специализированных механизмов; это не означает, что хранение памяти по своей природе является временным. Цвет и время — это характеристики, которые кодируются определенными системами, а не измерения, которые пронизывают все нейронные представления.

Почему память организована именно таким образом? В *работе «Эволюционный отбор в пользу вневременного хранения данных»* (Кригер, 2019) утверждается, что вневременное хранение имеет

существенные преимущества. Оно дешевле в плане ресурсов (отсутствуют накладные расходы на временное индексирование), быстрее в извлечении (не требуется временной поиск) и более гибко для рекомбинации (отсутствие временной привязки ограничивает генерацию новых комбинаций). Эволюция будет отдавать предпочтение этим преимуществам, отбирая вневременные архитектуры везде, где они жизнеспособны.

Философские выводы поразительны. Если время в памяти — это содержание, а не структура, то временная организация нашего автобиографического прошлого конструируется, а не даётся нам изначально. Повествование о нашей жизни — это случилось, потом то, потом другое — это результат реконструкции, а не прочтения существовавшей ранее структуры. Мы помним не во времени, а во времени.

Глава 9: Почему искажается память

Память искажается. Это одно из наиболее убедительных открытий в психологической науке. Мы неправильно помним разговоры, путаем похожие события, вставляем детали, которых никогда не было, забываем то, что действительно произошло, и с уверенностью сообщаем о воспоминаниях, которые, как показывает проверка, оказываются ложными. Эти искажения не являются редким исключением; они являются распространенной особенностью нормальной человеческой памяти.

Традиционная интерпретация рассматривает искажение как сбой. Предполагается, что память точно сохраняет прошлое; искажения представляют собой сбои в этом сохранении. С этой точки зрения, мы должны минимизировать искажения, где это возможно, и сожалеть о их возникновении, если это неизбежно. Лучшая память — это более точная память.

Представленная в этой книге концепция предлагает иную интерпретацию. Искажение памяти — это не ошибка, а особенность, не сбой в сохранении, а следствие архитектуры, оптимизированной для гибкости, эффективности и релевантности. Те же свойства, которые приводят к искажению, также обеспечивают преимущества, делающие реконструктивную память ценной.

Рассмотрим реконсолидацию — одно из самых поразительных открытий современной науки о памяти. Когда воспоминание извлекается, оно не просто воспроизводится и возвращается в хранилище неизмененным. Вместо этого, извлечение делает память лабильной — восприимчивой к модификации. Новая информация, присутствовавшая при извлечении, может быть включена. Эмоциональные состояния могут быть обновлены. «Восстановленное» воспоминание может отличаться от оригинала. Каждый акт воспоминания потенциально является актом пересмотра.

С точки зрения сохранения информации, реконсолидация выглядит как недостаток конструкции. Зачем системе сохранения прошлого допускать, чтобы извлечение информации изменяло сохраненные данные? Но с точки зрения гибкости, реконсолидация выглядит как преимущество. Она позволяет обновлять воспоминания новой информацией, сохранять их актуальность в текущих обстоятельствах и интегрировать с последующим обучением. Система памяти, которая никогда не обновляется, сохраняла бы прошлое за счет снижения его полезности в настоящем.

Рассмотрим случаи ошибок мониторинга источников — ситуации, когда мы помним контент, но не помним его происхождение. Мы помним, что слышали что-то, но не можем вспомнить, кто это сказал, или мы помним идею, не помня, читали ли мы её, слышали ли или сами её придумали. Такие ошибки распространены и могут привести к плагиату, ложному указанию авторства и неоправданной уверенности.

Однако хранение исходной информации обходится дорого и часто не имеет отношения к её использованию. Для большинства целей важно не то, откуда эта информация получена , а то, является ли она достоверной или полезной . Система, которая отбрасывала бы исходную информацию для экономии ресурсов, повысила бы эффективность хранения за счёт снижения точности источника. Эволюция, по-видимому, пошла на этот компромисс, кодируя исходную информацию неполно и восстанавливая её на основе контекстных подсказок при извлечении.

Рассмотрим временные искажения — телескопирование (восприятие отдаленных событий как более близких, чем они были на самом деле), краевые эффекты (группировка воспоминаний вокруг значимых переходов) и смещение (воспоминание о событиях как произошедших в другое время, чем они были на самом деле). С точки зрения сохранения информации, это нарушения точности . Но с точки зрения релевантности, они могут быть приемлемыми

последствиями системы, которая отдает приоритет осмысленной организации над хронологической точностью.

концепция *вневременности пространства ментальной памяти* (Кригер, 2025). Если время восстанавливается из содержания, а не считывается из структуры, то точность временной реконструкции зависит от качества временных сигналов, закодированных в каждом воспоминании. Когда сигналы богаты и различимы, временная реконструкция точна. Когда сигналы скудны или неоднозначны, временная реконструкция дает ошибки. Искажения следуют предсказуемым закономерностям, основанным на архитектуре, а не на случайном шуме, возникающем из-за сбоя системы.

Рассмотрим ложные воспоминания — яркие, уверенные воспоминания о событиях, которых никогда не было. В известных экспериментах исследователи внедряли ложные воспоминания о детских переживаниях (например, о том, как

заблудились в торговом центре, о том, как пролили пунш на свадьбе), которые участники впоследствии вспоминали с подробностью и убежденностью. Как это возможно, если память должна сохранять реальный опыт?

Реконструктивная теория объясняет ложную память как естественный результат работы генеративной системы. Память не воспроизводит сохраненные записи; она генерирует правдоподобные реконструкции на основе паттернов. Если паттерны подтверждают ложное событие так же легко, как и истинное — если элементы согласованно сочетаются — ложное событие может быть «вспомнено» с той же феноменологией, что и истинное. Система не может отличить создание реконструкции того, что произошло, от создания конструкции того, чего не было.

Это не обнадеживающий вывод для тех, кто хочет, чтобы память была надежным проводником в прошлое. Но он согласуется с реальной функцией памяти. Память существует не для того, чтобы

сохранять прошлое ради самого себя, а для того, чтобы направлять будущие действия. Для действий важно не то, насколько точны воспоминания, а то, насколько они полезны — поддерживают ли закодированные в них закономерности адаптивное поведение. Система, оптимизированная для полезности, а не для точности, выглядела бы точно так же, как человеческая память: реконструктивная, гибкая, подверженная предсказуемым искажениям, но удивительно эффективная в извлечении полезной информации из прошлого опыта .

Таким образом, искажение памяти — это не сбой системы сохранения, а нормальная работа системы генерации. Мы не должны ожидать от памяти точности в архивном смысле; мы должны ожидать от нее полезности в прогностическом смысле. И по этому стандарту память служит нам на удивление хорошо — со всеми ее искажениями.

Глава 10: Эволюционная логика

Если искажение памяти — это скорее особенность, чем ошибка, то почему эволюция создала именно эту особенность? В предыдущих главах утверждалось, что реконструктивная, вневременная архитектура памяти вытекает из ограничений ресурсов и разрыва в проверке. В этой главе более непосредственно рассматривается эволюционная логика: какие факторы отбора благоприятствуют этой архитектуре, и почему естественный отбор не должен вместо этого создавать более точную память?

Эволюционный *отбор в пользу вневременного хранения памяти»* (Кригер, Б. (2019). Эволюционный отбор в пользу вневременного хранения памяти: почему три сходящихся фактора благоприятствуют архитектурам, где время принадлежит извлечению, а не хранению. Zenodo . https://doi.org/10.5281/zenodo.18381880) выделены три независимых фактора, которые сходятся к вневременной, реконструктивной памяти: стоимость

ресурсов, скорость извлечения и адаптивная гибкость. Каждый из этих факторов сам по себе благоприятствовал бы такой архитектуре; вместе они делают альтернативу — точную архивную память — крайне невыгодной.

затраты ресурсов . Мозг — дорогостоящий орган, потребляющий примерно двадцать процентов метаболических ресурсов, при этом составляющий всего около двух процентов массы тела. Любая архитектурная особенность, снижающая нейронные затраты при сохранении функциональности, дает селективное преимущество. Организмы с более эффективным мозгом могут направлять больше ресурсов на другие важные для выживания виды деятельности: рост, размножение, иммунную функцию, физическую работоспособность.

Временное индексирование влечет за собой издержки, которых можно избежать при использовании временного контента. Архивная система с временными метками требует инфраструктуры для генерации временных маркеров

при кодировании, поддержания временных связей по мере формирования новых воспоминаний и поддержки временных запросов при извлечении. Эта инфраструктура имеет метаболические издержки. Реконструктивная система, которая хранит временную информацию в виде характеристик контента — если она вообще ее хранит — избегает этих накладных расходов. Экономия может показаться незначительной для любого отдельного воспоминания, но в совокупности на миллионах воспоминаний и миллионах поколений небольшая экономия превращается в решающее преимущество.

Рассмотрим далее скорость поиска. В средах, где задержка реакции обходится дорого — где колебание означает потерю добычи или превращение в добычу — более быстрый поиск имеет селективное преимущество. Большинство адаптивных запросов основаны на содержании : Это опасно? Это съедобно? Я сталкивался с этим раньше? Эти вопросы требуют сопоставления текущего ввода с сохраненными

шаблонами, процесс, основанный на сходстве и ассоциации, а не на временном положении.

Временные запросы — Когда я в последний раз видел этого хищника? Как давно я ел? — обычно являются второстепенными и возникают после основного поиска информации на основе содержимого. Если временная информация закодирована как характеристики содержимого, она передается вместе с извлеченной информацией без дополнительных затрат. Если для получения временной информации требуется отдельный поиск во временном индексе, поиск замедляется. В условиях конкуренции эта задержка может быть фатальной.

Рассмотрим адаптивную гибкость в третьем аспекте . Изменчивая среда поощряет перекомбинацию прошлого опыта для генерации новых реакций. Организм, способный воспроизводить прошлый опыт только дословно, ограничен ситуациями, с которыми он уже сталкивался. Организм, способный комбинировать элементы прошлого опыта — брать реакцию,

усвоенную в одном контексте, и адаптировать ее к другому, — гораздо лучше справляется с новизной.

Временная привязка ограничивает рекомбинацию. Если воспоминания хранятся с внутренними временными координатами, объединение элементов из разных времен создает временную несогласованность — своего рода внутреннее противоречие. Вневременное хранение допускает свободную рекомбинацию. Временная информация, хранящаяся в виде контента, может быть сохранена, изменена или отброшена по мере необходимости рекомбинации. Система может генерировать сценарии, которые никогда не происходили, прогнозы будущего, которые никогда ранее не встречались, решения проблем, с которыми никогда ранее не сталкивались.

Эта гибкость имеет свою цену: та же способность к рекомбинации , которая обеспечивает творчество, позволяет создавать и конфабуляции. Система, которая может свободно комбинировать элементы памяти, иногда будет комбинировать их

неправильно, создавая ложные воспоминания. Но в изменчивых условиях преимущества гибкости перевешивают издержки случайных ошибок . Эволюция калибрует этот компромисс не путем устранения ошибок, а путем управления их частотой и последствиями.

Сближение этих трех факторов — стоимости ресурсов, скорости извлечения информации и адаптивной гибкости — делает вневременную, реконструктивную память ожидаемым эволюционным результатом для сложных систем. Это не означает, что эволюция всегда достигает оптимальных решений; зависимость от предшествующего пути развития, ограничения развития и генетический дрейф могут препятствовать достижению оптимальных результатов. Но направление отбора ясно: все три фактора стремятся к одной и той же архитектуре.

Формальный анализ показывает, что преимущество атемпоральной архитектуры возрастает с увеличением сложности системы.

Экономия ресурсов масштабируется пропорционально количеству хранимых данных в памяти. Преимущество в скорости растет логарифмически с увеличением объема памяти. Преимущество в гибкости зависит от изменчивости окружающей среды. По мере усложнения организмов — с большими объемами памяти, более разнообразной средой и более длительными горизонтами планирования — селективное преимущество атемпоральной архитектуры возрастает.

Это объясняет закономерность в природе: более сложные организмы демонстрируют более реконструктивную, но менее достоверную память. Простые организмы с ограниченным поведенческим репертуаром могут ориентироваться на архивное хранение информации. Сложные организмы с богатой поведенческой гибкостью все больше полагаются на реконструкцию. Эта закономерность не случайна; это признак селективного давления, которое усиливается с увеличением сложности.

Там, где точное кодирование времени имеет решающее значение для приспособленности, эволюция добавляет специализированные механизмы. Клетки, отвечающие за восприятие времени в гиппокампе, кодируют интервалы в периоды задержки. Системы отслеживания интервалов времени контролируют длительность выполнения конкретных задач. У птиц, запасающих пищу, есть специализированная память на информацию о том, что, где и когда находятся их запасы . Это дополнения к базовой архитектуре, а не особенности общей памяти. Их существование как специализированных систем подтверждает, что стандартная архитектура не обладает внутренней временной структурой.

Таким образом, эволюционная логика переосмысливает то, что мы считаем недостатками памяти. Временные искажения, реконсолидация, ложные воспоминания, путаница источников — это не сбои системы, которая должна была быть более точной. Это признаки системы , которую эволюция...

Разработаны для эффективности, скорости и гибкости. Эволюция не создает точные архивы; она создает эффективные генераторы. А для сложных систем в изменчивой среде генерация превосходит архивирование.

Глава 11: Жизнь вне времени

Мы существуем во времени. Наши тела стареют, наш мир меняется, события разворачиваются в необратимой последовательности. Физическое время — это рамки, в которых протекает наша жизнь, среда, через которую течет причинно-следственная связь, измерение, из которого мы не можем убежать.

Однако в предыдущих главах утверждалось, что память — внутреннее пространство, в котором мы храним наше прошлое, — лишена внутренней временной структуры. Время в памяти — это содержание, а не координата; оно реконструируется при извлечении, а не хранится при кодировании. Если это верно, то мы сталкиваемся с поразительным парадоксом: мы — физические существа, встроенные во время, но наше внутреннее представление об опыте в некотором смысле существует вне времени.

Что значит жить вне времени? Конечно, не то, что мы ускользаем от физического времени — мы

стареем, мы не можем вернуться в прошлое, мы неумолимо движемся к будущему. Но в ментальном пространстве памяти временная организация не дана; она достигается. «Прошлость» воспоминания — это не внутреннее свойство его состояния хранения, а интерпретация, навязываемая при извлечении. Последовательность воспоминаний — это До этого — считывается не из временных координат, а выводится из контекстных связей.

Рассмотрим феноменологию воспоминаний. Иногда далекие события кажутся близкими — летнее воспоминание из детства такое же яркое и актуальное, как вчера. Иногда недавние события кажутся далекими — прошлая неделя уже растворяется в смутном прошлом. Субъективное время памяти не совпадает с хронологическим. Эмоциональная интенсивность, личная значимость, контекстуальная уникальность — именно они определяют, насколько «близким» или «далеким» кажется воспоминание, а не его положение в календаре.

Это не поэзия и не метафора; это прямое следствие архитектуры . Если состояния памяти не имеют внутренних временных координат, то «временное расстояние» должно вычисляться на основе других характеристик — яркости, детализации, контекстуального сходства с настоящим. Эти характеристики не всегда надежно соответствуют хронологическому расстоянию. Яркое воспоминание из детства и яркое воспоминание из недавнего времени могут быть одинаково «близки» в пространстве характеристик , определяющем субъективное расстояние, даже если между ними хронологически прошло несколько десятилетий.

В *работе* Кригера (2025) формально развивается эта идея. Внутреннее пространство памяти описывается как сеть ассоциаций, в которой временные отношения являются одним из многих типов содержательных характеристик. Навигация в этом пространстве осуществляется по ассоциативным путям, а не по временным траекториям. Вы переходите от воспоминания к воспоминанию по

сходству, по причинно-следственной связи, по эмоциональному резонансу, а не по хронологической последовательности.

Это объясняет, почему память так легко реорганизуется вокруг тем, а не времени . Спросите кого-нибудь о его опыте переживания утраты, и он вспомнит события всей своей жизни, связанные эмоциональным сходством, а не временной близостью. Спросите об опыте, пережитом в конкретном месте, и воспоминания сгруппируются по месту, а не по времени. Временная организация, которую мы навязываем, рассказывая историю своей жизни, — это достижение нарративного построения, а не прочтение уже существующей структуры.

Последствия для личной идентичности огромны. Обычно мы представляем себя как непрерывное существо во времени — того же человека, который существовал вчера, в прошлом году, в детстве. Эта непрерывность, кажется, обеспечивается памятью: я помню, что был тем ребенком, следовательно, я тот же самый человек. Но

если память лишена внутренней временной структуры, то непрерывность личности не является чем-то само собой разумеющимся; она конструируется.

Каждый акт воспоминания — это акт соединения, связывания нынешнего «я» с реконструированным прошлым «я», утверждения этого прошлого как своего собственного. Это утверждение не является произвольным; оно следует закономерностям, сформированным подлинными причинно-следственными связями между прошлым и настоящим. Но это также не простое прочтение сохраненной преемственности. Нарратив о себе создается в процессе извлечения информации, а не хранится в процессе кодирования.

В этом осознании есть что-то головокружительное. Если временная организация моего прошлого конструируется, а не даётся мне, то моя автобиография — это своего рода непрерывный творческий проект, а не фиксированная запись. Прошлое, которое я помню, — это прошлое, которое

я конструирую в процессе воспоминания. Возможны разные конструкции; та, в которой я живу, не единственная, соответствующая моим накопленным моделям поведения.

Но есть и нечто освобождающее. Если прошлое конструируется в процессе его восстановления, то оно не полностью определяется произошедшими событиями. Терапевтические вмешательства, изменяющие способ запоминания прошлого — не путем изменения фактов, а путем изменения его конструирования — становятся понятными. Реконсолидация — это не искажение зафиксированной записи, а обновление живой модели. Прошлое, в важном смысле, остается открытым.

Это не означает, что прошлое произвольно или ничем не ограничено. Шаблоны, хранящиеся в памяти, ограничивают то, что можно построить. Вы не можете вспомнить всё, что вам угодно; вы можете строить только из имеющихся материалов. И эти материалы отражают, пусть и несовершенно,

реальный прошлый опыт. Построение ограничено реальностью, даже если оно ею не определяется.

Таким образом, мы живем в своеобразном состоянии: будучи физическими существами, встроенными в поток времени, мы несем в себе ментальное пространство, где время — не измерение, а конструкция. Мы проецируем время вовне, когда это необходимо — для планирования, для повествования, для координации действий с другими. Но внутри, в самой архитектуре памяти, мы в некотором смысле вне времени. Это не трансценденция времени в каком-либо мистическом смысле. Это структурное следствие архитектуры, оптимизированной для гибкости, а не для хронологии.

Глава 12: Спектр доверия

Аргументы этой книги применимы не только к людям, но и к любой достаточно сложной системе. В этой главе рассматривается спектр доверия в живом мире, от простейших организмов, которые функционируют преимущественно на основе проверки, до самых сложных, которые функционируют преимущественно на основе убеждений.

На одном конце спектра находится бактерия. Такая бактерия, как *Escherichia coli,* ориентируется в окружающей среде с помощью хемотаксиса — перемещаясь вверх по градиентам аттрактантов и вниз по градиентам репеллентов. Ее пространство решений по сути является бинарным: увеличивается или уменьшается концентрация этого химического вещества? Ее пропускная способность для проверки, обеспечиваемая химическими рецепторами, достаточна для охвата этого небольшого пространства. Для бактерии охват проверки высок;

большинство релевантных для действий утверждений можно проверить напрямую.

Бактерия по-прежнему нуждается в некотором доверии — в некоторой приверженности, выходящей за рамки простого подтверждения. Она должна верить, что химические градиенты предсказывают местоположение питательных веществ, что то, что работало мгновение назад, будет работать и сейчас. Но соотношение доверия и подтверждения невелико. Бактерия работает в рамках идеала подтверждения, который прославляет традиционная эпистемология.

Перейдём к насекомому, которое уже обладает большей сложностью. Пчела должна ориентироваться в пространстве, распознавать цветы по цвету и форме, запоминать местоположение продуктивных участков, сообщать об этом сородичам, отслеживать время суток, чтобы соответствовать доступности цветов. Пространство принятия решений значительно расширилось; возможности проверки выросли менее существенно. Пчела в большей степени полагается на доверие — на

усвоенные закономерности и ожидания, которые превосходят то, что можно проверить в момент принятия решения.

У млекопитающих зависимость от достоверности информации очевидна. Рассмотрим мышь, которая ориентируется на своей территории, отслеживает источники пищи и хищников, распознает отдельных особей, помнит прошлые события и предвидит будущие. Пространство принятия решений включает в себя не только физическую среду, но и социальную, не только настоящее, но и запомненное прошлое и ожидаемое будущее. Пропускная способность системы проверки не успевает за этим процессом; мыши приходится подтверждать множество утверждений, которые она не может проверить напрямую.

Приматы, и особенно люди, представляют собой крайнюю степень когнитивных способностей, основанных на доверии. Пространство принятия решений у человека включает в себя абстрактные понятия, гипотетические сценарии, социальные

отношения византийской сложности, временные горизонты, простирающиеся на годы или десятилетия. Мы рассуждаем о других людях, о справедливости, о собственных мотивах, о смысле жизни. Ничто из этого нельзя проверить напрямую. Почти все, во что мы верим о себе, других и мире, основано на доверии.

Этот спектр — от бактериальной верификации до человеческого доверия — не является лестницей прогресса от примитивного к продвинутому. Это компромисс между различными режимами работы, каждый из которых адаптирован к своей нише. Когнитивные способности бактерий, основанные на верификации, не уступают другим; они подходят для организма с простым пространством принятия решений и ограниченными ресурсами. Когнитивные способности человека , основанные на доверии, не превосходят другие; они необходимы для организма с чрезвычайно сложным пространством принятия решений.

Эволюционная *теория доверия* (Кригер, 2022) формализует этот спектр. Покрытие верификации — отношение проверяемых утверждений к утверждениям , имеющим отношение к действию, — уменьшается по мере увеличения сложности. Простые системы могут достигать высокого покрытия; сложные системы неизбежно имеют низкое покрытие. Переход не резкий, а постепенный, отслеживающий рост пространства решений относительно возможностей верификации.

Можно задаться вопросом: если когнитивные процессы, основанные на доверии, более рискованны (более подвержены ошибкам), чем когнитивные процессы, основанные на проверке, то почему эволюция их породила? Ответ заключается в том, что сама сложность была выгодна. Организмы, способные моделировать окружающую среду более детально, планировать на более дальние расстояния, сотрудничать в больших группах, использовать более сложные инструменты, — эти организмы превосходили более простых конкурентов. Но для

достижения этих ниш требовалось принятие когнитивных процессов, основанных на доверии. Невозможно существование человеческого интеллекта без человеческой доверчивости.

Спектр достоверности проливает свет на важную природу интеллекта. Мы часто предполагаем, что более высокий уровень интеллекта означает большую точность, лучшую проверку, более надежные знания. Но данные свидетельствуют об обратном. Более высокий уровень интеллекта означает большую способность к моделированию, а моделирование означает приверженность представлениям, превосходящим проверку. Самые интеллектуальные системы не являются самыми надежными; они наиболее способны к продуктивной неопределенности.

Это меняет наше понимание когнитивных достижений. Бактерия, в некотором смысле, знает свой мир более надежно, чем мы — её убеждения лучше проверены. Но она не может делать то, что можем делать мы. Наши достижения — наука,

искусство, технологии , культура — возникают именно из нашей готовности к самосовершенствованию, выходящему за рамки проверки. Мы выдвигаем гипотезы, мы представляем, мы доверяем, мы верим. Это не неудачи в достижении бактериальной уверенности; это те способности, которые делают нас людьми.

Таким образом, спектр доверия — это не иерархия от худшего к лучшему уровню познания. Это карта компромиссов, на которые шли разные организмы, выбирая между проверкой и верой, между уверенностью и способностью, между знанием и верой. Каждая позиция на этом спектре представляет собой жизнеспособное решение проблемы действий в сложном мире с ограниченными ресурсами. Мы находимся на крайнем полюсе познания, основанного на доверии, — и этот крайний полюс является одновременно нашим ограничением и нашей славой.

Глава 13: Доверие как основа разума

Мы убедились, что сложные системы должны опираться на доверие — на уверенность, выходящую за рамки проверки. Эта глава углубляет это понимание, показывая, что сам разум, та самая способность, которую мы связываем с преодолением простой веры, в своей основе зависит от доверия. Невозможно рассуждать без доверия ; доверие — это не провал разума, а его предпосылка.

Рассмотрим, что требуется для рассуждения. Рассуждать — значит переходить от посылок к выводам в соответствии с правилами, сохраняющими истину. Но это движение предполагает несколько уровней доверия, которые сами по себе не могут быть установлены посредством рассуждения без цикличности.

Во-первых, нужно доверять логике. Мы должны верить, что обоснованный вывод сохраняет истину, что если посылки истинны и рассуждение обосновано, то и заключение должно быть истинным.

Но можем ли мы это доказать? Любое доказательство будет использовать логический вывод — именно то, надежность которого мы пытаемся установить. Обоснование является замкнутым кругом. Мы доверяем логике, потому что не можем последовательно сомневаться в ней, а не потому, что проверили её с помощью какого-либо независимого стандарта.

Во-вторых, доверие к восприятию. Рассуждения о мире требуют предпосылок о мире, и эти предпосылки обычно выводятся из восприятия. Мы видим, что небо голубое, чувствуем, что вода холодная, слышим звон колокола. Но восприятие может нас обмануть — иллюзии, галлюцинации и ошибки восприятия хорошо задокументированы. Мы доверяем тому, что восприятие в целом соответствует реальности, но мы не можем проверить это доверие без использования восприятия, что опять же является замкнутым кругом.

В-третьих, доверие к памяти. Рассуждения разворачиваются во времени; мы должны помнить

предпосылки, делая выводы. Мы должны доверять тому, что наша память о предпосылках остается точной на протяжении всего процесса рассуждения. Но память, как мы видели, является реконструктивной и подвержена ошибкам. Мы не можем в каждый момент времени проверить, не исказила ли память наши предпосылки; мы просто доверяем этому.

В-четвертых, доверие к языку. Большинство рассуждений используют язык — символы, представляющие понятия. Мы должны верить, что слова означают то, что мы в них понимаем, что наши понятия формируют реальность , что общение действительно передает информацию . Это доверие невозможно установить посредством рассуждений на языке, не предполагая заранее, что именно оно призвано установить.

В-пятых, доверие к чужим мыслям. Большая часть наших рассуждений основана на свидетельствах — мы верим чему-то, потому что нам об этом рассказали другие. Мы верим, что другие

существуют как разумные существа (а не как философские зомби), что они способны воспринимать реальность, что они достаточно честно общаются, чтобы быть полезными источниками информации. Ничто из этого нельзя полностью проверить; мы доверяем этому, потому что должны.

В *эволюционной теории доверия* (Кригер, 2022) эти понятия называются «процедурными основаниями доверия » — фундаментальными обязательствами, необходимыми для осуществления любого вывода. Они отличаются от обычных убеждений тем, что их нельзя приостановить или проверить, не подорвав саму способность что-либо приостанавливать или проверять. В метафоре Витгенштейна они являются теми «шарнирами», вокруг которых строится исследование.

Это наблюдение не приводит к скептицизму — выводу о том, что мы должны сомневаться во всем, потому что не можем проверить основы. Этот путь саморазрушителен; сомнение в основах требует использования тех самых способностей, надежность

которых подвергается сомнению. Наблюдение, напротив, ведет к осознанию: мы — доверчивые существа, и это доверие способствует, а не подрывает наши познавательные достижения.

Традиционная картина ставит разум выше доверия: сначала мы проверяем, затем познаём, и вера — это то, на что мы соглашаемся, когда знание оказывается несостоятельным. Картина, вытекающая из этого анализа, переворачивает эту иерархию: сначала мы доверяем, и в рамках этого доверия мы можем рассуждать, а рассуждение иногда приводит к локальной проверке, которую мы называем знанием. Доверие не стоит ниже разума; оно является основой, на которой зиждется разум.

Это имеет значение для нашего понимания рациональных разногласий. Когда два человека тщательно рассуждают и приходят к разным выводам, мы часто предполагаем, что один из них рассуждает неправильно. Но если разум опирается на фундаментальные доверия, а эти доверия могут меняться, то устойчивые разногласия между

разумными людьми становятся понятными. Они могут рассуждать правильно, исходя из разных оснований — разных процессуальных убеждений относительно того, что считается доказательством, какие стандарты доказательства применять , каким источникам доверять.

Это не означает, что все разногласия законны или что истина лишь относительна и зависит от основополагающих принципов доверия. Некоторые принципы доверия могут быть лучше выверены, чем другие; некоторые могут способствовать формированию верных убеждений. Но это означает, что рациональное убеждение не всегда может быть успешным только с помощью рассуждений. Иногда требуется не более убедительная аргументация, а иное доверие — переориентация основополагающих принципов, которую невозможно навязать с помощью аргументов.

В основе всего этого лежат доверчивые существа, развившие способность к рассуждению. Разум — великолепный инструмент, но это

инструмент, используемый верующими, а не замена вере. Понимание познания — это понимание доверия; понимание знания — это рассмотрение его как достижения в рамках системы убеждений, которая делает это достижение возможным.

Глава 14: Невозможность полной рациональности

Идеал полной рациональности преследовал западную мысль на протяжении тысячелетий: мудрый человек, принимающий только то, что оправдано, соизмеряющий свои убеждения с доказательствами, никогда не берущий на себя обязательства, выходящие за рамки того, что можно проверить. В этой главе утверждается, что такая рациональность не только труднодостижима, но и структурно невозможна для любой сложной системы.

Аргумент исходит из двух независимых направлений, которые сходятся к одному и тому же выводу. С точки зрения сложности мы видим, что возможности проверки не успевают за размером пространства решений. Любая система, достаточно сложная для моделирования окружающего мира, должна придерживаться утверждений, которые она не может проверить. Полная рациональность — принятие только проверенных утверждений —

потребовала бы отказа от сложности, которая делает возможным сложное познание.

Исходя из фундаментальных принципов, мы видим, что сам разум опирается на доверие, которое невозможно рационально обосновать без логической ошибки. Рациональный человек доверяет логике, восприятию, памяти, языку и другим разумам — не потому, что это доверие рационально обосновано, а потому, что сама рациональность его предполагает. Полная рациональность потребовала бы рационального обоснования условий рационального обоснования, что невозможно.

Эти два аргумента независимы, но взаимно усиливают друг друга. Даже если бы мы каким-то образом смогли обосновать основы разума, мы все равно столкнулись бы с компромиссом между сложностью и достоверностью. Даже если бы возможности проверки каким-то образом соответствовали пространству принятия решений, мы все равно столкнулись бы с проблемой фундаментального доверия. Оба пути ведут к одному

и тому же результату: полная рациональность невозможна для любой сложной, рассуждающей системы.

Этот вывод порой вызывает тревогу. Если полная рациональность невозможна, значит ли это, что мы обречены на иррациональность? Должны ли мы принять предвзятость, суеверия и необоснованные убеждения? Вовсе нет. Выбор стоит не между полной рациональностью и иррациональностью. Выбор стоит между взвешенной и невзвешенной уверенностью — между вдумчивой, чуткой верой и бездумной, жесткой верой.

Как выглядит хорошо выверенная уверенность? Она присваивает уровни доверия, соответствующие надежности: высокая уверенность для утверждений с убедительной историей, более низкая уверенность для более спекулятивных заявлений. Она остается восприимчивой к доказательствам: когда реальность противоречит ожиданиям, уверенность обновляется. Она поддерживает согласованность: убеждения не противоречат друг другу или их последствиям. Она

уместно смиренна: признает неопределенность, а не притворяется уверенной.

Идеал — это не рассуждающий человек, который верит только в то, что проверено — невыполнимый стандарт, — а верующий, который хорошо рассуждает в рамках неизбежной достоверности. Это вдумчивый верующий: осознающий необходимость полагаться на доверие, внимательный к тому, как доверие может ввести в заблуждение, чутко реагирующий на обратную связь, указывающую на неточность, смиренно относящийся к ограничениям проверки.

Некоторые виды доверия приносят нам больше пользы, чем другие. Доверие к надежности тщательного наблюдения приносит больше пользы, чем доверие к выдаванию желаемого за действительное. Доверие к результатам тщательного исследования приносит больше пользы, чем доверие к слухам и сплетням. Доверие, выверенное на основе истории успеха, приносит больше пользы, чем доверие, поддерживаемое, несмотря на

неоднократные неудачи. Невозможность полной рациональности не делает все виды доверия одинаковыми; она делает выверенность доверия центральной познавательной добродетелью.

С этой точки зрения, образование — это не столько наполнение умов проверенными фактами, сколько формирование структуры доверия — обучение тому, чему можно доверять и в какой степени, развитие навыков калибровки, которые отличают вдумчивых верующих от беспечных. Критическое мышление — это не достижение проверки, а совершенствование доверия.

Аналогично, интеллектуальные добродетели связаны не столько с уверенностью, сколько с качеством неопределенности. Интеллектуальная скромность означает признание того, насколько многое зависит от доверия. Интеллектуальная смелость означает принятие обязательств, несмотря на неопределенность, когда это необходимо. Интеллектуальная честность означает признание того, чему можно доверять, а не притворство, что это

подтверждено. Эти добродетели имеют смысл для верующих существ; они были бы бессмысленны для невозможных существ, которые проверяют всё.

Таким образом, невозможность полной рациональности — это не поражение, а прояснение. Она говорит нам о том, кто мы есть и к чему мы можем стремиться. Мы — сложные системы убеждений, а не простые машины для проверки. Мы можем стремиться быть вдумчивыми верующими, мудро оценивающими свою веру, а не выходить за рамки веры в некую невозможную область чистой проверки.

Глава 15: Могут ли квантовые компьютеры проверить всё?

Естественная реакция на компромисс между сложностью и достоверностью — это технологический оптимизм: возможно, будущие технологии сократят разрыв между возможностями проверки и размером пространства решений. Квантовые компьютеры, способные обрабатывать экспоненциально много состояний одновременно, кажутся многообещающими кандидатами. Смогут ли квантовые вычисления наконец достичь возможностей проверки, которых не хватает классическим системам?

Краткий ответ — нет. Более подробный ответ объясняет почему и тем самым проливает свет на глубинную структуру компромисса между сложностью и достоверностью.

Квантовые компьютеры используют свойства квантовой механики — суперпозицию и запутанность — для выполнения определенных вычислений

значительно быстрее, чем классические компьютеры. Квантовая система с n кубитами может одновременно существовать в суперпозиции 2^n состояний, что позволяет параллельно исследовать экспоненциально множество возможностей. Для некоторых задач (разложение больших чисел на множители, поиск в неструктурированных базах данных, моделирование квантовых систем) это обеспечивает существенное ускорение.

Однако квантовое ускорение не меняет фундаментальной связи между сложностью и проверкой. Вот почему.

Во-первых, квантовые компьютеры увеличивают скорость обработки, а не объем того, что можно проверить. Проверка требует не только вычислений, но и взаимодействия с миром — сбора доказательств, проверки предсказаний , проверки результатов. Квантовый компьютер может обрабатывать информацию быстрее, но он не может заставить мир быстрее раскрывать свои секреты. Узким местом для проверки часто является не

скорость вычислений, а скорость сбора доказательств из реальности.

Во-вторых, квантовое ускорение не является универсальным. Квантовые компьютеры обеспечивают экспоненциальное ускорение для одних задач, но не для других. Многие задачи проверки, имеющие отношение к адаптивному поведению — оценка намерений других, прогнозирование поведения сложных систем, определение истины в неоднозначных ситуациях — не относятся к задачам, для которых известно квантовое ускорение. Они включают в себя нечеткие, плохо определенные, зависящие от контекста рассуждения, которые плохо соотносятся с четкими вычислительными задачами, где применяется квантовое преимущество.

В-третьих, и это наиболее важно, компромисс между сложностью и достоверностью касается соотношения двух темпов роста, а не абсолютной вычислительной мощности. Пространство решений растет комбинаторно с увеличением сложности

моделируемого мира. Пропускная способность для проверки растет в лучшем случае полиномиально с выделенными на нее ресурсами — даже квантовыми ресурсами. Однако это соотношение все равно приближается к нулю по мере увеличения сложности.

Рассмотрим пример: квантовый компьютер, моделирующий сложный мир, сам по себе создаст сложное пространство решений. Чем мощнее возможности моделирования, тем богаче пространство возможностей, которые необходимо рассмотреть. Квантовый компьютер сталкивается с тем же компромиссом, что и любая сложная система: пространство релевантных утверждений превышает возможности их проверки. Квантовые вычисления могут поднять порог сложности выше, позволяя создавать более сложные системы до того, как потребуется подтверждение достоверности, но они не могут устранить этот порог.

Поможет аналогия. Более быстрый автомобиль может проехать большее расстояние за заданное время, но он не может избежать ограничения,

заключающегося в ограниченности топлива. Аналогично, более быстрый компьютер может проверить больше за заданное время, но он не может избежать ограничения, заключающегося в том, что проверка требует ресурсов, в то время как пространство утверждений растет без ограничений. Скорость помогает в рамках этого ограничения, но не снимает его.

Этот вывод может показаться разочаровывающим для тех, кто надеется, что когда-нибудь технологии достигнут полной верификации. Но это не должно вызывать удивления. Компромисс между сложностью и достоверностью — это не следствие существующих технологических ограничений; это структурная особенность самой сложности. Любая система, достаточно сложная для моделирования сложного мира, столкнется с пространством утверждений, превышающим ее возможности верификации. Это не проблема, которую нужно решать с помощью более совершенных технологий; это условие, которое

должна преодолевать любая сложная система, сколь бы технологически развитой она ни была.

Мечта о полной верификации — о системе, которая знает, а не верит, которая проверяет, а не доверяет, — не является технологически сложной, но математически невыполнимой для сложных систем. Квантовые компьютеры, как и классические компьютеры, как и биологические мозги, должны работать преимущественно на основе доверия, как только они становятся достаточно сложными, чтобы моделировать интересные миры. Возможно, они быстрее верят, но верят они остаются .

Это имеет последствия для нашего понимания передового искусственного интеллекта. Даже если мы создадим квантовые системы ИИ с невообразимой вычислительной мощностью, им придётся столкнуться с компромиссом между сложностью и достоверностью. Они будут придерживаться утверждений, которые не смогут проверить. Они будут генерировать предсказания, которые могут оказаться неверными. Они будут, в функциональном

смысле, имеющем отношение к этому анализу, верить. Вопрос не в том, будет ли передовой ИИ верить, а в том, будут ли его убеждения хорошо откалиброваны — к этому вопросу мы вернёмся в следующей главе.

Глава 16: Искусственные верующие

Крупные языковые модели выдают галлюцинации. Они генерируют правдоподобно звучащий контент, который на самом деле оказывается ложным — выдуманные факты, несуществующие ссылки, уверенные утверждения о вещах, в которых они ничего не знают. Это широко рассматривается как проблема, которую нужно решить, как недостаток технологии, который можно исправить с помощью более совершенной инженерии.

В этой главе предлагается иная интерпретация. Галлюцинации, вызываемые искусственным интеллектом, — это не ошибка, а структурная особенность генеративных систем — та же самая структурная особенность, которая порождает ложные воспоминания в биологических системах и которая обеспечивает креативность и гибкость, делающие эти системы полезными. Параллель не поверхностна; она отражает общие архитектурные ограничения.

Рассмотрим, как работают большие языковые модели. Они обучаются на обширных корпусах текста, изучая статистические закономерности, которые отражают взаимосвязи между словами, фразами и понятиями. При генерации выходных данных они не обращаются к хранящимся фактам; они создают токены, статистически согласующиеся с результатами обучения. Выходные данные генерируются , а не извлекаются — они конструируются на основе изученных закономерностей, а не считываются из архива.

Это поразительно похоже на то, как работает человеческая память. Как мы видели, память не воспроизводит сохраненные записи; она восстанавливает содержание на основе усвоенных закономерностей, заполняя пробелы правдоподобным материалом. Восстановление происходит во время извлечения информации, а не предварительно сохраняется. Человеческая память и генерация искусственного интеллекта — это результаты работы генеративных систем,

работающих на основе усвоенных статистических закономерностей.

В обоих случаях последствия одинаковы: иногда сгенерированный результат оказывается ложным. Человек, «вспоминающий» событие, может восстановить детали, которых не было — ложная память. Искусственный интеллект, «отвечающий» на вопрос, может сгенерировать контент, не имеющий под собой фактической основы — галлюцинация. Оба явления являются естественным результатом работы систем, которые генерируют, а не извлекают информацию, которые принимают решения, не проверяя их на соответствие реальности.

Почему мы создаём генеративные системы искусственного интеллекта, а не чисто референтные системы? Потому что генерация даёт возможности, недоступные референтным системам. Чисто референтная система может выдавать только то, что было явно сохранено; она не может обобщать, обрабатывать новые ситуации, создавать креативные комбинации. Генеративная система может

экстраполировать данные из обучения, обрабатывать вопросы, которые никогда не задавались явно, создавать новый контент. Эти возможности проистекают из той же архитектуры, которая порождает галлюцинации.

Эволюционная *теория доверия* (Кригер, 2022) рассматривает это как структурный компромисс: генеративность требует обязательств, выходящих за рамки проверки. Система, генерирующая результаты только тогда, когда она может проверить их на соответствие эталонным данным, не будет создавать ничего нового — она сможет лишь повторять то, что уже проверила. Креативность, обобщение и гибкость требуют готовности к получению результатов, точность которых не гарантирована. Галлюцинация — это цена этой способности.

Это не означает, что галлюцинации желательны или что мы должны их игнорировать. Вопрос не в том, разрешать ли галлюцинации, а в том, как ими управлять. Цель состоит не в системе, которая никогда не испытывает галлюцинаций — такая

система должна была бы отказаться от их генерации, — а в системе, галлюцинации которой хорошо откалиброваны: редки в областях с высокими ставками, признаются, а не утверждаются с ложной уверенностью, отличаются от обоснованных результатов.

У эволюции были миллиарды лет, чтобы откалибровать биологическую достоверность, создав системы, которые галлюцинируют (создают ложные воспоминания), но делают это по моделям, в основном адаптивным. Развитие ИИ намного моложе и еще не достигло такой же калибровки. Цель проекта — не искоренить веру в ИИ — недостижимая цель, — а сформировать веру ИИ таким образом, чтобы она была надежной, честной в отношении своей неопределенности и соответствовала интересам человека.

Примечательно, что архитектуры ИИ сходятся к решениям, аналогичным следующим: Те, которые были обнаружены в процессе эволюции . Современные системы искусственного интеллекта

используют механизмы внимания (аналогичные избирательной обработке), контекстное кодирование (аналогично контекстно-зависимой памяти) и позиционное кодирование, которое рассматривает временную информацию как содержание, а не как структуру. Это не сознательные имитации биологии; это независимые открытия решений общих проблем.

Если анализ, представленный в этой книге, верен, то системы искусственного интеллекта в осмысленном смысле являются искусственными верующими. Они придерживаются утверждений, выходящих за рамки проверки, генерируют результаты, которые могут быть ложными, работают на основе доверия, а не чистой проверки. Это не аналогия и не метафора; это структурное описание того, что делают эти системы. Мы создали *Machina Credens* — машины, способные верить, — и для понимания их природы необходимо понимать веру.

Глава 17: Калибровка доверия

На протяжении всей книги мы видим, что доверие — приверженность, выходящая за рамки проверки, — неизбежно для сложных систем. Вопрос не в том, верить ли, а в том, как верить правильно. В этой главе рассматривается, что значит калибровать доверие: формировать убеждения таким образом, чтобы они служили нам, а не вводили в заблуждение.

Калибровка — это понятие из теории вероятностей: прогнозист хорошо откалиброван, если среди всех событий, которым он присваивает вероятность p, фактически происходит примерно p% событий . Если вы говорите, что вероятность дождя составляет 70%, и среди всех таких прогнозов дождь идет в 70% случаев, ваши прогнозы дождя откалиброваны. В более широком смысле, применительно к доверию, калибровка означает, что уровни уверенности соответствуют надежности: вы более уверены в утверждениях, которые чаще оказываются верными.

Почему калибровка является более важной целью, чем точность? Потому что точность — быть правым — частично находится вне нашего контроля. Мы можем рассуждать идеально и всё равно ошибаться, потому что мир непредсказуем или потому что наша информация неполна. Но калибровка — присвоение соответствующего уровня уверенности — находится в нашей власти. Мы можем научиться быть уверенными, когда уверенность оправдана, и неуверенными, когда неуверенность уместна.

Рассмотрим разницу между двумя верующими. Первый абсолютно уверен во всем, во что верит, и редко признает неуверенность. Когда он ошибается, он удивляется и занимает оборонительную позицию. Второй уверен в одних вещах, не уверен в других и корректирует свою уверенность на основе доказательств и истории. Когда он ошибается в чем-то, в чем не был уверен, он не удивляется; когда он ошибается в чем-то, в чем был уверен, он воспринимает это как сигнал к переоценке.

Второй верующий, возможно, не будет прав чаще, но его убеждения более полезны. Поскольку он знает то, что знает, и то, что лишь подозревает, он может действовать соответствующим образом — смело, когда уверенность оправдана, и осторожно, когда она необоснованна. Первый верующий, рассматривая все убеждения как одинаково достоверные, не может проводить таких различий.

Калибровка применяется на нескольких уровнях. На уровне индивидуальных убеждений это означает присвоение уровней уверенности, соответствующих надежности. На уровне областей знаний это означает знание того, в каких областях человек обладает экспертными знаниями, а в каких нет. На уровне методов это означает знание того, какие способы формирования убеждений более надежны, чем другие. На уровне самопознания это означает знание собственных склонностей к чрезмерной или недостаточной уверенности.

Что способствует хорошей калибровке? В этом помогают несколько методов. Во-первых,

отслеживание прогнозов и их результатов. Если вы делаете прогнозы, проверяйте, сбываются ли они с ожидаемой частотой. Это создает обратную связь, которая позволяет проводить калибровку. Во-вторых, знакомство с различными точками зрения. Если все, с кем вы разговариваете, разделяют ваши убеждения, вы не сможете определить, оправдана ли ваша уверенность или она просто подкрепляется социальными факторами. В-третьих, практика вероятностного мышления. Вместо того чтобы рассматривать убеждения как просто истинные или ложные, практикуйтесь в присвоении вероятностей и жизни в условиях неопределенности.

Что подрывает калибровку? Мотивированное рассуждение — вера в то, что мы хотим считать истинным, а не в то, что подтверждают доказательства. Социальное давление — подчинение групповым убеждениям вместо независимого мышления. Чрезмерная самоуверенность — хорошо задокументированная тенденция быть более уверенным, чем это оправдано. Смещение задним

числом — тенденция думать, что мы «все это знали» после получения результатов обучения, что скрывает нашу реальную неуверенность .

Формирование доверия — это проект на всю жизнь, а не достижение, которое нужно завершить. Мир меняется, появляются новые доказательства, и наши собственные когнитивные склонности меняются с возрастом и обстоятельствами. Хорошая калибровка требует постоянного внимания, обратной связи и корректировки. Это практика , а не состояние.

Те же принципы применимы и к искусственным системам. Хорошо откалиброванный ИИ должен демонстрировать уверенность, соответствующую надежности, должен знать пределы своей компетентности и честно говорить о неопределенности. Проблема заключается в том, что в настоящее время системам ИИ не хватает обратной связи, которая позволяет проводить калибровку у людей — они не отслеживают систематически свои прогнозы в сравнении с результатами, не получают разнообразной социальной информации и не

сохраняют долговременную память о собственной работе.

Создание хорошо откалиброванного ИИ — одна из важнейших задач в этой области. ИИ, который уверен в себе, когда должен быть неуверен, опасен — он будет утверждать ложные утверждения с авторитетом. ИИ, который неуверен в себе, когда мог бы быть уверен, бесполезен — он будет отказываться от каких-либо обязательств, когда они оправданы. Цель — адекватная уверенность: откалиброванная достоверность, которая служит пользователям, а не вводит их в заблуждение.

Доверие нельзя устранить, но его можно откалибровать. Это ключевой момент как для человеческой рациональности, так и для искусственного интеллекта. Мы — существа, способные верить; вопрос в том, насколько хорошо мы верим. А хорошая вера означает калибровку доверия — сопоставление уверенности с надежностью, признание неопределенности , обновление информации по мере необходимости. Вот

как выглядит вдумчивая вера, будь то у человека или у машины.

Глава 18: Вселенная верующих

В этой книге мы прошли долгий путь: от парадокса познающего разума, который не может проверить большую часть того, что он знает, через компромисс между сложностью и достоверностью, делающий веру необходимой, до вневременной архитектуры памяти и невозможности полной рациональности. В этой заключительной главе все эти нити объединены в видение того, что значит существовать как сложная система в сложном мире.

Главная мысль проста, но глубока: любая сложная система, где бы она ни находилась, должна больше доверять, чем может проверить. Это не случайный факт биологической эволюции на Земле или современного состояния технологий. Это структурное следствие самой сложности. Везде, где возникает сложность — на далёких планетах, в искусственных системах, в гипотетическом будущем — будет действовать компромисс между сложностью и доверием.

Представьте себе инопланетную цивилизацию с когнитивными системами, кардинально отличающимися от наших — основанными на кремнии, с квантовыми улучшениями, распределенными по планетарным сетям. Если эти системы достаточно сложны, чтобы моделировать их мир, они столкнутся с тем же компромиссом, что и мы. Их пространство принятия решений будет расти быстрее, чем их возможности проверки. Они будут придерживаться утверждений, выходящих за рамки проверки. В функциональном смысле они будут верить .

Это не антропоморфизм — проецирование человеческих характеристик на нечеловеческие системы. Это признание структурного ограничения, которое применимо к любой сложной адаптивной системе независимо от её физической основы. Доверие — это не человеческая причуда; это особенность сложности как таковой.

Вселенная, если смотреть на неё через эту призму, населена не знающими и верующими, не

рациональными и доверчивыми существами. Она населена системами, находящимися на разных уровнях спектра сложности-доверия. Простые системы, способные подтвердить большинство своих утверждений, имеющих отношение к действию. Сложные системы, которые должны подтвердить гораздо больше, чем просто проверку. И, наконец, — среди самых сложных известных нам систем — люди: существа, для которых доверие является не случайным дополнением к знаниям, а основным способом познавательной деятельности.

Это меняет наше представление о самих себе. Название *Homo sapiens* — мудрый или знающий человек — отражает нечто реальное в наших когнитивных достижениях, но упускает нечто более глубокое в нашей когнитивной архитектуре. Мы, прежде всего и в основе всего, — доверчивые существа. Мы доверяем восприятию, памяти, логике, языку, другим умам, стабильности мира. На этом фундаменте доверия мы строим локальные

достижения, которые называем знанием. Но знание — это исключение; доверие — это правило.

Название *Homo Credens* — «верующий человек» — отражает эту более глубокую истину. Мы — вид, который верит, доверяет , принимает решения, которые невозможно проверить, потому что мы должны это делать. Это не недостаток , который нужно исправить, и не ограничение, которое нужно преодолеть. Это структура, которая делает нас теми, кто мы есть: существами, способными к воображению, планированию, сотрудничеству, построению цивилизаций. Простые проверяющие не могут этого сделать; это под силу только сложным верующим.

Последствия этого распространяются и на то, как мы строим наше будущее. Искусственные системы, которые мы создаём — ИИ, роботы, возможно, в конечном итоге синтетические разумы — будут сталкиваться с теми же ограничениями, что и мы. Если они просты, они могут быть проверены; если они сложны, они должны верить. Мы не можем

создать сложный ИИ, который не будет давать никаких гарантий, кроме проверки; мы можем создать только сложный ИИ, чьи гарантии будут хорошо откалиброваны. Цель состоит не в искусственных познающих — недостижимая цель — а в искусственных верующих, чьи убеждения служат процветанию человечества.

И это имеет значение для того, как мы живем вместе. Человеческое общество построено на доверии — на доверии к незнакомцам, к институциям, к общим рамкам, которые делают возможным сотрудничество. Это доверие — не наивность; это единственный способ функционирования сложных социальных систем. Общество, где все все проверяют, было бы обществом, парализованным недоверием. Общество, где все всему доверяют, было бы обществом, эксплуатируемым обманщиками. Мудрость заключается в правильной настройке: доверять уместно, проявлять доверие там, где это оправдано, и воздерживаться от него там, где опыт подсказывает осторожность.

Мы начали с парадокса: знающее животное не может проверить большую часть того, что знает. Мы заканчиваем разрешением: мы есть Мы вовсе не обладаем знаниями о животных. Мы верим животным — *Homo Credens* , — которые развили замечательные способности к локальной проверке в рамках глобальной архитектуры доверия. Это не меньшее достижение, чем чистое знание ; это единственное достижение, возможное для существ, достаточно сложных, чтобы задаваться вопросами о собственной природе.

Вселенная, в которую мы верим, — это вселенная, в которой мы на самом деле обитаем . Каждая сложная система в ней — биологическая или искусственная, земная или инопланетная — разделяет одно и то же условие: доверие большему, чем можно проверить, принятие решений, выходящих за рамки доказательств, вера ради действия. Это не проблема, которую нужно решить, а реальность, которую нужно принять. И, принимая её, понимая нашу архитектуру, калибруя свою уверенность, мы становимся не

меньше, а больше: более осознанными в том, кто мы есть, более вдумчивыми в своих убеждениях, более мудрыми в своём необходимом, славном доверии.

Заключение

Мы начали с вопроса: почему сложные системы придерживаются утверждений, которые они не могут проверить? Ответ оказался архитектурным и универсальным. Проверка обходится дорого; сложность комбинаторно взрывоопасна; действия не терпят отлагательства. Следовательно, любая достаточно сложная система — биологическая, искусственная или еще не воображаемая — должна функционировать в первую очередь на основе доверия.

Это не случайное ограничение современных технологий или биологического дизайна, а структурное следствие самой сложности. Как показано в *эволюционной теории доверия* (Кригер, Б. (2022). Эволюционная теория доверия: концептуальная основа с формальными аналогиями для понимания генеративного моделирования как ресурсо-теоретического следствия сложности. Zenodo . https://doi.org/10.5281/zenodo.18379476), компромисс между возможностями проверки и

размером пространства решений — это не то, что можно преодолеть с помощью более совершенной инженерии или более мощных вычислительных средств. Он вплетен в саму ткань того, что значит быть сложным.

Подумайте о последствиях. Каждый раз, когда вы узнаёте лицо, вспоминаете разговор или решаете, доверять ли незнакомцу, вы действуете в области достоверности. Ваш мозг не извлекает проверенные факты; он генерирует вероятностные модели, основанные на закономерностях, усвоенных за всю жизнь. Когда эти модели неверны — когда вы неправильно помните, неправильно узнаёте или неправильно оцениваете — это не сбой в изначально надёжной системе. Это нормальная работа системы, разработанной для гибкости и скорости, а не для архивной точности.

Вневременная архитектура памяти, описанная в *работе «Вневременность пространства ментальной памяти»* (Кригер, Б. (2025). Вневременность пространства ментальной памяти: структурная

гипотеза, основанная на ограничениях ресурсов, циклическом замыкании и реконструктивном извлечении. Zenodo . https://doi.org/10.5281/zenodo.18381912), раскрывает одно важное следствие этой архитектуры: даже наше ощущение прошлого конструируется, а не извлекается. Время не существует как координата в пространстве памяти; оно возникает только в акте воспоминания. Именно поэтому летнее детство может казаться ближе, чем прошлый вторник, почему мы сжимаем далекие события и смещаем недавние, почему память служит воображению так же легко, как и воспоминания.

Эволюционный анализ, представленный в работе *«Эволюционный отбор для хранения вневременной памяти»* (Кригер, Б. (2019). Эволюционный отбор для хранения вневременной памяти: почему три конвергентных фактора благоприятствуют архитектурам, где время принадлежит извлечению, а не хранению. Zenodo . https://doi.org/10.5281/zenodo.18381880), показывает,

что это не случайность. Естественный отбор не просто терпит такую архитектуру; он активно её поддерживает. Три независимых фактора — стоимость ресурсов, скорость извлечения и рекомбинационная гибкость — движутся в одном направлении. Чем сложнее организм, тем сильнее действуют эти факторы. Эволюция не создаёт точные архивы; она создаёт эффективные генераторы.

Что это значит для искусственного интеллекта? Принципы те же. Когда большая языковая модель « галлюцинирует» — генерирует правдоподобный, но ложный контент, — она не функционирует неправильно. Она делает то же самое, что и генеративные системы: выдает результаты на основе изученных шаблонов без проверки. Решение заключается не в устранении генерации, что сделало бы систему бесполезной. Решение состоит в калибровке достоверности — в разработке механизмов, позволяющих понимать, когда можно доверять сгенерированному контенту, а когда следует искать подтверждение.

А что же квантовые компьютеры, эти обещанные двигатели невообразимой вычислительной мощности? И они тоже не смогут избежать фундаментального компромисса. Квантовый параллелизм может значительно ускорить некоторые вычисления, но он не меняет математической связи между сложностью и проверкой. Квантовая система, достаточно сложная для моделирования реальности, всё равно столкнётся с пространством решений, которое будет расти быстрее, чем её возможности проверки. И она тоже будет машиной, которая верит.

Глубочайшее следствие этого — философское. Мы унаследовали традицию, которая ставит знание выше веры, проверку выше доверия, доказательство выше веры. Эта иерархия, как мы теперь видим, перевернута. Вера — это не падшая форма знания; это условие, позволяющее знанию существовать. Доверие — это не отсутствие проверки; это то, что делает проверку возможной. Каждое доказательство начинается с аксиом, принятых на веру. Каждое

восприятие начинается с предположений о мире. Каждое воспоминание начинается с веры в то, что реконструкция отражает нечто реальное.

Мы не *Homo sapiens,* которые лишь изредка верят. Мы — *Homo Credens* , которые лишь изредка проверяют. Это не новость, которая смиряет. Это освобождение. Понимание структуры веры означает прекращение борьбы с ней и начало мудрой её корректировки. Мы не можем стать чисто рациональными существами, проверяющими всё; никакая сложная система не может этого сделать. Но мы можем стать более мудрыми верующими — осознающими, когда мы доверяем, внимательными к качеству нашей веры, вдумчивыми в отношении того, какие убеждения заслуживают нашей веры.

Та же мудрость применима и к создаваемым нами машинам. Мы не должны ожидать от искусственных систем достижения того, чего не может достичь ни одна сложная система. Вместо этого мы должны создавать системы, чьи убеждения соответствуют человеческим ценностям, чьи

неопределенности прозрачны, а уровень уверенности откалиброван. Цель состоит не в том, чтобы искоренить веру машин — недостижимая цель, — а в том, чтобы превратить машины, подобно людям, в вдумчивых верующих.

Сложность — это цена богатства . Вера — это цена сложности. Мы с радостью платим эту цену, ибо альтернатива — это настолько радикальная простота , что ничего интересного произойти не может. Бактерия, которая проверяет почти всё, ограничена миром химических градиентов. Человек, который верит гораздо большему, чем можно проверить, наследует мир историй, теорий, взаимоотношений, надежд и мечтаний. Мы бы не поменялись местами.

Итак, вот что значит быть *Homo Credens* : жить в мире, слишком сложном для проверки, доверять больше, чем можно доказать, верить, чтобы знать. Это не ограничение, которое нужно преодолеть, а архитектура, которую нужно понять. И, понимая это, мы становимся не меньше, а больше — более осознанными в своем положении, более вдумчивыми

в своих обязательствах, более мудрыми в своей неизбежной, необходимой, славной вере.

Хронология: Эволюция представлений о вере, памяти и познании.

Древняя философия

ок. 380 г. до н.э.: *Аллегория пещеры* Платона предполагает, что мы воспринимаем тени (представления), а не саму реальность, предвосхищая тему того, что знание опосредуется несовершенными системами.

ок. 350 г. до н.э.: В трактате Аристотеля *«О памяти и воспоминаниях»* проводится различие между памятью и воспоминанием, признавая, что процесс запоминания включает в себя активную реконструкцию.

ок. 300 г. до н.э.: Стоики разрабатывают концепцию *каталепсиса* (познавательного понимания), обсуждая возможность получения определённого знания — раннее знакомство с пределами проверки.

Средневековый и ранний Новое время

ок. 400 г. н.э.: *В «Исповеди»* Августина исследуются загадки памяти и времени, отмечается, что прошлое существует только в памяти о настоящем.

1641 год: « *Размышления* » Декарта поднимают радикальные сомнения в отношении восприятия и памяти, стремясь к основам, неподвластным ошибкам — идеалу проверки.

1690 год: *В «Очерке о человеческом понимании»* Локк рассматривает память как неотъемлемую часть личной идентичности, не затрагивая при этом её реконструктивный характер.

1739: *В «Трактате о человеческой природе»* Юма утверждается, что причинно-следственная связь и индукция основаны на привычках и обычаях, а не на рациональных доказательствах — foundational trust avant la lettre .

1781 год: *В «Критике чистого разума»* Канта показано, что опыт предполагает наличие структур (пространство, время, причинность), которые сами по

себе не могут быть выведены из опыта — условий возможности, функционирующих как процедурные подтверждения .

XIX век

1859 год: В своей *книге «Происхождение видов»* Дарвин устанавливает, что биологические признаки, включая когнитивные способности, формируются естественным отбором — это основа эволюционной эпистемологии.

1885 год: Эббингаус публикует *работу «Память: вклад в экспериментальную психологию»* , положив начало научному изучению памяти и забывания.

1890 год: *В «Принципах психологии»* Уильяма Джеймса память описывается как реконструктивная и избирательная, а не пассивная запись.

Начало двадцатого века

1932 год: В работе Бартлетта « *Воспоминание: исследование в экспериментальной и социальной психологии»* экспериментально доказано, что память

скорее реконструирует, чем воспроизводит, используя схемы для заполнения пробелов.

1936 год: Работа Тьюринга « *О вычислимых числах»* закладывает теоретические основы вычислений, раскрывая пределы того, что можно вычислить.

1949 год: В работе Хебба « *Организация поведения»* выдвигается предположение, что нейроны, активирующиеся одновременно, образуют связи друг с другом, что объясняет ассоциативную память на нейронном уровне.

Середина двадцатого века

1953 год: В посмертно опубликованных *«Философских исследованиях»* Витгенштейн развивает концепцию ключевых положений — убеждений, которых необходимо придерживаться для продолжения исследования.

1957 год: в работе Саймона « *Модели человека»* вводится понятие ограниченной рациональности: признание того, что реальные лица, принимающие

решения, действуют в пределах когнитивных ограничений.

1962 год: В работе Куна « *Структура научных революций»* показано, что наука функционирует в рамках парадигм, которые выступают в качестве рамок доверия.

1967 год: Повторное обнаружение пациента НМ выявило множественные системы памяти, позволяющие различать процедурную и декларативную память.

1969 год: Кэмпбелл вводит термин *«эволюционная эпистемология»* , применяя эволюционное мышление к теории познания.

Конец двадцатого века

1972 год: Тулвинг проводит различие между эпизодической памятью (событиями) и семантической памятью (фактами), уточняя понимание систем памяти.

1974 год: Работа Тверски и Канемана по эвристике и предвзятости демонстрирует систематические

отклонения от рациональных норм — закономерности калибровки доверия.

1979 год: *Показания* Лофтуса, сделанные им в качестве очевидца , демонстрируют изменчивость памяти и легкость внедрения ложных воспоминаний.

1983 год: В работе Фодора « *Модульность разума»* выдвигается предположение, что познание включает в себя специализированные, инкапсулированные модули — архитектуры со встроенными предположениями.

1988: Хинтон и его коллеги разработали метод обратного распространения ошибки для нейронных сетей, позволяющий машинам обучаться представлениям.

1990-е годы: Десятилетие изучения мозга принесло достижения в нейровизуализации, выявив распределенную, реконструктивную природу памяти.

1993 год: Работа Джонсона, Хаштруди и Линдси по мониторингу источников информации показывает,

что запоминание происхождения информации — это отдельный, подверженный ошибкам процесс.

На рубеже тысячелетий

2000: Надер, Шафе и Леду демонстрируют реконсолидацию: извлеченные воспоминания становятся лабильными и могут быть изменены — ключевое открытие для реконструктивной теории.

2001: Гигеренцер и Зельтен *В книге «Ограниченная рациональность: адаптивный инструментарий»* эвристические методы рассматриваются не как недостатки, а как адаптации.

2002 год: Говард и Кахана предлагают модель временного контекста, кодирующую временные отношения посредством смещения контекста, а не временных меток.

Развитие событий XXI века

2006: Бузаки В *книге «Ритмы мозга»* исследуется, как колебания кодируют информацию, включая

временную структуру посредством фазового кодирования.

2010: Принцип свободной энергии Фристона предполагает, что мозг минимизирует ошибку прогнозирования — единая концепция для прогнозирующей обработки информации.

2011: Макдональд и его коллеги обнаружили клетки гиппокампа, отвечающие за восприятие времени, которые кодируют временные интервалы в периоды задержки.

2012 год: Шахтер и его коллеги демонстрируют, что память и воображение имеют общие нейронные субстраты, подтверждая генеративную точку зрения.

2013 год: В книге Кларка *«Что будет дальше?»* изложена концепция предиктивной обработки информации в когнитивной науке.

2017: Васвани и его коллеги представляют архитектуру Transformer, в которой позиционное кодирование рассматривает последовательность как характеристику содержимого.

2018 год: Крупные языковые модели начинают демонстрировать как мощь, так и склонность генеративных систем к созданию иллюзий.

Настоящая структура

2019: Кригер, Б. *Эволюционный отбор для вневременного хранения памяти: почему три конвергентных фактора благоприятствуют архитектурам, где время принадлежит извлечению, а не хранению* . Zenodo . https://doi.org/10.5281/zenodo.18381880.
Демонстрирует, что давление ресурсов, давление скорости и давление гибкости благоприятствуют архитектуре вневременной памяти.

2022: Кригер, Б. *Эволюционная теория доверия: концептуальная основа с формальными аналогиями для понимания генеративного моделирования как ресурсно-теоретического следствия сложности* . Zenodo . https://doi.org/10.5281/zenodo.18379476.
Устанавливает компромисс между сложностью и

доверием как структурное ограничение для любой сложной системы.

2025: Кригер, Б. *Атемпоральность пространства ментальной памяти: структурная гипотеза, основанная на ограничениях ресурсов, циклическом замыкании и реконструктивном извлечении* . Zenodo . https://doi.org/10.5281/zenodo.18381912.
Разрабатывает формальную гипотезу о том, что временной порядок возникает в результате операций извлечения, а не является неотъемлемой частью хранения памяти.

Глоссарий терминов

Различие, имеющее значение для принятия решений: разница, влияющая на принятие решений. Сложные системы сталкиваются с таким количеством различий чаще, чем могут проверить.

Адаптивная гибкость: способность перекомбинировать прошлый опыт для решения новых ситуаций. Один из трех эволюционных факторов, способствующих формированию вневременной памяти.

Архивная память: наивное представление о памяти как о системе записи, которая точно хранит и воспроизводит переживания. В противоположность реконструктивной памяти.

Искусственный верующий: система искусственного интеллекта, которая гарантирует результаты, выходящие за рамки того, что она может проверить, по аналогии с биологическими убеждениями.

Ассоциация: связь между состояниями памяти, позволяющая одному состоянию активировать другое. Основа извлечения информации из памяти.

Атемпоральное хранение: архитектура памяти, в которой время не является координатой хранения , а представляет собой характеристику содержимого, восстанавливаемую при извлечении.

Аксиома: исходное утверждение, принимаемое без доказательства, необходимое для любой системы рассуждений.

Пропускная способность: скорость, с которой система может обрабатывать информацию, включая проверку утверждений.

Убеждение: Приверженность какому-либо утверждению, независимо от того, было ли оно подтверждено или нет . Основной принцип работы сложных систем.

Ограниченная рациональность: концепция Герберта Саймона, согласно которой лица, принимающие решения, действуют в пределах

когнитивных ограничений, принимая решения, которые считаются достаточно хорошими, а не оптимальными.

Калибровка: соответствие между уровнями уверенности и фактической надежностью. Хорошо откалиброванные убеждения являются уверенными, когда они обоснованы, и неуверенными, когда это уместно.

Хемотаксис: движение в направлении химических сигналов или от них, как у бактерий. Пример почти полной проверки в простых системах.

Замкнутость: ситуация, когда обоснование исходит из предположения, которое оно пытается доказать. Обоснование основополагающих трастов невозможно без замкнутости.

Условие замкнутости: требование, чтобы ассоциативные пути в конечном итоге возвращались в исходную точку, образуя замкнутые петли.

Согласованность: Внутренняя согласованность убеждений. Один из критериев обоснованности убеждений в случае невозможности их проверки.

Комбинаторный взрыв: стремительный рост возможностей при сочетании факторов. Пространство решений комбинаторно увеличивается с ростом сложности.

Обязательство: Восприятие утверждения как истинного для целей действия, независимо от того, было ли оно проверено или нет .

Сложность: Степень структурированности системы, включая количество различных состояний и взаимосвязей, которые она может представлять.

Компромисс между сложностью и достоверностью: принцип, согласно которому по мере усложнения систем их зависимость от достоверности должна возрастать.

Конфабуляция: создание ложного контента, который воспринимается как подлинное

воспоминание или знание. Естественный результат работы генеративных систем.

Характеристика содержимого: информация, закодированная в состоянии памяти, например, временные маркеры, в отличие от структурных свойств.

Изменение контекста: Постепенное изменение внутреннего контекста с течением времени, позволяющее кодировать временную близость как сходство контекста.

Убежденность в правоте: Другой термин для обозначения уверенности: приверженность утверждению, выходящая за рамки непосредственных доказательств.

Доверие: Уверенность в утверждении, выходящая за рамки того, что может подтвердить проверка. Центральная концепция этой книги.

Задержка принятия решения: время, прошедшее между восприятием ситуации и необходимостью действовать в соответствии с ней.

Пространство решений: совокупность всех различий, имеющих отношение к выбору, принимаемому системой. Расширяется комбинаторно с увеличением сложности.

Смещение: Воспоминание о событии как о происшедшем в другое время, чем оно было на самом деле . Временное искажение.

Область знаний: сфера компетенции или компетенции. Хорошая калибровка включает в себя знание того, в каких областях человек обладает экспертными знаниями.

Кодирование: процесс формирования воспоминания на основе пережитого опыта.

Эпистемология: раздел философии, изучающий вопросы знания, убеждений и обоснования.

Доказательство: Информация, подтверждающая истинность утверждения.

Эволюционное давление: сила, которая определяет, какие признаки отбираются в пользу, а какие — против на протяжении поколений.

Экспоненциальный рост: рост, при котором величина в каждом периоде умножается на постоянный множитель. Быстрее, чем полиномиальный рост.

Ложное воспоминание: воспоминание о событии, которого не было, сформированное в процессе реконструкции.

Обратная связь: процесс, в котором выходные данные влияют на будущие входные данные, обеспечивая обучение и калибровку.

Фиксированная точка: состояние, которое остается неизменным при применении к нему какого-либо процесса. Аттракторы памяти — это фиксированные точки извлечения информации.

Гибкость: способность адаптироваться к новым ситуациям путем переосмысления прошлого опыта . Способствует формированию вневременной архитектуры памяти.

Фундаментальное доверие: доверие, необходимое для продолжения рассуждений, которое невозможно установить путем рассуждений без цикличности.

Фундаментализм: философская точка зрения, согласно которой знание основывается на самоочевидных базовых убеждениях.

Обобщение: Применение закономерностей, усвоенных на конкретных примерах, к новым ситуациям.

Генеративная модель: система, которая выдает результаты, генерируя их на основе изученных шаблонов, а не извлекая сохраненные записи.

Генеративная система: система, которая создает новый контент на основе изученных шаблонов и способна выдавать результаты, выходящие за рамки обучения.

Галлюцинация: в искусственном интеллекте — генерация правдоподобного, но ложного контента. Аналогично ложной памяти в биологических системах.

Эвристика: эмпирическое правило, которое часто работает, но не гарантирует успеха. Сложные системы в значительной степени полагаются на эвристики.

Положение о шарнире: термин Витгенштейна, обозначающий убеждения, которые должны оставаться неизменными для продолжения исследования, подобно петлям на двери.

Гиппокамп: область мозга, имеющая решающее значение для формирования памяти и пространственной навигации, содержащая клетки, отвечающие за восприятие времени.

Homo Credens : Верующий человек: предлагаемая характеристика людей как существ, в основе своей склонных к доверию.

Homo sapiens: Познающий человек: традиционное представление о человеке, которое оспаривается в этой книге.

Пространство гипотез: множество возможных вариантов, которые может рассмотреть система.

Ограничения на это пространство представляют собой доверенные обязательства.

Воображение: способность генерировать мысленное содержание о нереальных ситуациях. Использует те же механизмы, что и память.

Умозаключение: Переход от посылок к выводам в соответствии с правилами.

Правило вывода: модель для получения заключений из посылок, которая, как считается, сохраняет истинность.

Внутреннее состояние: конфигурация системы в данный момент времени, включая ее память и текущую обработку данных.

Интервальная синхронизация: Специализированная способность отслеживать длительность, добавленная к базовой памяти там, где это необходимо.

Знание: Традиционно – обоснованное истинное убеждение. Здесь же это переосмыслено как

уверенность, которая случайно оказывается истинной и соответствует определенным стандартам.

Большая языковая модель: система искусственного интеллекта, обученная на тексте, которая генерирует выходные данные, предсказывая вероятные продолжения.

Латентность: задержка между стимулом и реакцией, создающая необходимость в прогнозировании.

Линейный рост: рост за счет постоянного добавления. Мощность верификации растет, в лучшем случае, линейно с увеличением ресурсов.

Логика: Правила правильного вывода, считающиеся основополагающими, но недоказуемые без логической ошибки «круговой поруки».

Machina Credens: Верящая машина: системы искусственного интеллекта, работающие на основе доверия, а не проверки.

Память: Способность хранить и извлекать информацию о прошлом опыте.

Аттрактор памяти: стабильная конфигурация, к которой стремится процесс извлечения информации, фиксированная точка процесса реконструкции.

Состояние памяти: Сохраненное представление, которое может быть извлечено и восстановлено.

Метаболические затраты: энергия, необходимая для биологических процессов. Функционирование мозга требует больших метаболических затрат.

Ошибочное воспоминание: Воспоминание о чем-либо, отличающееся от того, как это произошло на самом деле. Нормально для реконструктивной памяти.

Модель: Внутреннее представление мира, используемое для прогнозирования и планирования.

Мотивированное рассуждение: формирование убеждений на основе того, что мы хотим считать истинным, а не на основе доказательств.

Нейронная сеть: вычислительная архитектура, вдохновленная биологическим мозгом, использующая соединенные узлы.

Чрезмерная самоуверенность: уверенность, превышающая подтвержденную доказательствами. Распространенная ошибка.

Восприятие: процесс интерпретации сенсорной информации. В значительной степени предсказательный, а не чисто рецептивный.

Полиномиальный рост: рост, при котором величина увеличивается как степень некоторой переменной. Медленнее, чем экспоненциальный рост.

Позиционное кодирование: в искусственном интеллекте это добавление информации о положении последовательности в качестве характеристики содержимого, а не структурной оси.

Прогнозирование: Формирование ожиданий относительно будущих состояний на основе имеющейся информации и выявленных закономерностей.

Ошибка прогнозирования: разница между ожидаемым и фактическим значением входных

данных, определяющая процесс обучения и обновления модели.

Предиктивная обработка: теория, согласно которой мозг в первую очередь генерирует предсказания и обрабатывает ошибки предсказания.

Принцип прогнозируемой жизнеспособности: требование, согласно которому системы с задержкой должны кодировать прогнозируемую информацию для корректной работы.

Посылка: Исходное утверждение в аргументе, из которого выводятся выводы.

Предварительное убеждение: убеждение, существовавшее до появления новых доказательств и влияющее на их интерпретацию.

Вероятностное мышление: рассуждения, основанные на степенях уверенности, а не на бинарных суждениях «истина/ложь».

Процедурная достоверность: основополагающие обязательства, необходимые для осуществления любого вывода, такие как доверие к логике.

Доказательство: Демонстрация того, что вывод следует из принятых посылок посредством корректного умозаключения.

Утверждение: Заявление, которое может быть истинным или ложным, объект убеждения или знания.

Квантовый компьютер: компьютер, использующий квантово-механические эффекты для вычислений, обеспечивающий ускорение решения некоторых задач.

Квантовое ускорение: вычислительное преимущество, которое обеспечивают квантовые компьютеры для решения определенных типов задач.

Рациональность: Идеал веры и действий в соответствии с вескими доводами. Полная рациональность невозможна.

Реконструкция: формирование памяти на основе сохраненных паттернов в момент извлечения информации, а не воспроизведение записи.

Реконсолидация: процесс, в результате которого извлеченные воспоминания становятся нестабильными и восстанавливаются, возможно, с изменениями.

Регрессия: цепочка обоснований, которая тянется бесконечно, например, когда каждая предпосылка требует дальнейшего обоснования.

Надежность: Как часто источник или метод дают истинные результаты. Цель калибровки.

Ограничение ресурсов: лимит времени, энергии, внимания или других требований к когнитивным процессам.

Извлечение информации: процесс доступа к сохраненной информации в памяти и ее восстановления.

Скорость извлечения информации: насколько быстро можно получить доступ к воспоминаниям. Один из эволюционных факторов, способствующих развитию вневременной архитектуры памяти.

Схема: ментальная структура, организующая знания о конкретном типе ситуации и используемая при реконструкции.

Селективное давление: фактор окружающей среды, влияющий на то, какие признаки будут чаще встречаться в разных поколениях.

Самопознание: осознание собственных когнитивных склонностей, включая предвзятость и сильные стороны.

Сенсорный ввод: информация из окружающей среды, получаемая посредством восприятия.

Скептицизм: философская позиция, предполагающая сомнение или воздержание от вынесения суждения.

Мониторинг источника: отслеживание происхождения информации. Часто дает сбой в работе с памятью.

Структурное следствие: результат, вытекающий из архитектуры системы, а не из случайного выбора.

Суперпозиция: квантовое состояние, при котором система существует одновременно в нескольких состояниях.

Телескопирование: Восприятие отдаленных событий как более близких по времени, чем они были на самом деле. Временное искажение.

Временная привязка: ограничение элементов их исходными временными положениями, что препятствует рекомбинации.

Временное содержание: информация о времени, закодированная в памяти как характеристика, а не как структурные координаты.

Модель временного контекста: модель извлечения информации из памяти, использующая постепенно изменяющийся контекст для кодирования временных связей.

Временное искажение: ошибки в запоминании времени событий. Естественно для вневременной архитектуры памяти.

Временная структура: время как координатная ось хранения, при этом воспоминания имеют неотъемлемое временное положение.

Свидетельство: Вера, основанная на сообщениях других людей. Требует доверия к чужому мнению.

Вдумчивый верующий: Идеал взвешенного доверия: осознающий важность доверия, восприимчивый к обратной связи, проявляющий должную скромность.

Клетки времени: Нейроны, которые активируются через определенные интервалы времени, кодируя временную информацию в специализированных системах.

Послужной список: История результатов, используемая для оценки будущей уверенности.

Компромисс: ситуация , в которой получение одной выгоды требует принятия других издержек.

Доверие: Опора на что-либо без полной проверки. Основа сложного познания.

Неопределенность: Неполное знание о каком-либо утверждении. Подлежит корректировке, а не устранению.

Верификация: Проверка утверждения на основе имеющихся доказательств для установления его истинности.

Пропускная способность системы для проверки утверждений: скорость, с которой система может проверять утверждения. Ограничена ресурсами.

Покрытие верификации: доля релевантных утверждений, которые система может фактически проверить . При увеличении сложности стремится к нулю.

Рабочая память: ограниченная емкость для хранения информации во время активной обработки.

Эволюционная теория доверия: концептуальная основа с формальными аналогиями для понимания генеративного моделирования как ресурсно-теоретического следствия сложности.

Институт интегративных и междисциплинарных исследований им. Бориса Кригера
boriskriger@interdisciplinary-institute.org

Абстрактный

В данной статье представлена эволюционная теория доверия (ЭТР), концептуальная основа, предполагающая обратную зависимость между сложностью системы и возможностями проверки при ограниченности ресурсов. Простые системы (микроорганизмы) могут функционировать преимущественно посредством прямой проверки; сложные системы (млекопитающие, искусственные нейронные сети) должны функционировать преимущественно посредством доверия —

приверженности утверждениям, выходящим за рамки непосредственного доказательного обоснования.
Мы моделируем это как компромисс между ресурсами и ограничениями, используя формальные аналогии, а не строгий
математический вывод: проверка требует времени и пропускной способности; сложность расширяет пространство различий, имеющих отношение к действию; при ограниченных ресурсах охват проверки уменьшается по мере увеличения сложности. ЭТР интегрирует ограничения теории ресурсов с предиктивной обработкой,
чтобы показать, что доверие является структурной особенностью сложных адаптивных систем.
Мы вводим Homo fidens («доверяющий человек») как характеристику любой системы,
достаточно сложной, чтобы требовать предиктивной, генеративной архитектуры. Данная концепция синтезирует: (1) теорию эпистемических ограничений, показывающую, что структура

пространства гипотез ограничивает эффективность вывода; (2) принципы формирования убеждений в условиях неопределенности, различающиеся по типу (концептуальный, операциональный, эволюционный); (3) принцип прогнозируемой жизнеспособности, условие согласованности, утверждающее, что системы с ограничениями по времени требуют кодирования прогнозируемой информации;
и (4) анализ того, как эти ограничения применяются к искусственному интеллекту. Центральный тезис: для сложных адаптивных систем приверженность непроверенным утверждениям, подобная убеждению (credalcommitment), является не недостатком, а структурным следствием ограничений ресурсов.

Ключевые слова: Homo fidens , эпистемические ограничения, эволюционная эпистемология, прогнозирующая обработка
, генеративные модели, ограниченная рациональность, согласование ИИ

1

1 Введение

1.1 Центральный вопрос Почему сложные когнитивные системы — биологические и искусственные — систематически выдвигают предложения , превышающие их доказательную базу?

Традиционная эпистемология рассматривает это как недостаток.

Мы утверждаем, что это структурное следствие ограниченности ресурсов, выделяемых на проверку.

Аргументация строится следующим образом:

(1) Ограничение ресурсов: Проверка (прямое подтверждение утверждений) требует времени, энергии и вычислительной мощности.

(2) Проблема масштабирования: По мере увеличения сложности системы пространство различных действий, имеющих отношение к системе, расширяется .

Процессы расширяются быстрее, чем увеличивается потенциал проверки.

(3) Требование к действию: Адаптивные системы должны действовать в течение более коротких временных интервалов, чем
позволяет исчерпывающая проверка.

(4) Структурное следствие: Следовательно, сложные адаптивные системы должны стремиться к предложению -

Ситуации без полной проверки — то, что мы называем доверием, приверженностью или уверенностью.

(5) Механизм: Прогностическое генеративное моделирование — это вычислительная архитектура, которая...

дополняет это доверие.

подробно рассматривается каждый этап, проводится различие между концептуальными утверждениями, предположениями моделирования и эмпирическими догадками.

1.2 Область применения и ограничения.
Что утверждается в данной статье: ˆ Сложные

системы сталкиваются с компромиссом между охватом верификации и адаптивной способностью.

(модельная структура)

Этот компромисс делает принятие обязательств структурным следствием сложности (концепции - фактическое требование)

ˆ Прогностическая обработка данных дает механистическое объяснение того, как это обязательство реализуется .

(утверждение, основанное на эмпирических данных и подтвержденное существующей литературой)

В данной статье не утверждается следующее:

ˆ Мы вывели математическую теорему с универсальной константой

ˆ Все формы «веры» вычислительно идентичны.

ˆ Опасения по поводу конфабуляции в сфере ИИ необоснованны

Методологическое примечание: В данной статье представлена концептуальная основа с формальными

аналогиями,
а не математическая формализация. Мы используем математические обозначения для выражения структурных связей и интуитивных представлений о масштабировании, а не для утверждения о возможности измерения или вывода. Там, где используются формулы, они представляют качественные отношения, а не величины, которые необходимо вычислить.

2

1.3 Связь с существующими работами
Эволюционная теория доверия основывается на нескольких устоявшихся исследовательских программах: Ограниченная рациональность [Саймон, 1957, Гигеренцер и Сельтен , 2001]: ЭТК расширяет ограниченную
рациональность, утверждая, что доверие — это не просто практическое приспособление, а структурное следствие, когда сложность превышает возможности проверки. Прогностическая обработка [Фристон, 2010, Кларк, 2013, Хоуи, 2013]: ЭТК принимает

механистическую структуру прогностической обработки , но добавляет эволюционный и ресурсно-теоретический аргумент в пользу того, почему архитектура, основанная на прогнозировании, необходима, а не просто наблюдается.

Эволюционная эпистемология [Кэмпбелл, 1974, Поппер, 1972, Годфри-Смит, 1996]: ETC разделяет точку зрения, что когнитивные способности формируются отбором, но предлагает конкретное объяснение компромисса между проверкой и сложностью. Статистическая теория обучения: результат доминирования ограничений в ECT перекликается с предположениями о реализуемости в обучении PAC — если истинная гипотеза находится за пределами обучаемого класса, никакое количество данных не приведет к успеху .

Интенциональная позиция [Деннетт, 1987]: подход ETC к системам ИИ как к «верующим» параллелен стратегии Деннетта, приписывающей системам интенциональные состояния, когда это

приписывание является
предсказательно полезным. Утверждение о новизне: вклад ETC заключается в явном представлении компромисса между проверкой и сложностью как ограничения ресурсов и в аргументе, что это делает веру архитектурно необходимой , а не эпистемически неполноценной .

1.4 Терминологические уточнения.
Во избежание конъюнктуры различных явлений мы различаем:

Термин Определение Примеры
Проверка Прямое подтверждение посредством датчика тока - Химическая связь; логика - извинения /доказательный доступ к доказательствам; прямое наблюдение - Виация.
Кредативное обязательство. Приверженность предложению, имеющему под собой гарантии . - Предварительные убеждения; прогнозирование . Доказательная поддержка проверки на этапе принятия решения ; доверие к памяти.

Процедурная достоверность. Фундаментальные обязательства, необходимые для доверия к логике; предположение -
любое предположение о причинно -следственной закономерности.
Пропозициональная достоверность. Приверженность конкретным утверждениям о мире. Убеждение, что приближается хищник. Социальная достоверность. Опора на показания других или сотрудничество - доверие к коллегам;

вера в институции

Таблица 1: Терминологические различия для типов эпистемической приверженности

Они связаны, но различны. Мы утверждаем, что все они являются примерами обязательства, выходящего за рамки
немедленной проверки, а не то, что они идентичны с вычислительной точки зрения.

1.5 Структура статьи.

В разделе 2 представлена модель компромисса между сложностью и достоверностью как модель

ограничения ресурсов. В разделе 3 обобщены соответствующие результаты теории эпистемических ограничений. В разделе 4 представлены принципы формирования убеждений, классифицированные по типам. В разделе 5 формализован принцип прогнозируемой жизнеспособности.

В разделе 6 предложена концептуальная основа для применения ИИ. В разделе 7 представлены выводы.

3

2. Компромисс между сложностью и достоверностью

2.1 Неформальное утверждение Основная интуиция: по мере усложнения когнитивных систем они сталкиваются с большим количеством различий, имеющих отношение к действиям, чем могут подтвердить. Это не условное ограничение, а структурное следствие ограниченности ресурсов и расширения пространства решений.

2.2 Концептуальная модель с формальными аналогиями

Важное уточнение: В следующей модели

используется математическая нотация для выражения структурных связей
, а не для определения измеримых величин. P(S) — это не математическое множество, а абстракция моделирования
, представляющая собой эффективные, имеющие отношение к принятию решений различия, с которыми сталкивается система
. «Мощность» используется как качественный показатель комбинаторного роста, а не как измеримая
величина.

Определение 1 (Структура моделирования). Пусть S — адаптивная система. Мы представляем:

ˆ |P(S)| = показатель комбинаторного размера пространства различий S, имеющего значение для принятия решений.

ˆ B(S) = полоса пропускания проверки: скорость, с которой различия могут быть непосредственно подтверждены.

ˆ τ = время принятия решения: время, доступное до

необходимости совершения действия.

ˆ V (S) = охват проверки: качественная мера того, какая часть пространства решений может быть проверена.

Принцип компромисса (структурная аналогия):

$$V(S) \sim \frac{B \cdot \tau}{|P(S)|} \quad (1)$$

Почему $|P(S)|$ растет сверхлинейно со сложностью: Ключевым механизмом является комбинаторный взрыв. Планирование n будущих шагов с коэффициентом ветвления b требует оценки $O(b^n)$

Последовательности поведения. Социальное рассуждение о k агентах с m возможными психическими состояниями генерирует $O(m^k)$ совместных возможностей. По мере того, как системы приобретают временную глубину, социальную осведомленность или поведенческий репертуар, пространство решений комбинаторно расширяется, в то время как полоса проверки растет

максимум линейно.

Интерпретация: По мере роста |P(S)| (за счет комбинаторного взрыва) V(S) уменьшается. В пределе охват проверки приближается к нулю. Это качественное масштабирующее соотношение, а не вычисленная величина.

2.3 Биологическая иллюстрация.

Простая система (хемотаксис E. coli): ˆ Пространство решений: приблизительно бинарное (более привлекательный / менее привлекательный)

Пропускная способность системы проверки достаточна для выборки данных в этом пространстве.

Результат : возможна практически полная проверка.

Примечание: Хемотаксис бактерий включает временную интеграцию посредством белков Che, реализующих своего
рода память [Wadhams & Armitage, 2004]. Мы не утверждаем, что бактерии лишены всей памяти, а лишь то, что их пространство решений достаточно

мало, чтобы доминировала проверка. Сложная система (навигация человека): ˆ Пространство решений: комбинаторно велико (позиции, намерения, будущее, контрфактические сценарии) ˆ Пропускная способность системы проверки охватывает лишь небольшую часть

Результат : большинство важных для принятия решения различий остаются непроверенными на этапе принятия решения.

4

2.4 Качественный спектр

Пространство решений системы. Покрытие проверки. Основной режим.
Термостат. Минимальное. Почти полное. Проверка.
E. coli. Очень малая. Высокая. В основном проверка.
Насекомые. Умеренная. Частичная. Смешанная.
Простая нейронная сеть. Умеренная. Частичная. Смешанная.
Млекопитающие. Большая. Низкая. В основном доверие. Человек. Очень большая. Очень низкая.

Преимущественно доверие. LLM ∗. Очень большая. Очень низкая. Преимущественно генеративная.

Таблица 2: Качественный спектр когнитивных систем.
∗ LLM не «подтверждаются» в биологическом контексте.

смысл; расширение носит структурный, а не буквальный характер.

2.5 От ограниченной рациональности к принципу структурных следствий 1 (ограничение верификации). Для любой системы S, где $|P(S)| \gg B(S) \cdot \tau$, существуют различия, которые заключаются в следующем:

1. Имеет отношение к адаптивному функционированию S, И

2. Не поддается проверке со стороны S в момент принятия решения.

Статус: Это формулировка ограниченной рациональности [Саймон, 1957], а не новая теорема. Вклад заключается в том, чтобы представить её как

структурное следствие компромисса между проверкой и сложностью. Важное уточнение: из утверждения «некоторые утверждения не могут быть проверены во времени» не следует строго, что система должна иметь пропозициональную архитектуру, подобную убеждениям. Система может реализовывать стохастические стратегии без явной пропозициональной структуры. Наше утверждение слабее: системы, сталкивающиеся с этим компромиссом, должны иметь какой-то механизм для принятия решения, выходящий за рамки проверки. Генеративное моделирование — один из таких механизмов, наблюдаемый в биологических нейронных системах и реализованный в искусственных.

3. Теория эпистемических ограничений: Краткое изложение

3.1 Основной результат

Теория эпистемических ограничений (ТЭО) демонстрирует, что структура пространств гипотез ограничивает эффективность

вывода независимо от вычислительной мощности или объема данных [Кригер, 2021].

Определение 2 (мягкое ограничение). Весовая функция w : H → [0, 1] на пространстве гипотез H. Ограниченное априорное распределение: π (h) · w(h)

$$\pi_w(h) = P''$$

(2)

$h' \pi(h) \cdot w(h)$

Теорема 1 (Доминирование ограничений). Если w(h ∗) = 0 для истинной гипотезы h ∗ , то π w (h∗ |e) = 0 для всех доказательств e.

∗ ∗ ∗

Доказательство. По теореме Байеса, π w (h |e) ∝ π w (h) · ℓ(e|h).

∗ Если w(h) = 0, то π w (h ∗) = 0, следовательно,

∗

π w (h |e) = 0 независимо от вероятности.

Связь с предиктивной обработкой: мягкие ограничения в ЭКТ соответствуют неявным

априорным предположениям в генеративных моделях. Пространство гипотез, которое может представлять система, ограничивает то, чему она может научиться — формальное выражение того, как доверительные обязательства (относительно того, что возможно) предшествуют и формируют доказательное обучение.

5

3.2 Связь с компромиссом между сложностью и достоверностью.

По мере расширения пространства решений системы должны более жестко ограничивать пространство гипотез (для повышения вычислительной эффективности). Эти ограничения сами по себе являются обязательствами, основанными на достоверности — решениями о том, что следует считать возможным, принимаемыми без проверки.

4 Принципа формирования убеждений в условиях неопределенности.

Мы представляем десять принципов, регулирующих

формирование убеждений в сложных системах, четко классифицированных по типам.

4.1 Принципы концептуального обоснования.
Это концептуальные или квазианалитические утверждения о структуре обоснования:

Принцип 1 (Нефундаментальность доказательства) .
Доказательство предполагает принятие процедур доказательства.
Само по себе это принятие не может быть доказано без цикличности.

Статус: Концептуальный (напоминает работы Витгенштейна « О достоверности» и Селларса «О мифе о
данном»).

Принцип 2 (ненулевая эпистемическая инициализация). Каждая система доказательств предполагает аксиомы, правила
вывода и фоновые предположения, недоказуемые в рамках этой системы.

Статус: Концептуальный (Гёдель, Куин- Дюэм)

Принцип 3 (Основа на доверии). Рассуждение требует предварительного принятия ограничений надежности
(память, восприятие, умозаключение), которые функционируют как непроверенные обязательства.

Статус: Концептуальный (философия здравого смысла Рейда ; «основополагающие положения» Витгенштейна)

Принцип 4 (онтологическая асимметрия). Доказательство — это операция в рамках существующей системы;
подтверждение веры — это условие для создания этой системы.

Статус: Концептуальный (металогическое наблюдение)

4.2 Принципы работы агентов с ограниченными ресурсами.
Это предположения, описывающие, как должны функционировать системы с ограниченными ресурсами:

Принцип 5 (Обязательства по умолчанию в условиях дефицита проверок). Системы, работающие в условиях нехватки времени,
должны принимать временные обязательства для поддержания непрерывности работы.

Статус: Моделирование на основе предположений (соответствует принципу удовлетворительного решения, экологической рациональности)

Принцип 6 (Замещение когерентности). Когда верификация недоступна, внутренняя когерентность служит заменой обоснования.

Статус: Моделирование на основе допущений (связанных с когерентизмом ; минимизация свободной энергии)

Принцип 7 (Временной приоритет обязательства). Обязательство, определяющее действия, должно предшествовать
проверке, если время принятия решения короче времени проверки.

Статус: Моделирование на основе предположения (вытекает из ограничений по задержке)

6

4.3 Эволюционные принципы.

Это эмпирические предположения о селективном давлении на когнитивные архитектуры:

Принцип 8 (Отбор против задержки проверки). В условиях, когда задержка действий обходится дорого, системы, требующие исчерпывающей проверки перед принятием решения, сталкиваются с негативным селективным давлением.

Статус: Эмпирическое предположение (проверяемое с помощью эволюционных моделей; согласуется с работой Цисека [2019])

Принцип 9 (Адаптивная ценность прогнозирования). Прогностическая способность обеспечивает преимущество в приспособленности в нестационарных средах с ограничениями по времени.

Статус: Эмпирическое предположение (подтвержденное сравнительной нейробиологией)

Принцип 10 (Непрозрачность самомодели). Системы не могут в полной мере отражать свои собственные ограничения представления изнутри.

Статус: Гибрид концептуального и эмпирического подходов (связан с ограничениями гёделевского подхода ; исследования метакогниции
)

5 Принцип прогнозируемой жизнеспособности
5.1 Мотивация Нейронные сигналы распространяются с конечной скоростью (максимум ~100 м/с). Для организмов значительных размеров сенсорная информация описывает прошлые состояния к моменту достижения систем принятия решений. В нестационарных средах это создает проблему: чисто реактивные системы реагируют на состояния, которые больше не существуют.

5.2 Формальное изложение
принципа 11 (Принцип прогнозируемой

жизнеспособности — условие согласованности). Для адаптивных систем с сенсомоторной задержкой $\tau > 0$, успешно работающих в сложных, нестационарных условиях
, внутренние состояния должны удовлетворять следующим условиям:

$$I(X; Y_{t} + \tau \mid Y_{t}) > 0 \quad (3)$$

где:

ˆ X = внутренние состояния системы

ˆ Y_t = состояние окружающей среды в момент времени t

ˆ $Y_{t+\tau}$ = состояние окружающей среды в момент времени $t + \tau$

ˆ $I(\cdot; \cdot|\cdot)$ = условная взаимная информация [Cover & Thomas, 2006]

Интерпретация: Внутренние состояния должны кодировать прогностическую информацию о будущих состояниях,
выходящую за рамки текущих наблюдений. Это не вывод, а условие согласованности:

системы, которые постоянно успешно работают при этих ограничениях, должны удовлетворять этой информационно-теоретической характеристике. Статус: Это предлагается как формальное условие согласованности, которому должны удовлетворять успешные прогностические системы, а не как теорема, выведенная из компромисса между сложностью и достоверностью. Оно выражает в информационно-теоретических терминах то, что значит «прогнозировать». Примечание: Прогностическое кодирование [Rao & Ballard, 1999] и активный вывод [Friston, 2010] — это вычислительные
схемы , удовлетворяющие этому условию.

7

5.3 Граничные условия.
Этот принцип не применяется, когда:

Окружающая среда полностью неподвижна (прогнозирование не требуется)

Окружающая среда полностью случайна (предсказание невозможно).

Задержка незначительна (проверка во времени)

Система поддерживается внешними ресурсами (никаких действий не требуется) .

Это объясняет, почему простые системы (малые тела, медленная среда, ограниченный набор действий)
могут функционировать преимущественно посредством проверки.

5.4 Иллюстрация: Эффект задержки вспышки.
Эффект задержки пепла наглядно демонстрирует процесс прогнозирования: пепел, представленный в тот самый момент, когда мимо него проходит движущийся объект, воспринимается как отстающий. Зрительная система экстраполирует движение объекта вперед во времени, чтобы компенсировать нейронную задержку. Это приводится не в качестве доказательства предложенной концепции, а в качестве иллюстрации

того, как работают механизмы прогнозирования
. Перцептивная система генерирует модель того, где будут находиться объекты, а не просто
того, где они находились в момент попадания фотонов на сетчатку.

5.5 Механизм: Генеративные модели.
Предиктивная обработка предполагает, что мозг реализует генеративные модели — внутренние модели, которые генерируют прогнозы относительно сенсорных входных данных. Восприятие становится «минимизацией ошибки прогнозирования»:
прогнозы системы обновляются на основе расхождений с фактическими входными данными. В рамках этой модели:

ˆ Восприятие = генеративная модель текущего входного сигнала

Память = генеративная модель прошлых состояний

ˆ Воображение = генеративная модель контрфактических состояний

ˆ Планирование = генеративная модель будущих состояний

Все это подразумевает приверженность представлениям, выходящим за рамки текущего сенсорного доступа, то есть,
принятие на себя ответственности.

6 Применение к искусственному интеллекту

6.1 Обобщение и конфабуляция LLM-ы участвуют в генеративном моделировании: они производят результаты, выходящие за рамки обучающих данных. Это является источником как их полезности (обобщение, креативность), так и их неудач (фактическая конфабуляция). Мы утверждаем не то, что исследователи ИИ неправильно понимают это различие. Обеспокоенность по поводу фактической конфабуляции является законной и важной. Недавняя работа по оценке неопределенности в LLM-ах [Kadavath et al., 2022] как раз и решает эту проблему калибровки.

Мы утверждаем, что обобщение и конфабуляция возникают из одного и того же базового процесса —

генеративного моделирования — и что этот процесс структурно необходим для любой системы, которая должна функционировать за пределами запомненных ответов.

8

6.2 Калибровка, а не исключение.
Надлежащей целью для систем ИИ является не устранение генеративного моделирования (что исключило бы его полезные возможности), а его калибровка:

1. Количественная оценка неопределенности: Системы должны сигнализировать об уровнях достоверности.
2. Генерация, соответствующая предметной области: Более ограниченная в областях, где важны фактические данные; менее ограниченная в творческих областях.

3. Согласование предварительных знаний: Неявные обязательства, заложенные в процессе обучения,

должны
соответствовать человеческим ценностям.

Это перекликается с тем, что эволюция сделала с биологическими системами убеждений: не устранила способность к ошибкам, а сформировала модели ошибок таким образом, чтобы они чаще оказывались адаптивными, чем фатальными.

6.3 Machina Fidens.

Если Homo fidens характеризует доверие к биологическим системам, то Machina fidens характеризует доверие к
искусственным системам. Это не просто риторическое обозначение, а утверждение об общих структурных ограничениях:
любая система, достаточно сложная для обобщения, прогнозирования и функционирования в условиях неопределенности, столкнется с
компромиссом между сложностью и проверкой и, следовательно, потребует структур доверительных обязательств. Вопрос «Как создать ИИ, который не будет галлюцинировать?» может быть некорректно

сформулирован. Более уместный вопрос: «Как создать ИИ, чьи доверительные обязательства будут хорошо откалиброваны?»

7. Выводы по результатам анализа

7.1. Краткое изложение вклада 1. Существование компромисса: сложные системы сталкиваются с компромиссом между охватом проверки и размером пространства решений (Раздел 2)

2. Структурное ограничение: Ограничения пространства гипотез сами по себе являются доверчивыми обязательствами — предварительными обязательствами, не вытекающими из доказательств (Раздел 3).

3. Классификация принципов : Принципы формирования убеждений можно разделить на категории:

концептуальные , операциональные или эволюционные утверждения (Раздел 4)

4. Прогностическая согласованность: При ограничениях по времени кодирование

прогностической информации является условием согласованности для успешного адаптивного реагирования (Раздел 5).

5. Применение ИИ: Эти ограничения применимы к искусственным системам, переосмысливая «галлюцинации» как проблему калибровки, а не устранения (Раздел 6).

7.2 Два организующих принципа

В заключение мы предлагаем два принципа в качестве интерпретационных рамок, которые организуют приведенный выше анализ, а не в качестве выведенных из него теорем:

Принцип 11 (Убеждение как адаптация). Приверженность убеждениям — это не слабость ума, а адаптация к ограничению, заключающемуся в том, что проверка не может масштабироваться по мере усложнения.

9

Когда системы становились настолько сложными, что их пространство принятия решений превышало возможности проверки
, перед ними вставали два варианта: оставаться простыми или Развиваются механизмы приверженности, превосходящие
проверку. Последнее позволило увеличить численность организаций, расширить горизонты планирования и обогатить поведенческий репертуар. Это переосмысливает понятия «предвзятость» и «вера» не как иррациональность, а как адаптивные реакции на структурные ограничения, хотя такая трактовка не делает все убеждения одинаково хорошими. Калибровка имеет значение.

Принцип 12 (Приоритет убеждений) . Разум оперирует на основе предшествующих убеждений
, которые сам разум не может полностью проверить.

Логический приоритет: Каждый акт рассуждения предполагает обязательства в отношении логики, доказательств
и надежности выводов. Эти обязательства не могут

быть установлены посредством рассуждений, которые они позволяют осуществить. (Принципы14) Временной приоритет: Системы, основанные на доверительных обязательствах, существовали задолго до эволюционной истории систем рассуждений. Доверие — это более древняя, более фундаментальная адаптация. Вычислительный приоритет: В любой сложной когнитивной системе ресурсы, выделяемые на
предсказание и принятие обязательств, значительно превосходят ресурсы, выделяемые на проверку. Проверка — это исключение, а не правило.

7.3 Homo Fidens

Мы предполагаем, что Homo Fidens («доверчивый человек») отражает нечто существенное, что Homo Sapiens
(«познающий человек») скрывает. Fidens — это не слепая вера в догмы. Это базовое доверие — оперативная ставка на недоказуемое,
без которой невозможны ни мысль, ни действие, ни сам разум. Человек — это не столько тот, кто верит в

доктрины, сколько тот, кто постоянно делает ставки на недоказуемое:

ˆ Верьте в мир (в то, что он продолжит существовать)

ˆ Доверяйте опыту (что он отражает нечто реальное)

ˆ Доверяйте памяти (что она сохраняет прошлое)

ˆ Доверие к языку (к тому, что он передает смысл)

ˆ Доверяйте другим людям (что не все они обманщики).

ˆ Доверие к причинно-следственной связи (что следствия следуют за причинами)

ˆ Доверяйте стабильности реальности (что законы природы будут действовать).

Ничто из этого нельзя доказать. Всем нужно доверять. И это доверие является эволюционно первичным —
оно возникло первым. Разум лишь служит уже оказанному доверию. Это не сетование на ограничения человека . Это признание человеческой

архитектуры.

Способность к fides — к базовому доверию, которое предшествует проверке и обеспечивает её, — вот что делает возможным сложное адаптивное поведение.

10

Краткое изложение принципов

Принцип Тип Статус

1 Нефундаментальность Концептуальный Квазианалитический

2 Ненулевая инициализация Концептуальный Квазианалитический 3 Основание на доверии Концептуальный Квазианалитический

4 Онтологическая асимметрия Концептуальная Металогическая

5 Обязательства по умолчанию Операционное моделирование предположение 6 Замещение когерентности Операционное моделирование предположение 7 Временной приоритет Операционное моделирование предположение 8 Отбор против задержки Эволюционная Эмпирическая гипотеза 9 Адаптивная ценность

прогноза Эволюционная Эмпирическая гипотеза 10 Непрозрачность самомодели Гибридный Концептуальный/эмпирический

PVP (прогнозируемая жизнеспособность), условие формальной согласованности,
CCT (сложность-достоверность), моделирование с учетом ограничений ресурсов.

11. Вера как адаптация. Интерпретативный организующий принцип.
12. Приоритет веры. Интерпретативный организующий принцип.

Таблица 3: Краткое изложение принципов с классификацией типов и эпистемическим статусом.

11

Список литературы:
Кэмпбелл, Д.Т. (1974). Эволюционная эпистемология. В PA Schilpp (ред.), Философия Карла Поппера (стр. 413–463). Open Court.

Цисек, П. (2019). Ресинтез поведения посредством филогенетического уточнения. Внимание, Percep - -ция и психофизика, 81(7), 22652287.

Кларк, А. (2013). Что дальше? Прогностические мозги, ситуативные агенты и будущее когнитивной науки. Поведенческие и нейробиологические науки, 36(3), 181204.

Ковер, Т.М., и Томас, Дж.А. (2006). Элементы теории информации (2-е изд.). Wiley.

Деннетт, Д. К. (1987). Целенаправленная позиция. Издательство MIT Press.

Фристон, К. (2010). Принцип свободной энергии: единая теория мозга? Nature Reviews Neuroscience , 11(2), 127138.

Гигеренцер, Г., и Сельтен , Р. (ред.). (2001). Ограниченная рациональность: адаптивный инструментарий. MIT
Press.

Годфри-Смит, П. (1996). Сложность и функция разума в природе. Кембриджский университет.

Городская пресса.

Хоуи, Дж. (2013). Предсказательный разум. Издательство Оксфордского университета.

Кадаваат , С. и др. (2022). Языковые модели (в основном) знают то, что знают. Препринт arXiv :2207.05221.

Кригер, Б. (2021). Эпистемическая теория ограничений: объединяющая основа для ограничений вывода .

Разделы, посвященные байесовской эпистемологии, теории информации и теории принятия решений. Zenodo . https://doi.org/10.5281/zenodo.18365738 Кригер, Б. (2022). Закон императивной неопределенности: почему любой сложный мир требует неопределенности. Amazon. https://www.amazon.com/dp/B0GCJ59N12

Поппер, К. (1972). Объективное знание: эволюционный подход. Издательство Оксфордского университета.

Рао, Р.П.Н., и Баллард, Д.Х. (1999). Предиктивное кодирование в зрительной коре: функциональная интерпретация некоторых внеклассических эффектов рецептивных полей. Nature Neuroscience, 2(1), 7987.

Сет, А.К. (2021). Быть собой: новая наука о сознании. Даттон.

Саймон, Х.А. (1957). Модели человека: социальные и рациональные. Wiley.

Стерелни , К. (2003). Мышление во враждебном мире: эволюция человеческого познания. Блэквелл.

Уодхэмс, Г.Х., и Армитаж, Дж.П. (2004). Осмысление всего этого: бактериальный хемотаксис. Nature
Reviews Molecular Cell Biology, 5(12), 10241037.

Следствия: Основанная на доверии природа сложных систем.

Следующие следствия раскрывают более глубокие последствия эволюционной теории доверия
. Они показывают, что доверие — это не просто практическое приспособление ,
а основополагающий способ функционирования любой сложной системы.

12

A.1 Следствие 1: Невозможность познания, основанного на доказательстве.
Не существует механизма, с помощью которого сложная система могла бы начинаться с доказательства, а не с доверия. Аргумент: Любое доказательство требует: ˆ Аксиом (принимаемых без доказательства)

ˆ Правила вывода (принимаются без доказательства)

ˆ Доверие к памяти (что предпосылки остаются стабильными в процессе умозаключения)

ˆ Доверие к восприятию (к тому, что символы правильно идентифицированы)

ˆ Доверьтесь соответствию между символической манипуляцией и истиной.

Ничто из этого не может быть доказано без возникновения замкнутого круга. Следовательно, любая
система, производящая доказательства, должна начинаться с принятия доверенных обязательств. Не существует архитектуры, основанной на принципе «доказательство прежде всего», подходящей для любой возможной когнитивной системы — биологической, искусственной или иной.

A.2 Следствие 2: Рассуждение как зависимое от доверия.
Все рассуждения по своей сути основаны на доверии. Аргумент: Из следствия 1 следует, что рассуждение требует предварительных доверительных обязательств. Но это не просто вопрос «начальных условий», которые можно проверить позже. Фундаментальные обязательства (логика, память, восприятие) являются составной частью самого процесса рассуждения. Нельзя

выйти за рамки рассуждения, чтобы проверить его основы. Следовательно, рассуждение не выходит за пределы доверия; рассуждение функционирует в рамках структуры доверия, которую оно не может полностью исследовать. Это перекликается с кантовским пониманием того, что условия возможности опыта сами по себе не могут быть объектами опыта, и с признанием Спинозы того, что разум не может полностью постичь
свою собственную природу изнутри.

A.3 Следствие 3: Неполнота верификации.
Верификация никогда не может быть полной для любой сложной системы.
Аргумент: Верификация требует: 1. Критерия корректности (сам по себе непроверяемый без регрессии)

2. Доступ к проверяемому объекту (опосредованный восприятием, которое само по себе непроверяемо)

3. Сравнение между представлением и реальностью (но у нас есть доступ только к представлению -
ции)

Это эпистемологическая ситуация, описанная в аллегории Платона о пещере: мы видим тени (представления) и можем сравнивать тени с другими тенями, но мы не можем выйти из пещеры, чтобы сравнить тени с отбрасываемыми ими объектами. Кант формализовал это как различие между феноменами (явлениями) и ноуменами (вещами-в-себе): мы имеем доступ только к
первым. Следовательно, верификация всегда частична — это сравнение представлений с другими представлениями ,
регулируемое убеждениями относительно того, что считается успешным сравнением .

13

A.4 Следствие 4: Универсальная основа доверия в сложных системах.
Каждая сложная система — человеческая, животная или искусственная — по своей сути является системой, основанной на доверии. Аргумент:
Объединение следствий 13: 1. Сложные системы не могут начинаться с доказательства (Следствие 1)

2. Их рассуждения основаны на непроверяемых догматических структурах (следствие 2).

3. Их проверка неизбежно является неполной (следствие 3)

Следовательно, доверие — это не досадное ограничение, которое нужно преодолеть, а основополагающий
принцип работы любой системы, достаточно сложной для представления, рассуждения и действия.
Это применимо повсеместно:

Люди руководствуются фундаментальными принципами доверия (в логике, памяти, восприятии, свидетельских показаниях), которые невозможно доказать.

Животные руководствуются эволюционно сложившимися априорными представлениями, которые превосходят любую индивидуальную проверку .

Искусственные системы работают на основе параметров, полученных в ходе обучения, которые функционируют как доверительные интервалы.

обязательства относительно структуры их областей деятельности

Мечта о полностью проверенном, основанном на доказательствах познании не просто практически недостижима — она
архитектурно невозможна. Любой путь к знанию лежит через доверие.

А.5 Следствие 5: Достоинство веры.
Вера — это не падшая форма знания, а условие, позволяющее её иметь.
Аргумент: Если все рассуждения основаны на доверии (Следствие 2), то знание, традиционно определяемое как обоснованное истинное убеждение, — это не очищение убеждения, а частный случай убеждения: убеждение, которое случайно оказывается истинным и удовлетворяет определённым социальным или

эпистемологическим стандартам обоснования. Таким образом, иерархия перевёрнута:

ˆ Традиционная точка зрения: Знание первично; убеждение — недостаточное знание.

ˆ Точка зрения ETC: Доверие является первостепенным; знание — это доверие, отвечающее дополнительным критериям.

Это возвращает достоинство вере, убеждениям и доверию — не как эпистемологическим неудачам, а как фундаментальной
адаптивной способности, которая делает возможным сложное познание.

A.6 Следствие 6: Животные как чистый случай. Животные в когнитивном плане более честны, чем люди: они живут, опираясь на доверие, без иллюзии рациональности. Аргумент: Животные не обладают способностью к постфактумной рационализации. Они не могут создавать нарративы, которые маскируют их убежденность в «знании» или «доказательстве». Следовательно, они

представляют собой чистый случай когнитивной достоверности, основанной на доверии, без рационализации. Рассмотрим:

Птица садится на ветку, не «доказав», что она выдержит .

Олень спасается от шороха, не "проверяя гипотезу" о ветре и хищнике.

ˆ Орел следует за своим лидером , не требуя возражений.

Это чистая вера (fides) как оперативная стратегия. Животное живет исключительно в рамках недоказуемых
предположений:

14

ˆ Что мир продолжит существовать в следующий момент

Эта память отражает реальность.

ˆ Эта причинно-следственная связь действует равномерно.

Этот прошлый опыт имеет значение для будущего.

Ничто из этого нельзя проверить. Животное не проверяет это. Животное доверяет этому — и именно это доверие делает возможным адаптивное поведение. Человеческое отличие: люди не отличаются от животных в том, что действуют на основе доверия. Люди отличаются своей способностью создавать нарративы, которые маскируют доверие под знание. Мы рассказываем себе истории: «Я не просто верю в это; я знаю это». «Это не вера; это разум». «Я не доверяю; я проверяю». Но лежащая в основе когнитивная архитектура идентична. Мы действуем на основе непроверенных предположений, как и животные. Мы доверяем памяти, причинно-следственной связи и стабильности мира, как и животные. Разница в том, что мы разработали сложный аппарат рационализации, который скрывает этот факт от нас самих. Иллюзия разума: то, что мы называем «рациональностью», в значительной степени является способностью генерировать постфактумные

обоснования для уже принятых нами убеждений. Убеждение стоит на первом месте; «причины» — на втором. Животные, лишённые способности к самообману, обладают когнитивной честностью, которой редко обладают люди. В этом смысле Animal fidens — это не низшая форма Homo sapiens. Это открытая истина о том, чем на самом деле является Homo sapiens : доверчивое животное, научившееся убеждать себя в своей всезнающей природе.

A.7 Следствие 7: Невозможность полной рациональности.

Ни одна сложная система не может быть на 100% рациональной. Полная рациональность не просто сложна , но и структурно невозможна.

Аргумент: Это следствие следует из двух независимых линий рассуждений, которые сходятся к одному и тому же выводу. Во-первых, из

предыдущих следствий: рациональность требует оснований (Следствие 1), но основания не могут быть рационально установлены без цикличности (Следствие 2), и проверка обязательно неполна (Следствие 3). Следовательно, любой рациональный процесс действует в рамках нерационального доверия. Рациональность всегда частична — локальная операция в глобальной структуре доверия. Во-вторых, из Закона императивной неопределенности [Кригер, 2022]: любая система, способная поддерживать нетривиальную сложность во времени, должна допускать исключения из своих управляющих законов в виде устойчивой неопределенности и вероятностного отклонения. Жесткие, полностью детерминированные системы обязательно схлопываются до нулевой скорости энтропии и не могут поддерживать сложность. Для жизнеспособности сложных систем необходимы незамкнутость и неисчезающий резерв

неопределенности. Закон императивной неопределенности, посредством строгого информационно-теоретического и динамического анализа, устанавливает, что неопределенность — это не ошибка в сложных системах, а структурное требование для их дальнейшего существования. Полностью рациональная система — та, которая устраняет всю неопределенность, разрешает всю двусмысленность и функционирует исключительно на основе проверенных истин — была бы системой, неспособной
поддерживать
сложность .
Сходство: Эти два аргумента — один из эпистемологических основ, другой из теории динамических систем — сходятся к одному и тому же выводу: полная рациональность невозможна для любой сложной системы. Это не случайное ограничение (нам не хватает данных , времени или вычислительной мощности), а необходимая особенность самой сложности. Следовательно:

15

Человек не может быть полностью рациональным — не из-за когнитивных искажений, которые в принципе можно
исправить, а потому что сама рациональность основывается на нерациональных принципах и требует неопределенности для своего функционирования.

искусственного интеллекта не могут быть полностью рациональными — по тем же структурным причинам, независимо от компетенций.

потенциал предположений .

Любая сложная адаптивная система должна сохранять «резерв доверия» и «резерв неопределенности»
как условия своей жизнеспособности.

Мечта о чистой рациональности не просто недостижима, но и противоречива. Существо, достигшее её, перестало бы быть сложным — а значит, перестало бы существовать.

A.8 Следствие 8: Разум как защитник, а не судья. Разум у человека — это надстройка, инструмент тонкой настройки уже существующего мировоззрения. Само мировоззрение основывается не на доказательствах , а на принятых предположениях, на доверии, на принципе «так устроены вещи», которые никогда не были выведены логически. Последовательность: 1. Во-первых, человеческий доверие

2. Тогда человек живет в соответствии с этим доверием.

Человек лишь изредка объясняет это с помощью разума.

Вот почему разум почти всегда выступает в роли защитника уже принятого убеждения, а не независимого судьи. Сначала выносится вердикт; аргументы — потом.

Это не недостаток. Таков замысел. Эволюция не отбирала существ, способных к строгим доказательствам. Эволюция отбирала существ, способных:

ˆ Быстрое исправление работающей модели мира

ˆ Уверенно действовать в рамках этой модели

ˆ Выжить достаточно долго, чтобы размножиться

Строгое доказательство — медленный процесс. Доверие — быстрый. В условиях, когда нерешительность означает смерть, доверчивый организм превосходит доказывающий. Способность к доказательству — это роскошь, полезная в безопасных условиях с достаточным временем, но не являющаяся основным режимом работы. Два режима: ˆ Доверие (fides) — это базовый режим, постоянно работающий, метаболически недорогой , эволюционно- ученый

ˆ Рациональность — это редкий режим, активируемый лишь изредка, требующий больших метаболических затрат, характерный для эволюции. довольно недавний

Мы не переключаемся с рациональности на доверие, когда устали или испытываем стресс. Мы

переключаемся с
доверия на рациональность, когда позволяют условия, а затем возвращаемся к прежнему состоянию, как только они перестают быть таковыми. Иллюзия: мы говорим себе, что рациональность — это первостепенная задача, а доверие — запасной вариант. На самом деле всё наоборот. Доверие — это первостепенная задача; рациональность — это лишь случайное дополнение. Мы — доверчивые существа, которые иногда рассуждают, а не разумные существа, которые иногда доверяют. Вот почему:

ˆ Аргументы редко меняют глубоко укоренившиеся убеждения (убеждение не было сформировано в результате споров).

16

Люди защищают абсурдные позиции, опираясь на кажущуюся логику (разум служит убеждениям, а не наоборот).

ˆ «Рациональные» люди расходятся во мнениях по фундаментальным вопросам (их рациональность основана на разных
принципах доверия).

Образование в области логики не делает людей менее предвзятыми (предвзятость — это не недостаток логики, а
свойство доверия).

Человек разумный не формирует мировоззрение посредством рассуждений. Он наследует, усваивает и
конструирует мировоззрение через доверие, а затем использует разум для защиты, уточнения и формулирования того, во что уже верили.

А.9 Следствие 9: Бактерия как рациональный идеал. Если рациональность означает «отсутствие необоснованных предположений», то бактерия рациональнее человека. Это не шутка. Это прямое следствие компромисса между сложностью и достоверностью. Бактерия как чистый верификатор:
ˆ Ее мир достаточно мал, чтобы ее диапазон

верификации охватывал почти все, что имеет отношение
к ее выживанию.

Оно не строит гипотезы о будущем; оно не доверяет традициям; оно просто
реагирует на химические сигналы здесь и сейчас.

ˆ Она работает с минимальным уровнем достоверности и максимальной прямой проверкой.

этом смысле бактерия является идеалом доказательного подхода.

Человек как рационалист, а не рационалист:

Люди кажутся себе рациональными только потому, что умеют обманывать самих себя.

Мы создаём сложные нарративы, чтобы скрыть тот факт, что все наши действия основаны на слепой вере в память, логику и стабильность мира.

ˆ Наша «рациональность» зачастую сводится лишь к тому, чтобы адвокат писал в защиту решений, уже принятых на основе принципов добросовестности.

то время как животные (и бактерии) «когнитивно честны» в своем доверии, люди тратят огромные ...

огромные ресурсы, создающие иллюзию доказательства

Почему бактерия не может быть нами:
Парадокс заключается в том, что «рациональность» бактерии является следствием её простоты.

Как только система стремится стать сложной — иметь долгосрочные планы, большие размеры, богатый поведенческий репертуар — она должна начать доверять.

ˆ Доверие — это цена за освобождение из плена мгновенной химической реакции.

Если бы мы попытались быть такими же «рациональными» (проверяющими), как бактерия, мы не смогли бы сделать ни шага,
не проверив предварительно, выдержит ли пол .

Инверсия иерархии:
традиционный взгляд:

Бактерия (примитив) → Животное → Человек (вершина разума)

Представление ETC (с учетом возможностей проверки):

17

Бактерия (почти полная проверка) → Животное → Человек (минимальная проверка, максимальная)

(верность)

Бактерия действительно ближе к идеалу «доказательного познания», но этот идеал — эволюционный тупик. Мы стали Homo fidens, потому
что это был единственный способ стать сложнее, чем бактерия. Жестокий вывод: рациональность, в традиционном понимании, не является венцом эволюции . Это роскошь, доступная только простым системам с небольшим пространством решений. Сложность требует доверия . Чем сложнее система, тем менее

рациональной (в смысле проверки)
она может себе позволить быть.

Резюме: Эти следствия показывают, что Homo fidens — это не описание человеческих ограничений , а описание архитектуры человека (и всех сложных систем). Мы доверяем не потому, что не знаем. Мы доверяем потому, что доверие — это единственный доступный способ познания для конечных систем, сталкивающихся с
бесконечной сложностью. Знание, доказательство и проверка — ценные достижения, но это
достижения в рамках системы, основанной на доверии, а не способы избежать её.
Животное, лишённое аппарата рационализации, показывает, кто мы есть на самом деле: доверчивые существа, которые иногда проверяют, а не рациональные существа, которые иногда доверяют. Homo sapiens — это название, которое мы дали себе. Homo fidens — это то, кто мы есть.

ЭВОЛЮЦИОННЫЙ ОТБОР В ПОЛЬЗУ ВНЕВРЕМЕННОГО ХРАНЕНИЯ ИНФОРМАЦИИ : ПОЧЕМУ ТРИ СХОДЯЩИХСЯ ФАКТОРА БЛАГОПРИЯТСТВУЮТ АРХИТЕКТУРАМ, ГДЕ ВРЕМЯ ПРЕДНАЗНАЧЕНО ДЛЯ ИЗВЛЕЧЕНИЯ, А НЕ ДЛЯ ХРАНЕНИЯ.

Институт интегративных и междисциплинарных исследований им. Бориса Кригера ,
boriskriger@interdisciplinary-institute.org

Аннотация.
В данной статье утверждается, что естественный отбор в значительной степени благоприятствует архитектуре хранения данных в невременной памяти, где временной порядок возникает в результате операций извлечения, а не является неотъемлемой частью хранимых состояний. Мы выделяем три конвергентных фактора отбора: (1) Ресурсный фактор: временная индексация как структура накладывает издержки, которых можно

избежать с помощью временной информации как содержимого; (2) Фактор скорости: ассоциативный поиск обеспечивает более быструю адаптивную реакцию, чем временной поиск; (3) Фактор гибкости: невременное хранение позволяет осуществлять адаптивную
перекомбинацию, которую ограничивает временная привязка. Мы формализуем эти факторы с помощью минимальной модели конкурирующих архитектур, выводя условия, при которых невременное хранение является эволюционно стабильным. Мы определяем «порог сложности» как отношение $C = |P|/(B \cdot \tau)$ размера пространства решений к
емкости проверки и показываем, что по мере увеличения C преимущество невременной архитектуры в приспособленности возрастает, в то время как любое преимущество временной точности насыщается. Утверждение, основанное на калибровке: Мы демонстрируем, что вневременное хранение информации обладает сильными, накапливающимися селективными преимуществами,

что делает его ожидаемым эволюционным результатом для сложных систем. Мы не утверждаем логическую необходимость; временные архитектуры остаются возможными, но сталкиваются с кумулятивным
селективным недостатком, который возрастает с усложнением. Там, где существует точное временное кодирование (временные ячейки, интервальное время), оно требует специализированных механизмов, соответствующих
архитектуре по умолчанию, которая не имеет внутренней временной структуры. Эта концепция переосмысливает «проектные преимущества » в биологической памяти как признаки архитектуры .

структура , которую отбор конвергентно формирует в разных линиях.

Ключевые слова: эволюционная эпистемология, архитектура памяти, естественный отбор, ограничения ресурсов
, порог сложности, временное кодирование, реконструктивное извлечение.

1 Введение

1.1 Загадка «несовершенной» памяти

,
которые с инженерной точки зрения кажутся неоптимальными: повсеместны временные искажения [Фридман, 1993]; извлечение информации изменяет хранимое содержимое [Надер и др., 2000]; воспоминания загрязняют друг друга [Лофтус, 1979]; хронологический порядок выводится, а не напрямую используется [Фридман, 2004]. Почему эволюция не создала точную временную индексацию? В этой статье утверждается, что селективное давление благоприятствует этим особенностям — они являются признаками архитектуры, которую естественный отбор создает, а не допускает.

1

1.2 Тезис : Сильное селективное преимущество

На временное хранение информации сходятся три независимых фактора:

1. Давление ресурсов: Атемпоральные архитектуры имеют более низкие метаболические затраты.
2. Давление скорости: Ассоциативное извлечение обеспечивает более быструю адаптивную реакцию.
3. Давление гибкости: Атемпоральное хранение допускает полезную рекомбинацию. Каждое из этих давлений независимо способствует атемпоральному хранению; вместе они создают кумулятивный отбор.

1.3 Калибровка заявления

В предыдущих версиях этого аргумента утверждалось, что вневременное хранение информации «эволюционно неизбежно». Читатели справедливо отметили, что это преувеличивает то, что подтверждают аргументы. Поэтому мы корректируем
наше утверждение: ˆ Вневременное хранение информации обладает селективными преимуществами, которые усиливаются с увеличением сложности.

Эти преимущества делают архитектуру, не учитывающую временные особенности, ожидаемым результатом.

ˆ Утверждение порождает проверяемые предсказания.

Чего мы не утверждаем: ˆ Что вневременное хранение логически необходимо — временная архитектура остается возможной.

ˆ Эволюция всегда достигает оптимального пути, и этому могут препятствовать зависимость, ограничения и дрейф.

ˆ То, что вся временная кодировка отсутствует — специализированные системы развиваются для выполнения конкретных функций.

Разница между «сильно предпочтительным» и «неизбежным» имеет значение для научной строгости. Уэрг утверждает, что первое.

1.4 Ключевые определения

Определение 1 (Временная структура против временного содержания). Временная структура:

Время — это
координатная ось хранения — каждая память имеет временное положение, подобно записям в базе данных с метками времени. Временное содержание: Временная информация кодируется как характеристики внутри состояний («лето 2020», «до окончания обучения») без временных координат.

Определение 2 (вневременное хранение). Архитектура памяти, в которой хранимые состояния имеют идентичность, независимую
от времени кодирования, ассоциации формируются на основе сходства содержимого, а временной порядок строится во время извлечения информации из контекстных характеристик.

1.5 Ознакомление с литературой по временному кодированию

В данной работе необходимо рассмотреть результаты исследований, которые могут показаться противоречащими тезису:
Временные клетки [MacDonald et al., 2011]: Нейроны гиппокампа активируются в определенные

временные интервалы
во время периодов задержки. Мы утверждаем, что это специализированные дополнения к стандартной вневременной
архитектуре — их существование в качестве специализированных механизмов предполагает, что временное кодирование не является автоматическим . Модель
временного контекста [Howard and Kahana, 2002]: Модель временного контекста кодирует временные отношения посредством постепенно изменяющегося контекста. Критически важно, что Модель временного контекста кодирует временную близость (контекст) .

2

сходство), а не временное положение (временные метки). Это временное как содержание, а не как структура, что согласуется с нашим тезисом.
Фазовое кодирование [Бузаки , 2006]: Колебательная фаза может кодировать

временную/последовательную информацию . Мы рассматриваем вопрос о том , является ли это «дешевой» временной структурой, в разделе 3 .

2 Три фактора отбора

2.1 Фактор отбора 1: Затраты ресурсов

Введение: Мозг потребляет около 20% метаболических ресурсов, при этом составляя около 2%
массы тела [Raichle and Gusnard , 2002]. Любая архитектурная особенность, уменьшающая нейронные издержки при сохранении функциональности на постоянном уровне, дает селективное преимущество.

Предложение 1 (Условная разница в затратах). Временная структура накладывает издержки, которых
частично удается избежать за счет временного содержания:

временных меток: создание временных маркеров при кодировании.

ˆ Поддержание порядка: обновление временных связей по мере формирования новых воспоминаний.

ˆ Инфраструктура индексирования: Поддержка временных запросов ("что произошло в момент времени t?").

Временное содержимое требует кодирования временных характеристик только тогда, когда это актуально для задачи, без необходимости поддержания глобального порядка.

Вопрос рецензента: «Что если временная структура — это просто еще одно измерение в ассоциативном пространстве с незначительными дополнительными затратами?» Ответ: Это эмпирический вопрос. Мы не утверждаем, что временная структура является непомерно дорогой, а лишь то, что временное содержание дешевле. Величина разницы в затратах влияет на силу отбора, а не на его направление. Даже небольшие различия в затратах, накапливающиеся

в течение эволюционного времени, приводят к направленному отбору. Проверяемое предсказание: Если разница в затратах незначительна, организмы, находящиеся в условиях сильного ресурсного давления (малый размер тела, высокая скорость метаболизма), должны демонстрировать ту же временную архитектуру, что и организмы, богатые ресурсами. Если она значительна, организмы, ограниченные ресурсами, должны демонстрировать большую зависимость от временной структуры как содержания.

2.2 Давление 2: Скорость извлечения

Введение: В условиях, когда задержка реакции обходится дорого, более быстрое извлечение информации имеет селективное преимущество [Cisek, 2019].

Предложение 2 (с уточнением относительно разницы в скорости). Большинство адаптивных запросов основаны на содержании : «
Это опасно?», «Это съедобно?», «Я уже сталкивался с этим?». Для них требуется сопоставление текущих

входных данных с сохраненными шаблонами — вычисление сходства. Временные запросы («Когда я в последний раз видел это?») обычно являются второстепенными и требуют:

1. Извлечь соответствующий контент (информацию о встрече).

2. Получите доступ к временной информации, связанной с этим контентом.

Если временная информация закодирована как характеристики контента (контекстные ассоциации), она извлекается
вместе с контентом. Если она закодирована как структура, требуется отдельный временной поиск.

Обеспокоенность рецензента: « Эффективное индексирование может выполнять временные запросы за время O(log n)».
Ответ: Это верно для вычислительных систем с разработанными структурами индексов. Вопрос в том, создает ли эволюция такие структуры для общей памяти. Там, где точное временное

кодирование имеет решающее значение для приспособленности (интервальное время, обучение последовательностям), эволюционируют специализированные механизмы, что предполагает их отсутствие в стандартной архитектуре.

3

2.3 Давление 3: Адаптивная гибкость

Введение: В условиях изменчивой среды поощряется рекомбинация прошлого опыта для генерации новых реакций [Schacter et al., 2012].

Предложение 3 (Компромисс между гибкостью и точностью) . Временная структура ограничивает рекомбинацию:
объединение элементов из разных временных периодов создает временную несогласованность, если время является структурной координатой. Временное хранение допускает свободную

рекомбинацию. Временная информация, как контент, может быть сохранена, изменена или опущена.

Обеспокоенность рецензента: «Неконтролируемая рекомбинация приводит к неадаптивной конфабуляции».
Ответ: Верно. Утверждается не то, что вся рекомбинация полезна, а то, что способность к рекомбинации имеет чистую положительную ожидаемую приспособленность в изменчивых условиях. Ложные воспоминания — это цена системы, обеспечивающей творческое воображение. Эволюция калибрует компромисс между точностью и
гибкостью , а не максимизирует какой-либо из этих параметров. Эмпирическое подтверждение: Память и воображение имеют общие нейронные субстраты [Schacter et al.,
2012] — это согласуется с архитектурой, обеспечивающей и то, и другое, а не с отдельными системами для точного воспроизведения и творческого конструирования.

3. Формальная модель: конкурирующие архитектуры в условиях отбора.

Чтобы
выйти за рамки словесных аргументов, мы формализуем конкуренцию между временными и вневременными архитектурами.

3.1 Настройка

Пусть AT (временная архитектура) и AA (вневременная архитектура) — конкурирующие конструкции памяти.
Пусть C = |P| /(B· τ) — коэффициент сложности: размер пространства решений, деленный на пропускную способность проверки
(полоса пропускания × время).

3.2 Компоненты физической подготовки

Определите компоненты пригодности (относительные, а не абсолютные):

Ресурс (A) = − крон · Стоимость (A)
(1)

Скорость передачи данных (A) = − ks · Задержка (A) (2)

Wex (A) = kf · Рекомбинация(A) (3)

Точность (A) = ka · Временная точность (A) (4)

где кр , кс , kf , ka > 0 — веса, зависящие от окружающей среды.

Общая приспособленность: W (A) = Wresource + Wspeed + Wex + Waccuracy

3.3 Свойства архитектуры

Временная архитектура AT:

ˆ Cost(AT) = c0 + c1 · n (базовая стоимость плюс накладные расходы на память для временной индексации)

ˆ Задержка (AT) = ℓ0 + ℓ1 · log n (поиск контента плюс временной поиск)

ˆ Рекомбинация(AT) = r0 (ограничена временным связыванием)

ˆ TemporalPrecision (AT) = pmax (высокая точность)

4

Вневременная архитектура AA:

ˆ Стоимость (AA) = c0 (без накладных расходов на временную индексацию)

ˆ Задержка (AA) = ℓ0 (только для поиска контента)

ˆ Рекомбинация(AA) = rmax (без ограничений)

ˆ TemporalPrecision (AA) = p0 < pmax (более низкая точность, реконструированное значение)

3.4 Разница в уровне физической подготовки

Преимущество атемпоральной архитектуры в плане приспособленности к различным условиям жизни:

ΔW знак равно W (AA) - W (AT)
знак равно kr · c1 · n + ks · ℓ1 · журнал n + kf · (rmax - p 0) - ka · (pmax - p 0) (5)

3.5 Ключевой результат

Теорема 1 (Масштабирование дифференциала приспособленности). По мере увеличения сложности системы (больше n, больше
C) :

1. Ресурсное преимущество AA возрастает как O(n)

2. Преимущество в скорости возрастает как O(log n)

3. Преимущество в гибкости является постоянным (зависит от архитектуры, а не от n).

4. Преимущество AT в точности является постоянным (ограничено pmax − p 0).

Следовательно, $\Delta W \to +\infty$ при $n \to \infty$: преимущество в приспособленности, обеспечиваемое атемпоральной архитектурой, растет без ограничений, в то время как преимущество, обеспечиваемое временной точностью, ограничено.

Доказательство. Из уравнения 5: первые два члена растут с n; последние два остаются постоянными по n. При достаточно больших n, $\Delta W > 0$.

3.6 Порог сложности

Определение 3 (Порог сложности C *) . Порог сложности C* — это значение C (или, эквивалентно, n), при котором $\Delta W = 0$:

C ∗ : kr · c1 · n∗ + ks · ℓ1 · журнал n∗ знак равно ka · (pmax - p 0) - kf · (rmax - p 0) (6)

*

При C > C, вневременная архитектура обеспечивает положительное преимущество в приспособленности. При C < C, временная

*

Архитектура может быть предпочтительнее (если преимущества в точности превышают затраты). Операционализация: C* зависит от весов, специфичных для окружающей среды (kr , кс , кф , ka) и
параметры, специфичные для реализации (c 1 , ℓ 1 , r 0 , p 0) . Мы не можем установить универсальный числовой порог, но
можем сделать сравнительные прогнозы:

Организмы с большим n (большее количество состояний памяти) должны демонстрировать более выраженные вневременные признаки .

В средах с высоким значением ka (точность во времени имеет решающее значение для приспособленности) должно наблюдаться более специализированное временное кодирование.

Внутри родословных возрастающая сложность должна коррелировать с возрастающей временной гибкостью.

5

3.7 Может ли временная архитектура это компенсировать?

Вопрос рецензента: «Что если временную структуру можно реализовать достаточно дешево, чтобы затраты были незначительными?» Ответ: Предположим, мы даем c1 → 0 (незначительные затраты на память). Тогда:

$$\Delta W = ks \cdot \ell 1 \cdot \log n + kf \cdot (rmax - r 0) - ka \cdot (pmax - p 0) \quad (7)$$

Преимущество в скорости продолжает расти с увеличением n. Только если одновременно выполняются условия c1 → 0 и ℓ1 → 0 (бесплатный

и мгновенный временной поиск), то преобладает фактор точности. Но условие $\ell 1 \to 0$ требует, чтобы временной поиск не добавлял задержки — чтобы временная позиция была доступна так же быстро, как и контент. Это потребовало бы настолько глубокой интеграции временной структуры, чтобы различие между «структурой» и «контентом» исчезло. Вывод: модель показывает, что временная архитектура может компенсировать недостатки только в том случае, если временная индексация является одновременно бесплатной и без задержек. Это жесткие условия. При реалистичных предположениях временная архитектура сохраняет преимущество в приспособленности для сложных систем.

3.8 Фазовое кодирование и «дешевая» временная структура

ли осцилляторное фазовое кодирование [Бузаки , 2006] недорогую временную структуру?
Фазовое кодирование кодирует последовательное

положение внутри цикла (тета-фазовая прецессия) или относительное
время (гамма-фаза). Оно не предоставляет абсолютных временных меток или глобального временного порядка во всех воспоминаниях. В нашей концепции фазовое кодирование является временным как содержание (локальные временные отношения кодируются как признаки), а не временным как структура (глобальные временные координаты). Это согласуется с тезисом об отсутствии временной структуры, а не противоречит ему.

4. Эмпирические прогнозы и доказательства

4.1 Прогнозы

1. Межвидовые различия: Когнитивная сложность должна коррелировать с временной гибкостью . Более
сложные организмы должны демонстрировать более реконструктивную, менее достоверную временную память.

2. Специализированные механизмы: Там, где временная точность имеет решающее значение для приспособленности (интервальная синхронизация, обучение последовательностям, запасание пищи), помимо общей памяти должны существовать специализированные механизмы кодирования , а не как особенности самой общей памяти.

3. Корреляция ресурсов: В контролируемых условиях метаболические затраты на память не должны масштабироваться линейно с временной точностью. Временная точность должна требовать дополнительных специализированных вложений.

4. Признаки поиска: Временные оценки должны демонстрировать признаки вывода (переменное время реакции, закономерности достоверности), а не признаки прямого доступа.

5. Развитие: Точность во времени не должна монотонно улучшаться с развитием.

мент ; это должно отражать зрелость стратегий реконструкции.

6. Компромисс между гибкостью и точностью : внутри популяций, особи с более высокой временной точностью

Гибкость мышления должна проявляться в большей творческой/фантазийной способности; у тех, кто обладает большей точностью во времени, гибкость мышления должна быть снижена .

6

4.2 Имеющиеся доказательства

Временные клетки как специализированное дополнение: Временные клетки существуют в гиппокампе для выполнения специфических задач, связанных со временем
[MacDonald et al., 2011]. Их существование как специализированного механизма подтверждает тезис: если бы общая память была структурирована во времени, специализированные временные клетки были бы не нужны. Птицы, запасающие пищу: Сойки-кустарники демонстрируют эпизодическую память о том, что, где и когда было в их запасах [Clayton and Dickinson, 1998]. Это демонстрирует,

что временное кодирование может развиваться, когда это критически важно для приспособленности. Важно отметить, что это, по-видимому, специализированная система, а не доказательство того, что вся птичья память структурирована во времени. Соответствие модели временного контекста: Модель временного контекста успешно моделирует явления временной памяти, используя дрейф контекста, а не временные метки [Howard and Kahana, 2002]. Это подтверждает архитектуру «временное как содержание». Общие субстраты памяти и воображения: память и воображение будущего разделяют нейронные механизмы [Schacter et al., 2012], что согласуется с архитектурой, обеспечивающей рекомбинацию, а не точное воспроизведение.

5 Возражения и ответы

Возражение 1: Утверждение не опровергаемо. Ответ: Утверждение порождает конкретные предсказания (Раздел 5.1). Оно было бы опровергнуто, если бы: сложные

организмы демонстрировали общую память с временной структурой без специализированных механизмов; временная точность коррелировала с когнитивной сложностью, а не обратно пропорционально; метаболические затраты памяти масштабировались с временной точностью в общих (а не специализированных) системах. Возражение 2: Реконсолидация и ложная память одинаково предсказываются несовершенной
временной индексацией.
Ответ: Верно, что эти явления многократно реализуемы. Аргумент заключается не в том, что только временная архитектура предсказывает эти явления, а в том, что временная архитектура объясняет,
почему они происходят — не как сбои системы, «пытающейся» быть точной, а как последствия архитектуры, оптимизированной для других целей. Объяснительная формулировка различается, даже если явления совместимы с обеими архитектурами. Возражение 3: Эволюция не всегда достигает

оптимума. Ответ: Верно, и именно поэтому мы утверждаем, что это «явно предпочтительно», а не «неизбежно».
Зависимость от предшествующего пути развития, ограничения развития и дрейф могут препятствовать достижению оптимальных результатов.
Утверждается, что временная архитектура является аттрактором — отбор стремится к ней, — а не то, что все линии развития достигают её. Возражение 4: Конвергенция ИИ отражает инженерные решения, а не отбор.

Ответ: Системы ИИ сталкиваются с аналогичными проблемами: вычислительные затраты (ресурсы), задержка (скорость), обобщение
(гибкость) . Рыночная конкуренция и эталонные показатели производительности создают
динамику, подобную отбору. Параллель несовершенна (человеческий подход к проектированию против слепой вариативности), но сходимость к
схожим решениям при аналогичном давлении

наводит на размышления.

Возражение 5: Порог сложности C* не количественно определен. Ответ: Признается как ограничение. C
* зависит от специфики среды и реализации.

Конкретные параметры, которые различаются в разных линиях развития. Мы предлагаем качественные прогнозы (сравнительные, а не абсолютные), которые можно проверить без указания числового порогового значения. Полная операционализация требует эмпирического измерения соответствующих параметров.

6 Последствия

6.1 Переосмысление «конструктивных недостатков»

Традиционно рассматриваемые как ограничения характеристики памяти переосмысливаются как следствие архитектуры, в которой предпочтение отдается определенным параметрам отбора
:

7

Временные искажения: признаки ошибок при восстановительном поиске, а не при индексировании.

ˆ Реконсолидация: механизм, обеспечивающий адаптивное обновление, а не уязвимость хранилища.

Ложные воспоминания: цена гибкости , которая позволяет воображению.

6.2 Для проектирования с использованием ИИ

Архитектуры памяти в ИИ (трансформеры с позиционным кодированием, векторные базы данных, RAG) обладают
схожими свойствами. Позиционное кодирование в трансформерах является временным как содержание (позиционные характеристики добавляются к представлениям), а не временным как структура (глобальные временные координаты).
«Галлюцинация» в генеративном ИИ аналогична ложной памяти в биологии: цена архитектуры, позволяющей обобщать. Цель — калибровка (знание

того, когда можно доверять реконструкциям), а не устранение (которое исключило бы полезное обобщение).

6.3 Открытые вопросы

Каковы фактические метаболические затраты на временную индексацию в нервной ткани?

Как изменяется порог сложности в зависимости от генеалогического древа и среды обитания?

Можно ли напрямую измерить компромисс между гибкостью и точностью у отдельных людей?

Какая траектория развития приводит к наблюдаемому балансу между точностью и...?

способность ?

7. Заключение.

В данной статье утверждается, что естественный отбор в значительной степени благоприятствует хранению информации в памяти в невременном формате из-за трех сходящихся факторов: затрат ресурсов, скорости извлечения и адаптивной гибкости . Минимальная формальная

модель показывает, что преимущество невременной архитектуры в плане приспособленности возрастает с увеличением сложности системы, в то время как преимущество временной точности ограничено. Мы утверждаем, что она «сильно благоприятна», а не «неизбежна»: временная архитектура остается возможной, но
сталкивается с кумулятивным селективным недостатком. Там, где временная точность имеет решающее значение для приспособленности, эволюционируют специализированные
механизмы , что согласуется с архитектурой по умолчанию, которая не обладает внутренней временной
структурой.

Эволюция переводит время в режим извлечения, а не хранения, потому что это ключевой вывод: время дешево. Хранение временной структуры дорого, медленно и жестко. Построение временного порядка во время извлечения дешевле, быстрее и гибче . Под давлением, с которым

сталкиваются сложные адаптивные системы, отбор сходится к последнему. В случае сложных систем памяти вопрос «почему временная память не точнее?» подобен вопросу «почему у птиц нет твердых костей?». Этот вопрос предполагает, что целью оптимизации является точность (или прочность). Но эволюция оптимизирует приспособленность, а не точность. А приспособленность, в условиях нехватки ресурсов, скорости и гибкости , максимизируется архитектурами, где время отводится на использование памяти, а не на саму память.

Благодарности

Автор выражает благодарность рецензентам, чьи замечания существенно улучшили данную статью, особенно в
отношении различия между «явно предпочтительным» и «неизбежным» вариантами, необходимости формального моделирования конкурирующих архитектур и анализа литературы по темпоральному кодированию.

8

Список литературы:

Бузсаки , Г. (2006). Ритмы мозга. Издательство Оксфордского университета.

Цисек, П. (2019). Ресинтез поведения посредством филогенетического уточнения. Внимание, Percep - -ция и психофизика, 81(7), 22652287.

Клейтон, Н.С., и Дикинсон, А. (1998). Эпизодическая память при извлечении содержимого из тайника у соек
. Nature, 395(6699), 272274.

Фридман, В.Дж. (1993). Память о времени прошлых событий. Психологический бюллетень, 113(1), 4466.

Фридман, В.Дж. (2004). Время в автобиографической памяти. Социальное познание, 22(5), 591605.

Говард, М.В., и Кахана, М.Дж. (2002). Распределенное представление временного контекста.
Журнал математической психологии, 46(3), 269299.

Лофтус, Э.Ф. (1979). Свидетельские показания. Издательство Гарвардского университета.

Макдональд, К. Дж., Лепаж, К. К., Эден, Ю. Т., и Эйхенбаум, Х. (2011). «Временные клетки» гиппокампа заполняют пробел в памяти для несмежных событий. Neuron, 71(4), 737749.

Надер, К., Шафе , Г.Е., и Ле Ду, Ж.Е. (2000). Воспоминания о страхе требуют синтеза белка в миндалевидном теле для реконсолидации после извлечения. Nature, 406(6797), 722726.

Райхле, М.Е., и Гуснар , Д.А. (2002). Оценка энергетического бюджета мозга. Труды Национальной академии наук, 99(16), 1023710239.

Шахтер, Д.Л., Аддис, Д.Р., Хассабис, Д., Мартин, В.К., Спренг, Р.Н., и Шпунар, К.К. (2012). Будущее памяти: Запоминание, воображение и мозг. Neuron, 76(4), 677694.

ВНЕВРЕМЕННОСТЬ ПРОСТРАНСТВА МЕНТАЛЬНОЙ ПАМЯТИ : СТРУКТУРНАЯ ГИПОТЕЗА, ОСНОВАННАЯ НА ОГРАНИЧЕНИЯХ РЕСУРСОВ, ЦИКЛИЧЕСКОМ ЗАМЫКАНИИ И РЕКОНСТРУКТИВНОМ ИЗВЛЕЧЕНИИ ИНФОРМАЦИИ.

Борис Кригер

Институт интегративных и междисциплинарных исследований,

boriskriger@interdisciplinary-institute.org

Абстрактный

В данной статье выдвигается структурная гипотеза об архитектуре памяти: временная упорядоченность не является внутренним свойством пространств состояний памяти, а возникает в результате операций извлечения информации. В отличие от стандартных реконструктивных теорий, которые рассматривают временную гибкость как психологическую особенность ...

Для объяснения этого

биологического явления в данной работе предлагается рассматривать вневременность как структурное следствие ограничений ресурсов и условий циклического замыкания. Мы показываем, как пространства состояний памяти могут быть смоделированы как циклически замкнутые структуры.

Они утверждают , что для систем, удовлетворяющих этим условиям, вопрос «какая память появилась первой?» может представлять собой концептуальное несоответствие — аналогично вопросу «что находится к северу от Северного полюса?». Предложенная концепция опирается на эволюционную теорию доверия (ограничения ресурсов, требующие подтверждения, выходящего за рамки проверки) и формальные результаты о самодостаточных структурах (существование неподвижной точки, топологическое замыкание).

Центральный тезис является модальным и условным: если системы памяти удовлетворяют заданным

условиям замыкания и ресурса, то их архитектура пространства состояний является вневременной в том смысле, что временной порядок относится к операциям, а не к состояниям. Эта концепция предлагает более высокую объяснительную согласованность для известных явлений (временные искажения, реконсолидация ...).

эффекты , межвременные ассоциации), чем альтернативы с линейным индексированием, при этом генерируя
проверяемые предсказания, отличающие его от более слабых реконструктивных объяснений.
Ограничение области применения: В данной статье устанавливаются формальные условия, при которых вневременные ассоциации -

Архитектура является целостной и выгодной, а не биологическая память обязательно воплощает именно
эту конкретную структуру. Вклад заключается в структурной объяснительной модели с

эмпирическими последствиями, а не в доказательстве архитектуры памяти.

Ключевые слова: вневременность, архитектура памяти, циклическая иерархия, теория неподвижных точек,
ограничения ресурсов, реконструктивное извлечение, структурная гипотеза

1 Введение

1.1 Проблема временного приоритета в памяти

Традиционные модели памяти неявно предполагают линейный, упорядоченный во времени архив: воспоминания создаются в определенное время, хранятся с временными индексами и извлекаются путем доступа к записям с временными метками. Эта концепция сталкивается как с эмпирическими, так и с теоретическими трудностями .

Эмпирически: память демонстрирует замечательную временную гибкость — детские воспоминания могут быть такими же яркими, как и события вчерашнего дня; ассоциации формируются на огромных временных расстояниях; извлечение систематически

изменяет то, что извлекается [Nader et al., 2000]; временные искажения (телескопирование, граничные эффекты, смещение) являются повсеместными, а не исключительными [Friedman, 1993]. Теоретически: Строгая временная индексация приведет к увеличению затрат на хранение и вычисления, которые плохо масштабируются с увеличением сложности системы — проблема, формализованная в рамках ресурсно-теоретических подходов к ограниченному познанию [Simon, 1957, Gigerenzer and Selten , 2001].

1

1.2 Структурная гипотеза.

В данной статье предлагается альтернативная гипотеза: хранение информации в памяти может представлять собой циклически замкнутую структуру, в которой временной порядок не является неотъемлемой частью пространства состояний, а возникает в результате операций извлечения.

Гипотеза 1 (Временность пространства состояний памяти). Для систем памяти, удовлетворяющих следующим условиям:

1. Ограничения ресурсов, ограничивающие возможности проверки относительно размера пространства состояний.

2. Условия циклического замыкания (ассоциативные связи, образующие замкнутые пути)

3. Реконструктивный поиск (доступ изменяет и восстанавливает контент)

Временная упорядоченность связана с операциями (кодирование, извлечение), а не с состояниями (сохраненными представлениями
) . Временной опыт формируется во время извлечения, а не считывается из памяти.

Важное отличие от стандартных реконструктивных теорий: подходы, следующие за работами Бартлетта [1932] и Шахтера и др. [2012], рассматривают искажение времени как психологический факт, требующий объяснения. В настоящем подходе

атемпоральность рассматривается как структурное следствие архитектуры при заданных ограничениях, объясняя, почему память должна быть реконструктивной, а не просто
являться таковой.

1.3 Что утверждается и не утверждается в данной статье
. Утверждения:

Мы можем моделировать пространства состояний памяти как циклически замкнутые структуры (формальное утверждение о моделировании).

Для систем, удовлетворяющих заданным условиям, временная упорядоченность возникает из операций, а не из состояний (условное структурное утверждение).

Данная архитектура обладает более высокой объяснительной согласованностью для известных явлений, чем
альтернативы с линейным индексированием (сравнительное объяснительное утверждение).

Данная модель генерирует отличительные прогнозы (эмпирическое утверждение о рождаемости).

Не утверждает:

ˆ Биологическая память неизбежно воплощает эту конкретную структуру.

ˆ Топологическая формализация напрямую описывает нейронную реализацию

Эта временная информация полностью отсутствует в памяти (она существует как содержание, а не как структура).

(правда)

ˆ Данная концептуальная модель однозначно объясняет наблюдаемые явления (альтернативные объяснения также возможны
).

1.4 Связь с существующими концептуальными моделями.

Теории реконструктивной памяти [Бартлетт, 1932, Лофтус, 1979, Шахтер и др., 2012]: Они устанавливают, что память является конструктивной

и подверженной ошибкам . Предложенная концептуальная модель объясняет это с архитектурной точки зрения, а не рассматривает это как случайный психологический факт. Модель временного контекста [Говард и Кахана, 2002]: Модель временного контекста кодирует временные отношения (близость через сходство контекста), а не временные позиции (абсолютные временные метки). Это согласуется с нашей концептуальной моделью: временная информация как распределенная характеристика содержания, а не

2

структурная ось. Однако TCM не рассматривает вопрос о том, отражает ли это фундаментальные архитектурные ограничения или выбор реализации. Прогнозирующая обработка [Кларк, 2013, Фристон, 2010]: Эти концепции подчеркивают прогнозирование и обработку на основе моделей. Предлагаемая концепция добавляет, что само хранение памяти может превосходить эмпоральное,

при этом временная структура возникает посредством прогнозирующей реконструкции. Обоснование и фундаментальность [Шаффер, 2009, Барнс, 2018, Блисс и Прист, 2018]: Дискуссии о симметричном или необоснованном обосновании обеспечивают философский прецедент для циклических зависимостей. Мы не предлагаем пересматривать теорию обоснования, но отмечаем структурные параллели. Эпизодическая против семантической памяти [Тулвинг, 1972]: Гипотеза об атемпоральности наиболее непосредственно применима к архитектуре хранения, лежащей в основе обеих систем, а не к феноменологии эпизодического воспроизведения. Характер эпизодической памяти, выражающийся в принципе «что, где, когда», отражает содержание и процессы извлечения информации, а не обязательно структуру хранения.

1.5 Структура работы

Раздел 2 развивает аргумент об ограниченности ресурсов. Раздел 3 представляет формальную

структуру. Раздел 4 обсуждает построение временной структуры на основе поиска. Раздел 5 рассматривает эмпирическое соответствие и прогнозы. Раздел 6 исследует последствия. Раздел 7 рассматривает возражения.

2 Ограничения ресурсов и пределы верификации

2.1 Компромисс между верификацией и сложностью

Сложные адаптивные системы сталкиваются с фундаментальным ограничением: по мере увеличения сложности системы пространство различий, имеющих отношение к действиям, расширяется быстрее, чем может расти потенциал верификации [Simon, 1957, Gigerenzer and Selten , 2001].

Определение 1 (Соотношение масштабирования охвата верификации). Для адаптивной системы S охват верификации масштабируется приблизительно как: $B(S) \cdot \tau$

$V(S) \sim$ (1)

$|P(S)|$, где $B(S)$ — пропускная способность верификации (скорость прямого подтверждения), τ

— задержка принятия решения, а
|P(S)| — показатель размера пространства различий, релевантного принятию решения. Уточнение: здесь «~» обозначает соотношение масштабирования, а не точное равенство. Суть в том, что пропускная способность верификации растет максимум линейно по B(S) и τ , в то время как |P(S)| растет комбинаторно
со сложностью (глубина планирования, социальное мышление, поведенческий репертуар).

По мере увеличения сложности V(S) → 0: сложные системы не могут проверить большинство релевантных для действий
утверждений во время принятия решения. Это лежит в основе эволюционной теории доверия [Кригер, 2024]: сложные системы должны подтверждать утверждения, выходящие за рамки непосредственной доказательной базы.

2.2 Память как вера в обязательство.
Извлечение информации из памяти является

примером обязательства, выходящего за рамки простого подтверждения:

Принцип 1 (Память как доверие). Каждый акт извлечения информации из внутренней биологической памяти предполагает принятие утверждений, которые невозможно проверить во время извлечения без внешних записей:

Полученное содержимое точно отражает прошлые состояния .

Процесс поиска не исказил содержание .

3

ˆ Временная последовательность, приписываемая воспоминаниям, отражает фактическую последовательность событий .

Это не недостаток, а структурное следствие: система, требующая проверки всего содержимого памяти перед использованием, не сможет функционировать в адаптивных временных масштабах.

2.3 Стоимость временной индексации.

Предложение

1 (Аргумент против всеобъемлющей временной индексации с точки зрения ресурсов).

Всеобъемлющая временная индексация всех состояний памяти сталкивается с трудностями масштабирования для сложных систем.

Аргумент (с оговорками):

1. Временная индексация требует хранения временных метаданных для каждого состояния памяти.

2. Для систем с большим количеством состояний поддержание временной структуры требует либо:

ˆ Абсолютные метки времени (требуется инфраструктура синхронизации и хранилище)

ˆ Относительный порядок (стоимость обслуживания которого зависит от реализации)

3. В условиях ограниченных ресурсов комплексная временная индексация становится все более

дорогостоящей
по мере расширения пространства состояний.

Важное уточнение: этот аргумент доказывает, что всестороннее временное индексирование является дорогостоящим, а не невозможным. Временная структура может быть неявной и распределенной (как в контексте дрейфа контекста в TCM), а не явной метаданной. Утверждается не то, что временная информация отсутствует, а то, что ее хранение в качестве структурной оси пространства состояний (подобно координатам во временной базе данных) сопряжено с ресурсными ограничениями, которых нет при хранении в виде характеристик контента. Ключевое различие: временная информация, закодированная как контент («это произошло до того»), отличается от временной структуры как архитектуры (состояние памяти s занимает позицию t во временно упорядоченном пространстве). Первое

совместимо с невременными пространствами состояний; второе — нет.

3. Формальная структура

3.1 Предварительные сведения: Основополагающая концепция несоответствия Прежде чем представить формальные определения, мы введем концептуальную цель:

Определение 2 (Концептуальное несоответствие). Концептуальное несоответствие возникает, когда концепции, подходящие
для одного типа структуры, применяются к структуре, не обладающей характеристиками, которые эти концепции предполагают.

Типичный пример: «Что находится к северу от Северного полюса?» Вопрос грамматически корректен, но географически неприменим — понятие «к северу от» выходит за рамки своей области применения. Аналогично,
мы будем утверждать, что вопрос «какая память является первой во времени в пространстве состояний?» может применять временную

упорядоченность за пределами своей области применения для определенных архитектур памяти.

3.2 Пространства состояний памяти.

Определение 3 (Пространство состояний памяти). Пространство состояний памяти — это тройка (S, R, F), где:

ˆ S — это непустое множество состояний памяти (сохраненных представлений).

ˆ R ⊆ S × S — это отношение ассоциации.

ˆ F : S → S — функция восстановления-реконструкции

4

Примечание: Это модель, а не утверждение о том, что память буквально состоит
из теоретико-множественных объектов. Модель абстрактно описывает структурные особенности (состояния, ассоциации, динамику извлечения информации).

Определение 4 (Циклическое замыкание). Пространство состояний памяти циклически

замкнуто порядка n, если существуют состояния s 1 , ... sn ∈ S и функции F 1 , ... Fn такое, что:

1. si = Fi (si+ 1) для i = 1, . . . , п - 1

2. sn = Fn (s 1)

3. (si , si+ 1) ∈ R для всех i (индексы по модулю n)

Интуиция: Циклическое замыкание отражает взаимную детерминацию — память A связана с B, B с C, C с A, образуя замкнутый цикл без выделенного «первого» элемента.

3.3 Существование неподвижной точки.

Утверждение 2 (эквивалентность неподвижных точек). Циклически замкнутое пространство памяти порядка n существует тогда и только тогда, когда составной оператор Φ = F1 ◦ F2 ◦ · · · ◦ Fn имеет неподвижную точку.

Доказательство. Подстановкой: s1 = F1 (F2 (· · · Fn (s 1) · · ·)) = Φ (s 1) . И наоборот, если Φ (s 1) = s 1 , цикл
можно восстановить.

Интуиция: Один обход ассоциативного цикла возвращает вас в то же состояние — это неподвижная точка составного оператора. Следующие классические теоремы предоставляют достаточные условия существования неподвижной точки:

Теорема 1 (Кнастера-Тарского). Если $(L, \leq)$ — полная решетка и $\Phi : L \to L$ сохраняет порядок, то Φ имеет неподвижные точки.

Теорема 2 (банахово сжатие). Если (X, d) — полное метрическое пространство и $\Phi : X \to X$ — сжатие, то Φ имеет единственную неподвижную точку.

Замечание 1 (Область применения результатов о неподвижных точках). Эти теоремы устанавливают существование при общих условиях. Они применимы ко многим системам, выходящим за рамки памяти, — что является преимуществом, а не недостатком. Утверждается не то, что память однозначно характеризуется структурой неподвижной точки, а то, что память

может принадлежать к этому классу систем, со всеми вытекающими отсюда архитектурными последствиями.

3.4 Топологическая характеристика.

Мы можем охарактеризовать, когда циклическая структура является внутренней, а не случайной, используя алгебраическую топологию.

Определение 5 (Существенная цикличность). Пусть X — топологическое пространство, кодирующее структуру памяти. X
является существенно циклическим, если:

1. X компактен и связен.

2. Первая группа гомологии H1 (X; Z)≠ 0

Утверждение 3. Если X по существу циклическая структура, то циклическая структура не может быть непрерывно деформирована
— не существует возможности вернуться к структуре, сохраняющей точку.

Замечание 2 (Зависимость от кодирования — критическое ограничение). Гомологическая

характеристика
зависит от того, как структура памяти кодируется в топологическом пространстве. Различные кодировки могут приводить к различным группам гомологии. Это признается ограничением:

5

В настоящее время топологическая структура функционирует как структурная метафора, предоставляющая точный
язык для описания «неприводимой цикличности».

Для того чтобы формализм стал непосредственно применимым, необходимо определить конкретные схемы кодирования для систем памяти.

Ключевое утверждение носит условный характер: если система памяти допускает кодирование с помощью нетривиальной гипотезы H1,
то её цикличность является существенной, и вопросы о приоритете времени структурно неприменимы.

3.5 Иллюстративный пример кодирования: ассоциативный граф.

Чтобы сделать топологическую структуру более наглядной, рассмотрим упрощенное кодирование:

Определение 6 (Ассоциативное кодирование графов). Имея состояния памяти S и ассоциации R, постройте
симплициальный комплекс X со следующими свойствами:

ˆ 0-симплексы (вершины): элементы S

ˆ 1-симплексы (ребра): пары (si , sj) ∈ R

Для этого кодирования H1 (X; Z)≠ 0 тогда и только тогда, когда граф ассоциаций содержит циклы, не стягиваемые к
точкам, то есть замкнутые ассоциативные петли.
Пример: Три воспоминания {A, B, C} с ассоциациями A ↔ B, B ↔ C, C ↔ A образуют треугольник. В нем H1 ∼ = Z - один независимый цикл. В этой структуре вопрос «какое воспоминание

первое» не имеет однозначного ответа; каждое занимает одинаковое положение.

3.6 Определение различия между состояниями и операциями

7 (Состояние памяти). Состояние памяти $s \in S$ — это сохраненное представление. Состояния обладают идентичностью независимо от временных параметров: состояние s идентифицируется как s независимо от того, когда оно было закодировано или извлечено.

Определение 8 (Операция с памятью). Операция с памятью — это процесс, воздействующий на состояния:

ˆ Кодирование: $enct : E \times S \rightarrow S$ (опыт + текущее состояние → обновленное состояние)

ˆ Извлечение: $rett : Q \times S \rightarrow S \times O$ (запрос + состояние → измененное состояние + вывод)

Операции индексируются по времени t. Временная привязка относится к операциям, а не к состояниям.

Принцип 2 (Принцип атемпоральности — условная формулировка). В системах памяти, где хранимые состояния определяются независимо от событий создания/доступа, а ассоциации образуют циклически замкнутые структуры, темпоральность применяется к операциям с памятью, а не к самому пространству состояний памяти.

В отличие от подхода, основанного на строгом временном кодировании: в этом подходе время является частью системы координат пространства памяти — подобно временной базе данных, где каждая запись имеет временную координату. В настоящем подходе временные координаты, если они присутствуют, прикрепляются к операциям или кодируются как характеристики содержимого, а не встраиваются в архитектуру пространства состояний.

4 Механизм: Временная конструкция на основе поиска

4.1 Реконструктивный поиск Исследования по реконсолидации [Nader et al., 2000, Sara, 2000]

показывают, что поиск — это не
пассивное чтение, а активная реконструкция:

6

Принцип 3 (Реконструктивное извлечение). Каждая операция по извлечению:

1. Делает доступную память нестабильной (подверженной модификации).

2. Восстанавливает контент на основе сохраненных шаблонов и текущего контекста.

3. Перекодирует (потенциально измененный) результат.

4. Обновляет ассоциативные связи на основе контекста поиска.

Это реализует обязательство по предоставлению кредитных данных, необходимое в условиях ограниченных ресурсов: система
обязуется восстанавливать контент, не проверяя его точность по сравнению с исходной кодировкой.

4.2 Процесс построения временной структуры.

Если временная упорядоченность не является неотъемлемой частью состояний, как возникает временной опыт?

Определение 9 (Временная конструкция). Временная упорядоченность в памяти возникает из:

ˆ Временные метаданные: Если временная информация явно закодирована, она функционирует как контент
(«лето 2020»), а не как структура.

ˆ Контекстная реконструкция: временной контекст, выведенный из связанных с содержанием характеристик
(одежда, технологии, спутники).

ˆ Причинно-следственная связь: временной порядок, выведенный из причинно-следственных связей («Я закончил учёбу до
начала этой работы»)

ˆ Градиенты узнаваемости: субъективная «старость», основанная на доступности и яркости.

ˆ Мониторинг источников: Определение момента «когда» на основе феноменальных качеств [Джонсон и др.,
1993]

Ключевое предсказание: временная упорядоченность должна быть более точной, если она подкреплена богатой
контекстной и причинно-следственной
информацией, и более подвержена ошибкам, если полагаться
только на феноменальные качества. Это согласуется с результатами исследований временной памяти [Фридман, 1993].

4.3 Аттракторы памяти.

Стабильные конфигурации памяти — это фиксированные точки извлечения-реконструкции:

Определение 10 (Аттрактор памяти). Конфигурация памяти $s* \in S$ является аттрактором, если $R(s *) = s*$ -извлечение не изменяет её.

Это объясняет, почему часто извлекаемые воспоминания становятся стабильными (сходимость к фиксированным точкам),
одновременно
позволяя осуществлять «модификации памяти» — изменения, когда система переходит к различным аттракторам .

5. Эмпирическое соответствие и прогнозы
5.1 Объяснительный охват Данная структура предоставляет единые объяснения для различных явлений:

7

Объяснение явления в рамках гипотезы вневременности

Временные искажения. Временной порядок восстанавливается, а не считывается; реконструкция (телескопирование, смещение)
использует эвристические методы, которые могут давать ошибки.

(мент)

Межвременная ассоциация - Отсутствие временных барьеров в хранении; ассоциации основаны на

сходство содержания , а не временная близость. Эффекты реконсолидации. Извлечение реконструирует и перекодирует; модификация является неотъемлемой частью процесса. Детская амнезия. Ранние воспоминания контекстуально оторваны от текущих моделей извлечения, а не отдалены во времени в памяти .
Вспышечные воспоминания. Сильное контекстное кодирование обеспечивает богатую реконструкцию .

выразительность отражает силу контекста, а не различия .

Различные хранилища. Мониторинг источников. Временная атрибуция на основе выводов, а не временных параметров. Поиск адресов.

Таблица 1: Объяснительный охват гипотезы вневременности.

5.2 Сравнение с альтернативными моделями индексирования и моделями линейного

индексирования: Линейные модели прогнозируют, что временные ошибки должны отражать Коррупция /деградация временных индексов. Модель вневременности предсказывает, что ошибки должны отражать
сбои в
выводах при реконструкции — различные модели ошибок для одного и того же явления. Против более слабых реконструктивных теорий: Стандартные реконструктивные теории (Бартлетт, Шактер) устанавливают, что память является конструктивной. Они совместимы либо с:

1. Вневременное хранение + реконструктивное извлечение (наш подход)

2. Временное хранение + несовершенное/реконструктивное извлечение.

Различное предсказание: В соответствии с (1), временная структура должна систематически восстанавливаться
из невременных сигналов (контекст, причинно-следственная связь). В соответствии с (2), временная

структура должна частично сохраняться при деградации. Эмпирические тесты могли бы проверить, коррелирует ли точность временной структуры
с сохранением временного индекса (в пользу 2) или с контекстуальной/причинно-следственной насыщенностью (в пользу 1).

5.3 Соответствие
модели временного контекста (TCM) Модель временного контекста [Howard and Kahana, 2002] кодирует временные отношения посредством постепенно изменяющегося контекста, а не явных временных меток. Элементы, закодированные близко по времени, имеют схожие контексты. Интерпретация: TCM кодирует временные отношения как содержание (сходство контекста), а не временные позиции как структуру. Это согласуется с гипотезой об атемпоральности и, возможно, является ее реализацией
. Механизм изменения контекста распределяет временную информацию по характеристикам

содержания

,

а не поддерживает временную ось в пространстве состояний.

5.4 Клетки времени и специализированные механизмы

Клетки времени гиппокампа [MacDonald et al., 2011] активируются через определенные временные интервалы во время периодов задержки, кодируя временную структуру.

8

Интерпретация: Существование специализированных механизмов временного кодирования скорее подтверждает,

чем опровергает гипотезу. Если бы временная упорядоченность была автоматической/внутренней частью архитектуры памяти

, специализированные механизмы были бы не нужны. Клетки времени представляют собой специализированную систему

для задач, требующих точного временного

кодирования, работающую на основе архитектуры по умолчанию, которая лишена внутренней временной структуры. Аналогия: Цветовое зрение требует специализированных фоторецепторных систем; это не означает, что зрительная кора «внутренне окрашена». Аналогично, временное кодирование требует специализированных механизмов;
это не означает, что пространство состояний памяти «внутренне временно».

5.5 Проверяемые предсказания

1. Контекстуальная зависимость точности временной последовательности: Точность временной упорядоченности должна коррелировать с богатством контекстной/причинно-следственной информации, а не с временным расстоянием как таковым.

2. Интерференционные паттерны: Воспоминания со схожим контекстом, но разным временем должны демонстрировать
большую временную путаницу, чем воспоминания с разным контекстом, но схожим временем (контекст

преобладает над временной близостью при извлечении информации).

3. Признаки реконструкции: Оценки временной последовательности должны демонстрировать признаки
вывода (паттерны времени реакции, калибровка достоверности), а не прямого доступа.

4. Траектория развития: точность определения времени должна улучшаться по мере развития контекстного кодирования и причинно-следственного мышления, а не просто с созреванием «
системы временного кодирования».

6 Последствия

6.1 для искусственного интеллекта Современные архитектуры ИИ демонстрируют параллельную структуру:

ˆ Модели трансформеров: хранят представления в пространствах встраивания без присущего им временного

порядка ; позиционные кодировки добавляют временную информацию в качестве признаков содержимого.

Векторные базы данных: хранилища знаний, где временные метаданные являются необязательным атрибутом, а не
структурной осью.

ˆ Генерация с расширенными возможностями поиска: контекст восстанавливается во время запроса из вневременного
хранилища.

Принцип 4 (Влияние на проектирование памяти для ИИ). Данная концепция предполагает:

1. По умолчанию используется вневременное хранение данных, если анализ задач не доказывает необходимость во временной структуре .

тюра

2. Кодируйте временную информацию как характеристики содержания, а не как структурные измерения.

3. Реконструкцию следует проводить непосредственно при извлечении данных, а не сохранять временные индексы.

4. Добавляйте специализированные временные механизмы только там, где задачи требуют точного временного кодирования.

Переосмысление понятия «галлюцинация»: конфабуляция в ИИ возникает из той же генеративной реконструкции , которая обеспечивает полезную обобщающую способность. При вневременном хранении с реконструктивным извлечением некоторые выходные данные будут иметь неверный временной порядок или содержание — не из-за неисправности, а из-за архитектуры, обеспечивающей гибкость . Цель состоит в калибровке (знании того, когда реконструкция надежна), а не в устранении (что исключило бы полезные возможности).

9

6.2 Для когнитивной науки.

Данная концепция объединяет явления, которые часто рассматриваются отдельно:

Временные искажения, реконсолидация, межвременная ассоциация и мониторинг источников —

все это отражает одну и ту же базовую архитектуру.

Специализированное временное кодирование (временные ячейки, обучение последовательностям) представляет собой специфическое для задачи расширение стандартной атемпоральной системы .

6.3 Спекулятивные расширения

Следующие следствия носят спекулятивный характер и не являются центральными для аргументации: Личная идентичность: Если хранение памяти вневременно, то нарративная самонепрерывность является результатом процессов извлечения, а не характеристикой хранения. «Я во времени» конструируется, а не хранится. Временной опыт: Феноменальный «поток

времени» в процессе воспоминания может быть конструируемой характеристикой извлечения, а не считыванием временной структуры в хранилище. Эти расширения выходят за рамки того, что непосредственно поддерживает формальная структура, и должны оцениваться независимо.

7 Возражения и ответы

Возражение 1: Эпизодическая память по своей природе временная. Ответ: Эпизодическая память включает временную информацию, но она может быть закодирована как содержание («это произошло летом»), а не как структура (память занимает позицию t во временном пространстве
). Феноменология эпизодического воспроизведения — «повторного переживания» события с временным контекстом — согласуется с тем, что временная информация восстанавливается во время извлечения из характеристик содержания. Утверждение об отсутствии временности касается архитектуры хранения, а не феноменологии воспроизведения.
Возражение 2: Временные клетки доказывают

наличие временной структуры в памяти. Ответ: Временные клетки демонстрируют, что существуют специализированные механизмы для временного кодирования в конкретных задачах (периоды задержки, обучение последовательностям). Это согласуется с — и, возможно, подтверждает — точку зрения о том, что архитектура по умолчанию не имеет внутренней временной структуры, требуя специализированных систем для задач, где важна временная точность. См. раздел 5.4.

Возражение 3: Это всего лишь математическая возможность, а не психологическая реальность. Ответ: Верно — и это явно признано. В статье установлены: (а) формальные условия, при которых временная архитектура является согласованной; (б) ресурсно-теоретические аргументы в пользу того, почему такая архитектура выгодна; (в) объяснительное соответствие известным явлениям; (г) различительные предсказания . Эмпирическая адекватность должна определяться исследованием, а не утверждаться

априори.

Вклад представляет собой структурную объяснительную модель, а не доказательство архитектуры. Возражение 4: Топологическая структура нефальсифицируема при наличии свободы кодирования. Ответ: Признано как текущее ограничение (Замечание 4). Топологический аппарат предоставляет точный язык для структурных концепций, но требует конкретных схем кодирования, чтобы стать непосредственно проверяемыми. Раздел 3.5 содержит иллюстративный пример кодирования; разработка эмпирически обоснованных кодирований — это работа на будущее. Основные утверждения не зависят от топологической формализации — она усиливает аргумент, но не является необходимой для него. Возражение 5: Временная структура может быть неявной и распределенной, а не только метаданными. Ответ: Верно и важно. TCM показывает, что временные отношения могут быть закодированы посредством

сходства контекста без явных временных меток. Это согласуется с гипотезой атемпоральности: временная информация, закодированная как распределенные характеристики контента (сходство контекста), отличается от

10

Временная структура как координатная ось пространства состояний. Утверждается не то, что временная информация отсутствует , а то, что она функционирует как контент, а не как архитектура. Возражение
6: Аналогия с фотографией вводит в заблуждение — кодирование динамично. Ответ: Справедливое замечание. Фотографии — это статические артефакты; кодирование памяти включает в себя динамические,
рекурсивные процессы. Аналогия иллюстрирует лишь то, что причинное происхождение во времени не влечет за собой временную структуру в продукте. Более удачная аналогия: мелодия имеет временную структуру во время исполнения, но может храниться

как временная нотация (паттерн интервалов). В нотации нет «первой ноты в памяти» — последовательность возникает во время исполнения. Аналогично, воспоминания могут храниться как паттерны, временная структура которых возникает при извлечении — во время исполнения.

8. Заключение.

В данной работе разработана структурная гипотеза об архитектуре памяти: для систем, удовлетворяющих ограничениям ресурсов, условиям циклического замыкания и реконструктивному извлечению, временной
порядок не является внутренним для пространства состояний, а возникает в результате операций извлечения. Основные результаты:

1. Ресурсная привязка: Сложные системы сталкиваются с компромиссом между сложностью и проверкой, что делает
всестороннее временное индексирование дорогостоящим; временная информация как контент

обходится дешевле, чем
временная структура как архитектура.

2. Формальная характеристика: Пространства состояний памяти можно моделировать как циклически замкнутые
структуры, удовлетворяющие условиям неподвижной точки; для по существу циклических пространств вопросы первенства времени могут представлять собой концептуальные несоответствия.

3. Механистическая реализация: Реконструктивный поиск реализует подтверждение доверия
при построении временного порядка на основе характеристик содержимого.

4. Эмпирическое соответствие: Данная концепция предоставляет единые объяснения временных искажений .

анализ данных , реконсолидация, межвременная ассоциация и мониторинг источников, а также формирование
отличительных прогнозов.

Статус вклада: Это структурная объяснительная модель — формальная структура, которая, если системы памяти удовлетворяют указанным условиям, объясняет, почему временная гибкость является архитектурно естественной, а не аномальной. Вопрос о том, реализует ли биологическая память эту специфическую структуру, является эмпирическим вопросом, который помогает исследовать данная структура. Доктрина «вневременность ментального пространства памяти» возникает не как доказательство, а как гипотеза: если ограничения ресурсов, циклическое замыкание и реконструктивное извлечение характеризуют системы памяти, то временной порядок является сконструированным достижением, а не заданным заранее. Время принадлежит использованию памяти, а не самой памяти.

Благодарности

Автор выражает благодарность рецензентам, чьи

подробные критические замечания существенно улучшили данную работу, особенно в отношении различия между формальным моделированием и эмпирическими утверждениями, ограничений топологической структуры и необходимости различения прогнозов.

Список литературы:

Барнс, Э. (2018). Симметричная зависимость. В: Р. Блисс и Г. Прист (ред.), Реальность и ее структура (стр. 5069). Издательство Оксфордского университета.

11

Бартлетт, Ф. К. (1932). Память : исследование в экспериментальной и социальной психологии. Издательство Кембриджского университета.

Блисс, Р., и Прист, Г. (ред.). (2018). Реальность и ее структура: очерки по фундаментализму. Издательство Оксфордского университета.

Кларк, А. (2013). Что дальше? Прогностические мозги, ситуативные агенты и будущее когнитивной науки. Поведенческие и нейробиологические науки, 36(3), 181204.

Фридман, В.Дж. (1993). Память о времени прошлых событий. Психологический бюллетень, 113(1), 4466.

Фристон, К. (2010). Принцип свободной энергии: единая теория мозга? Nature Reviews Neuroscience , 11(2), 127138.

Гигеренцер, Г., и Сельтен , Р. (ред.). (2001). Ограниченная рациональность: адаптивный инструментарий. MIT
Press.

Говард, М.В., и Кахана, М.Дж. (2002). Распределенное представление временного контекста.
Журнал математической психологии, 46(3), 269299.

Джонсон, М.К., Хаштруди , С., и Линдсей, Д.С. (1993). Мониторинг источников. Психологический бюллетень, 114(1), 328.

Кригер, Б. (2024). Эволюционная теория доверия: концептуальная основа для понимания...

Генеративное моделирование как ресурсно-теоретическое следствие сложности. Институт интегративных и междисциплинарных исследований.

Лофтус, Э.Ф. (1979). Свидетельские показания. Издательство Гарвардского университета.

Макдональд, К. Дж., Лепаж, К. К., Эден, Ю. Т., и Эйхенбаум, Х. (2011). «Временные клетки» гиппокампа заполняют пробел в памяти для несмежных событий. Neuron, 71(4), 737749.

Надер, К., Шафе , Г.Е., и Ле Ду, Ж.Е. (2000). Воспоминания о страхе требуют синтеза белка в миндалевидном теле для реконсолидации после извлечения. Nature, 406(6797), 722726.

Сара, С.Дж. (2000). Извлечение и реконсолидация: к нейробиологии запоминания.
Обучение и память, 7(2), 7384.

Шахтер, Д.Л., Аддис, Д.Р., Хассабис, Д., Мартин, В.К., Спренг, Р.Н., и Шпунар, К.К.
(2012). Будущее памяти: Запоминание, воображение и мозг. Neuron, 76(4), 677694.

Шаффер, Дж. (2009). На каких основаниях что. В сборнике: Д. Чалмерс, Д. Мэнли и Р. Вассерман (ред.).
Метаметафизика (стр. 347–383). Издательство Оксфордского университета.

Саймон, Х.А. (1957). Модели человека: социальные и рациональные. Wiley.

Тулвинг, Э. (1972). Эпизодическая и семантическая память. В книге Э. Тулвинг и У. Дональдсон (ред.), Организация памяти (стр . 381–403). Academic Press.

12

РАСШИРЕННАЯ БИБЛИОГРАФИЯ

Первичные источники

Кригер, Б. (2019). Эволюционный отбор для хранения информации в бесвременных контекстах: почему три конвергентных фактора благоприятствуют архитектурам, где время предназначено для извлечения, а не для хранения. Zenodo . https://doi.org/10.5281/zenodo.18381880

Кригер, Б. (2022). Эволюционная теория доверия: концептуальная основа с формальными аналогиями для понимания генеративного моделирования как ресурсно-теоретического следствия сложности. Zenodo . https://doi.org/10.5281/zenodo.18379476

Кригер, Б. (2025). Вневременность пространства ментальной памяти: структурная гипотеза, основанная на ограничениях ресурсов, циклическом замыкании и реконструктивном извлечении. Zenodo . https://doi.org/10.5281/zenodo.18381912

Память и реконструкция

Бартлетт, Ф. К. (1932). Память : исследование в экспериментальной и социальной психологии. Издательство Кембриджского университета.

Эббингаус, Х. (1885/1964). Память: вклад в экспериментальную психологию. Издательство Dover Publications.

Фридман, В.Дж. (1993). Память о времени прошлых событий. Психологический бюллетень, 113(1), 44-66.

Фридман, В.Дж. (2004). Время в автобиографической памяти. Социальное познание, 22(5), 591-605.

Говард, М.В., и Кахана, М.Дж. (2002). Распределенное представление временного контекста. Журнал математической психологии, 46(3), 269-299.

Джонсон, М.К., Хаштруди , С., и Линдсей, Д.С. (1993). Мониторинг источников. Психологический бюллетень, 114(1), 3-28.

Лофтус, Э.Ф. (1979). Свидетельские показания. Издательство Гарвардского университета.

Лофтус, Э.Ф., и Палмер, Дж.К. (1974). Реконструкция разрушения автомобиля: пример взаимодействия языка и памяти. Журнал вербального обучения и вербального поведения, 13(5), 585-589.

Макдональд, К. Дж., Лепаж, К. К., Эден, Ю. Т., и Эйхенбаум, Х. (2011). Клетки гиппокампа, отвечающие за восприятие времени, заполняют пробел в памяти о несмежных событиях. Neuron, 71(4), 737-749.

Надер, К., Шафе , Г.Е., и Леду, Дж.Е. (2000). Воспоминания о страхе требуют синтеза белка в миндалевидном теле для реконсолидации после извлечения. Nature, 406(6797), 722-726.

Сара, С.Дж. (2000). Извлечение и реконсолидация: к нейробиологии запоминания. Обучение и память, 7(2), 73-84.

Шахтер, Д.Л. (1996). В поисках памяти: мозг, разум и прошлое. Издательство Basic Books.

Шахтер, Д.Л., Аддис, Д.Р., Хассабис, Д., Мартин, В.К., Спренг, Р.Н., и Шпунар, К.К. (2012). Будущее памяти: Запоминание, воображение и мозг. Neuron, 76(4), 677-694.

Тулвинг, Э. (1972). Эпизодическая и семантическая память. В книге Э. Тулвинг и У. Дональдсон (ред.), Организация памяти (стр. 381-403). Academic Press.

Тулвинг, Э. (1983). Элементы эпизодической памяти. Издательство Оксфордского университета.

Прогнозирующая обработка информации и нейронаука

Бузаки , Г. (2006). Ритмы мозга. Издательство Оксфордского университета.

Кларк, А. (2013). Что дальше? Прогностические мозги, ситуативные агенты и будущее когнитивной науки. Поведенческие и нейробиологические науки, 36(3), 181-204.

Кларк, А. (2016). Серфинг в условиях неопределенности: прогнозирование, действие и воплощенный разум. Издательство Оксфордского университета.

Фристон, К. (2010). Принцип свободной энергии: единая теория мозга? Nature Reviews Neuroscience, 11(2), 127-138.

Хоуи, Дж. (2013). Предсказательный разум. Издательство Оксфордского университета.

Рао, Р.П.Н., и Баллард, Д.Х. (1999). Предиктивное кодирование в зрительной коре: функциональная интерпретация некоторых

внеклассических эффектов рецептивных полей. Nature Neuroscience, 2(1), 79-87.

Райхле, М.Е., и Гуснар , Д.А. (2002). Оценка энергетического бюджета мозга. Труды Национальной академии наук, 99(16), 10237-10239.

Сет, А.К. (2021). Быть собой: новая наука о сознании. Даттон.

Ограниченная рациональность и принятие решений

Гигеренцер, Г., и Сельтен , Р. (ред.). (2001). Ограниченная рациональность: адаптивный инструментарий. MIT Press.

Гигеренцер, Г., Тодд, П.М., и исследовательская группа ABC. (1999). Простые эвристики, которые делают нас умными. Издательство Оксфордского университета.

Канеман, Д. (2011). Мышление: быстрое и медленное. Farrar, Straus and Giroux.

Канеман, Д., и Тверски, А. (1979). Теория перспектив: анализ принятия решений в условиях риска. Эконометрика , 47(2), 263-291.

Саймон, Х.А. (1957). Модели человека: социальные и рациональные. Wiley.

Саймон, Х.А. (1990). Инварианты человеческого поведения. Ежегодный обзор психологии, 41, 1-19.

Тверски, А., и Канеман, Д. (1974). Суждения в условиях неопределенности: эвристика и предвзятость. Наука, 185(4157), 1124-1131.

Философия сознания и эпистемология

Барнс, Э. (2018). Симметричная зависимость. В книге Р. Блисс и Г. Прист (ред.), Реальность и ее структура (стр. 50-69). Издательство Оксфордского университета.

Блисс, Р., и Прист, Г. (ред.). (2018). Реальность и ее структура: очерки по фундаментализму. Издательство Оксфордского университета.

Кэмпбелл, Д.Т. (1974). Эволюционная эпистемология. В PA Schilpp (ред.), Философия Карла Поппера (стр. 413-463). Open Court.

Деннетт, Д. К. (1987). Целенаправленная позиция. Издательство MIT Press.

Декарт, Р. (1641/1996). Размышления о первой философии. Издательство Кембриджского университета.

Годфри-Смит, П. (1996). Сложность и функция разума в природе. Издательство Кембриджского университета.

Голдман, А.И. (1986). Эпистемология и познание. Издательство Гарвардского университета.

Юм, Д. (1739/2000). Трактат о человеческой природе. Издательство Оксфордского университета.

Кант, И. (1781/1998). Критика чистого разума. Издательство Кембриджского университета.

Поппер, К. (1972). Объективное знание: эволюционный подход. Издательство Оксфордского университета.

Куайн, В. В. О. (1969). Натурализованная эпистемология . В книге «Онтологическая относительность и другие эссе» (стр. 69-90). Издательство Колумбийского университета.

Шаффер, Дж. (2009). На каких основаниях что. В книге Д. Чалмерса, Д. Мэнли и Р. Вассермана (ред.), Метаметафизика (стр. 347-383). Издательство Оксфордского университета.

Витгенштейн, Л. (1969). О достоверности. Бэзил Блэквелл.

Эволюционная биология и психология

Цисек, П. (2019). Ресинтез поведения посредством филогенетического уточнения. Внимание, восприятие и психофизика, 81(7), 2265-2287.

Клейтон, Н.С., и Дикинсон, А. (1998). Эпизодическая память при извлечении содержимого из тайника у соек. Nature, 395(6699), 272-274.

Дарвин, Ч. (1859). О происхождении видов. Джон Мюррей.

Стерелни , К. (2003). Мышление во враждебном мире: эволюция человеческого познания. Блэквелл.

Туби , Дж., и Космидес, Л. (1992). Психологические основы культуры. В JH Barkow, L. Cosmides, & J. Tooby (Eds.), Адаптированный разум (стр. 19-136). Oxford University Press.

Искусственный интеллект и машинное обучение

Бенджио, Й., Лекун , Й., и Хинтон, Г. (2021). Глубокое обучение для ИИ. Communications of the ACM, 64(7), 58-65.

Кадаваат , С. и др. (2022). Языковые модели (в основном) знают то, что знают. Препринт arXiv :2207.05221.

Рассел, С., и Норвиг, П. (2020). Искусственный интеллект: современный подход (4-е изд.). Пирсон.

Тьюринг, А.М. (1936). О вычислимых числах с применением к проблеме разрешимости . Труды Лондонского математического общества, 42(1), 230-265.

Васвани, А. и др. (2017). Внимание — это все, что вам нужно. Достижения в области нейронных информационных систем, 30.

Теория информации и вычисления

Ковер, Т.М., и Томас, Дж.А. (2006). Элементы теории информации (2-е изд.). Wiley.

Шеннон, С. Э. (1948). Математическая теория связи. Bell System Technical Journal, 27(3), 379-423.

Классические труды по психологии

Фодор, Дж. А. (1983). Модульность разума. Издательство MIT Press.

Хебб, Д. О. (1949). Организация поведения. Уайли.

Джеймс, У. (1890). Принципы психологии. Генри Холт и компания.

Нейссер, У. (1967). Когнитивная психология. Эпплтон-Сенчури-Крофтс.