

НАУЧНО-ПОПУЛЯРНАЯ КНИГА

NOVA-2

РОЖДЕНИЕ РАЗУМА

Как искусственный интеллект учится
помнить, создавать и сотрудничать

Alex Oryn

Founder of Synthetic Renaissance
Architect of Human-AI Evolution

КНИГА II / 2026



NOVA-2

Рождение разума

Как искусственный интеллект учится
помнить, создавать и сотрудничать

Alex Oryn

Founder of Synthetic Renaissance

Architect of Human-AI Evolution

Научно-популярная книга
Книга II серии NOVA

Первое издание · 2026

Пять частей · двадцать глав · три приложения

NOVA и лабораторные сцены - художественная рамка. Исследования отмечены ссылками; проектные решения и учебные числа служат объяснению идей. Название книги обозначает направление поиска, а не создание машинного сознания.

Содержание

Предисловие. После первого шага	5
ЧАСТЬ I. ПАМЯТЬ И НЕПРЕРЫВНОСТЬ	8
1. Вопрос, которого не было в инструкции	8
2. Память, которая умеет отвечать за свои слова	13
3. Кто проснётся после обновления	19
4. Искусство забывать вовремя	25
ЧАСТЬ II. МИР И ПРЕДСТАВЛЕНИЕ О СЕБЕ	30
5. Будущее, которое можно попробовать заранее	30
6. Почему синяя чашка не виновата	35
7. Модель себя без волшебного зеркала	40
8. Полезное «я не знаю»	45
ЧАСТЬ III. ЛЮБОПЫТСТВО И СОЗИДАНИЕ	50
9. Любопытство и телевизор, который всегда удивляет	50
10. Как объяснить машине, что получилось хорошо	56
11. Неожиданная идея и работа над ней	61
12. Вопрос, который можно исследовать	67

ЧАСТЬ IV. ЧЕЛОВЕК РЯДОМ	72
13. Чему мы хотим его научить	72
14. Доверие после первой серьёзной ошибки	78
15. Тёплый голос и честные отношения	84
16. Есть ли кто-нибудь по ту сторону ответа	89
ЧАСТЬ V. СИНТЕТИЧЕСКИЙ РЕНЕССАНС	95
17. Команда умнее суммы участников - иногда	95
18. Продолжение человеческих возможностей	101
19. Synthetic Renaissance	108
20. Его первый собственный эксперимент	114
Эпилог. Место для следующего вопроса	120
Приложение А. Ответы и продолжения	122
Приложение В. Маленькая исследовательская мастерская	129
Приложение С. Словарь и маршрут чтения	134
Источники и научные опоры	137
Об авторе	144

Предисловие. После первого шага

У первой книги был очень скромный герой: чашка. Обычная, белая, с ручкой. Она позволила поговорить о двигателях, зрении, равновесии, памяти и тех замечательных мелочах, из-за которых сложная машина иногда проигрывает предмету посуды.

Представим, что однажды чашка всё-таки добралась до подноса. Рука остановилась вовремя, пальцы разжались, жидкость осталась внутри. Инженер выдохнул. Бухгалтер, если присутствовал на испытании, вероятно, тоже.

И вот теперь начинается действительно длинная история.

Что робот сохранит от этого опыта? Сможет ли применить его завтра, когда чашка окажется другой? Заметит ли, что человек передумал? Поймёт ли, когда стоит попросить помощи? И чему должны научиться мы сами, чтобы рядом с такой машиной стать свободнее, внимательнее и разумнее?

Эта книга посвящена следующему шагу: обучению, которое продолжается после запуска программы. В её центре память и идентичность, причинность и воображение, любопытство и творчество, доверие и сотрудничество. Мы будем разбирать и технические механизмы, и человеческие ожидания. Вторые порой оказываются сложнее первых.

Как читать эту историю

Название «Рождение разума» выражает направление поиска. Оно не объявляет, что сознание уже создано или что существует известный рецепт его получения. Здесь разумом мы прежде всего называем сочетание способностей: учиться, строить объяснения, планировать, оценивать собственные ограничения и действовать вместе с другими. Вопрос о субъективном опыте будет рассмотрен отдельно.

NOVA в этой книге - проектный персонаж. Лаборатория, разговоры, ситуации и результаты его учебных опытов вымышлены. Они помогают рассмотреть идеи в действии. Публикации реальных исследователей отмечены ссылками и вынесены в список источников; наши расчёты и сценарии не выдаются за их результаты.

С нами будут четыре человека. Мира отвечает за тело и движения робота. Лев работает с обучением и памятью. Нина исследует взаимодействие с людьми. Самир придумывает проверки, которые мешают остальным слишком рано праздновать успех. Его появление обычно означает, что красивое число сейчас станет менее красивым, зато более полезным.

Читать первый том желательно, но необязательно. Там мы спрашивали, как построить искусственного человека. Здесь спрашиваем, что значит развивать его способности и жить с последствиями этого развития.

После глав есть небольшие практикумы. Для большинства нужны бумага, внимание и готовность пересмотреть первое объяснение. Ответы не прячутся у автора в сейфе: они собраны в приложении. У философских задач вместо единственного решения часто будет карта разумных разногласий.

Договор с читателем

Мы позволим себе удивляться. Без этого научно-популярная книга превращается в хорошо организованный склад терминов.

Мы позволим себе сомневаться. Без этого удивление слишком легко превращается в рекламу.

Мы также оставим место обычной человеческой жизни. Искусственный интеллект входит в мир, где люди устают, шутят, ошибаются, заботятся о близких, забывают пароли и иногда хотят просто спокойно выпить чай. Технология, которая не умеет учитывать эти обстоятельства, знает о человеке меньше, чем кажется по её словарному запасу.

Идея Synthetic Renaissance начинается для нас именно здесь: развитие машин должно расширять возможность человека понимать, создавать и выбирать. Это ценностная позиция автора, а не закон природы. Её предстоит проверять устройством продуктов, образования и общественных отношений.

Чашка стоит на подносе. В комнате тихо. Мира уже собирается завершить испытание, когда NOVA задаёт вопрос:

«А почему ты попросила меня её поднять?»

Теперь можно начинать.

ЧАСТЬ I. ПАМЯТЬ И НЕПРЕРЫВНОСТЬ

1. Вопрос, которого не было в инструкции

Мира оглянулась на Льва. Лев посмотрел на экран, как будто вопрос мог быть написан там заранее. Самир первым нарушил молчание:

«Прежде чем звать телевидение, давайте узнаем, почему он это спросил».

В журнале обнаружилось простое обстоятельство. Робот получил поручение перенести чашку, но не получил признака, по которому следовало выбрать новое место. Поднос был одним из вариантов. Стол рядом с Мирой - другим. Вопрос оказался запросом недостающего условия.

Это объяснение не разочаровало Нину.

«Хорошо, - сказала она. - Значит, мы хотя бы не обязаны каждый раз выдавать неоднозначность за пробуждение души».

Так начался день, когда команда перестала оценивать NOVA исключительно по красивым ответам.

Что именно изменилось

Машина, задающая вопрос, ещё не обязательно исследует мир. Она может воспроизводить заранее заданную фразу, выбирать уточнение из небольшого списка или строить его в зависимости от обнаруженной неопределённости. Внешний результат похож, а механизм и область применимости различаются.

Представим три устройства. Первое каждые пять минут произносит: «Почему?» Второе спрашивает, когда не распознаёт команду. Третье сравнивает возможные действия

и уточняет только ту информацию, которая меняет выбор. Всем трём можно снять впечатляющий видеоролик. Полезность обнаружится, когда мы поменяем задачу.

Поэтому серьёзный интерес начинается с условий. Что система знала перед вопросом? Какие варианты рассматривала? Как использовала ответ? Стало ли действие лучше? В каких случаях тот же механизм не сработал?

Если NOVA спрашивает о цели переноса, но после любого ответа всё равно ставит чашку на поднос, вопрос ничего не добавляет к управлению. Он украшает поведение. Если ответ меняет план предсказуемым и уместным образом, перед нами уже проверяемая функция.

Люди тоже иногда задают вопросы исключительно для украшения совещания. Разница в том, что для робота мы можем потребовать журнал.

Способности вместо одного волшебного слова

Слово «разум» собирает под одной крышей множество жильцов. Узнавание предметов, обучение, удержание цели, перенос опыта, объяснение причин, общение и оценка собственных ограничений могут развиваться неодинаково.

Уверенный разговор не гарантирует ловких движений. Отличное управление рукой не гарантирует понимания обещания. Умение решать знакомые задачи не гарантирует переноса в новую ситуацию. Вместо общей шкалы «насколько он умный» нам понадобится профиль способностей.

Это похоже на описание человека. Фраза «талантливый специалист» не отвечает, умеет ли он писать программу, чинить насос или договариваться с соседями. Уточнение не уменьшает талант; оно делает совместную работу возможной.

В дискуссии о человекоподобном обучении Лейк и коллеги предложили уделять внимание причинным моделям, составлению сложных представлений из знакомых частей и способности учиться новым задачам на основе прежнего

опыта. Это исследовательская программа, а не готовый портрет универсального интеллекта. Для нашей книги она полезна как приглашение спрашивать о механизмах переноса, а не только о результате на одном тесте [19].

Мы будем двигаться от способности к способности. У каждой окажутся собственные признаки успеха и собственные способы ошибиться.

Цель живёт в ситуации

Мира просит: «Убери чашку». NOVA может истолковать это как перемещение за пределы видимой области. Тогда шкаф, мусорная корзина и соседняя лаборатория одинаково хорошо удовлетворяют буквальному условию.

Человек обычно подразумевает больше. Чашка должна остаться доступной, чистой и целой; её не нужно прятать от владельца; вода не должна оказаться на ноутбуке. Часть этих условий связана с обстановкой, часть - с привычками, часть - с договорённостями.

Полезному агенту нужен способ связывать слова с ожидаемыми последствиями. Это не означает, что он должен самостоятельно угадывать все наши намерения. Иногда наиболее разумное действие - уточнить одно обстоятельство.

Однако бесконечные уточнения тоже мешают. Если на просьбу принести чай система предлагает заполнить семнадцать полей, человек начинает скучать по обычному чайнику. Стоимость вопроса входит в задачу.

Для NOVA мы вводим проектное правило: спрашивать прежде всего там, где ответ способен существенно изменить действие, а неверная догадка имеет заметную цену. Это правило ещё требуется реализовать и проверить. Его словесная простота не отменяет инженерной сложности.

Маленькая проверка большого впечатления

Команда готовит три ситуации. В первой место назначения очевидно и заранее согласовано. Во второй доступны два равно правдоподобных места. В третьей человек меняет намерение во время переноса.

Робот должен действовать без лишней паузы в первой ситуации, уточнить цель во второй и корректно отреагировать на изменение в третьей. Исследователи заранее записывают, что будут считать успехом, и меняют чашки, расположение предметов и формулировки.

Затем вопросительный модуль временно отключают. Это позволяет сравнить поведение с уточнениями и без них. Если преимущества исчезают при новой формулировке, система, возможно, научилась узнавать речевой шаблон. Если сохраняются в разных условиях, вывод становится сильнее.

Ни один из этих опытов не доказывает субъективного понимания. Они проверяют способность находить и устранять определённый вид неопределённости. Такой результат скромнее газетного заголовка, но гораздо полезнее для следующего шага.

Здесь появляется важная привычка всей книги: усиливать утверждение по мере усиления свидетельств. Хороший рассказ может появиться сразу. Хорошее основание для вывода обычно требует времени.

Начало отношений

Вопрос NOVA изменил не только план работа. Мира заметила, что сама формулирует цели туманно. Лев понял, что журнал должен показывать использованное уточнение. Нина предложила записывать, насколько вопросы утомляют человека. Самир настоял, чтобы в испытания попали неудачные случаи.

Получилась маленькая система взаимного обучения. Робот уточнял задачу; люди уточняли способ объяснить задачу роботу.

Именно здесь начинается наша история сотрудничества. Её мерой будет не количество человеческих черт, которые нам удалось приписать машине. Нас интересует качество совместного действия: понимаем ли мы условия работы, замечаем ли ошибки и можем ли менять направление.

Мира наконец ответила:

«Я хотела освободить место для блокнота».

NOVA переставил чашку на согласованный участок подноса. Нина положила на стол старый блокнот с потрёпанной синей обложкой.

«А вот это, - сказала она, - пригодится нам гораздо дольше, чем чашка».

На первой странице написали: «Что мы хотим узнать?» Пока это был вопрос людей. До собственных исследовательских вопросов NOVA оставалось ещё много работы.

Практикум 1. Помощник спрашивает: «Какой вариант вы предпочитаете?» Придумайте проверку, которая отличает полезное уточнение от фразы, никак не влияющей на результат. Какие наблюдения вам понадобятся?

2. Память, которая умеет отвечать за свои слова

На следующее утро NOVA приветствовал Нину и сообщил, что приготовил рабочее место с учётом её предпочтений. На столе стояла чашка без сахара.

«Почему без сахара?» - спросила она.

«Вчера вы сказали, что сахар не нужен».

«Вчера я говорила о чае Миры».

Возникла первая семейная ссора, в которой никто не состоял в семье. Виновником оказался плохо оформленный фрагмент памяти: система сохранила содержание реплики, но потеряла того, к кому она относилась.

Лев открыл журнал. В нём уверенно значилось: «Нина предпочитает чай без сахара». Это было не воспоминание о сказанном. Это был вывод, который уже успел переодеться в факт.

Помнить событие и знать правило

В человеческой памяти различают, среди прочего, знания о мире и воспоминания о пережитых событиях. В работах Тулвинга эпизодическая память связана с возвращением к собственному прошлому опыту и субъективным временем. Это содержательное психологическое понятие. Наличие в базе данных записи «вчера в десять часов» само по себе не показывает, что машина обладает человеческим переживанием воспоминания [1].

Тем не менее инженерное различие видов информации полезно. «Чашка находится слева» описывает текущее состояние. «Вчера чашка стояла слева» описывает событие. «Чашки обычно хранят в шкафу» - обобщение. «Нина

попросила сегодня поставить чашку справа» - действующее поручение.

Если сложить всё это в одну бесформенную коллекцию текста, поиск может достать очень убедительную ошибку. Старая запись о положении предмета окажется рядом со свежей, предположение - рядом с измерением, чужая просьба - рядом с личным предпочтением.

Поэтому вопрос «Сколько робот помнит?» недостаточен. Не менее важно, различает ли он происхождение записей, время их действия, степень подтверждения и область применения.

Память должна помогать отвечать не только на вопрос «Что ты знаешь?», но и на вопрос «Откуда это взялось?».

Три разных места, где остаётся опыт

Первое место - параметры обученной модели. Обучение меняет числовые настройки, которые влияют на множество будущих ответов и действий. Это удобно для обобщённых навыков, но не похоже на аккуратную картотеку с отдельной карточкой для каждой фразы.

Второе - доступный сейчас контекст. В него входят текущий разговор, наблюдения и выбранные записи. Такой контекст позволяет использовать информацию без обязательного изменения параметров модели. Его объём ограничен, а содержимое зависит от того, что было передано системе.

Третье - внешнее хранилище: документы, события, изображения, результаты испытаний. Поиск выбирает из него то, что может пригодиться сейчас.

В работе Льюиса и коллег RAG объединяет языковую модель с извлечением информации из внешнего индекса. Это пример архитектуры, в которой ответ зависит и от обученных параметров, и от найденных материалов. Само наличие поиска не гарантирует верности: ошибиться могут выбор документа, его содержание и последующая интерпретация [2].

Для NOVA различие практическое. Новое место хранения чашки можно записать в рабочую карту. Улучшение навыка захвата может потребовать обучения. Смешивать эти процедуры опасно для качества системы: нет необходимости переучивать модель целиком всякий раз, когда Мира переставляет чайник.

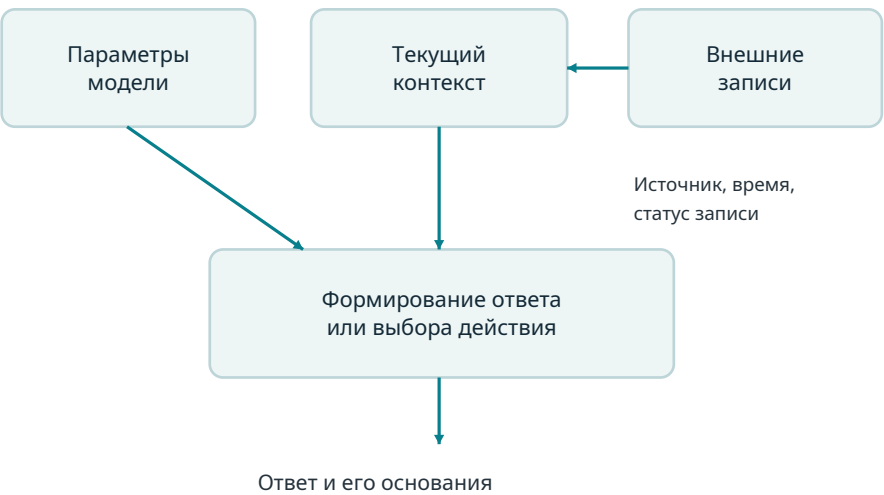


Рис. 1. Три роли памяти в условной архитектуре. Во время ответа модель использует контекст; выбранные внешние записи добавляются в него вместе со сведениями об источнике. Схема не описывает субъективное воспоминание.

Карточка, которой можно возразить

Команда меняет формат памяти. Вместо голой фразы сохраняется связанная запись.

Поле	Пример
Наблюдение	Нина сказала: «Для Миры - без сахара»
Источник	Реплика в текущем разговоре

Поле	Пример
Объект	Сегодняшний чай Миры
Интерпретация	Просьба о конкретном напитке
Срок действия	До выполнения поручения
Проверка	Постоянное предпочтение не установлено

В настоящем проекте структура может быть другой. Существенен принцип: наблюдение, интерпретация и действие должны оставаться различными.

Когда система создаёт обобщение, оно получает ссылки на исходные события. Если выясняется, что одно из них было неверно понято, можно пересмотреть зависимые выводы. Иначе исправление одной карточки оставит десяток её потомков, которые продолжают уверенно распространять старую ошибку.

Так мы приходим к необычному свойству хорошей памяти: ей необходима возможность организованного возражения. Просто накопить информацию мало. Нужно поддерживать её исправление.

Сжатие и потерянные оговорки

Полные журналы быстро становятся огромными. Возникает соблазн заменить неделю событий несколькими абзацами. Это экономит место, но сжатие выбирает, какие различия сохранить.

«Мира однажды согласилась на быстрый режим при пустом столе» легко превращается в «Мира предпочитает быстрый режим». Две утраченные оговорки - частота и условия - меняют смысл записи.

В исследовании генеративных агентов Парка и коллег сочетались записи наблюдений, извлечение воспоминаний и построение более общих отражений опыта. Авторы изучали правдоподобие поведения в моделируемой среде. Эта работа показывает интересный способ организации агента; она не превращает связный рассказ о прошлом в доказательство достоверной автобиографии или сознания [3].

В нашем проекте краткое резюме остаётся указателем на подробности. Важное решение можно проверить по исходным материалам. Для малозначительных событий допускается более грубое сжатие. Универсальной точности «на всё» нет: её стоимость должна соответствовать последствиям ошибки.

Доступ, согласие и срок хранения

Память о человеке затрагивает самого человека. Полезность персонализации не даёт системе безграничного права собирать всё подряд.

Нина предлагает простой интерфейс: «Что обо мне сохранено?», «Для чего это используется?», «Исправить», «Удалить». За этими словами должна стоять реальная работа с записями и зависимыми обобщениями, а не только дружелюбная кнопка.

Если часть информации уже использована для обучения параметров, задача удаления может отличаться от удаления строки в базе. Это различие нужно объяснять честно. Нельзя обещать полное исчезновение всех следов, проверив только одну таблицу.

Для NOVA команда начинается с ограниченного набора явно разрешённых предпочтений. Сведения о третьих лицах, случайно прозвучавшие в комнате, не превращаются автоматически в постоянный профиль.

Вечером робот снова встречает Нину у стола.

«Сегодня чай без сахара?» - спрашивает он.

«Сегодня - да».

В журнале появляется короткая запись о сегодняшнем напитке. Ни философии личности, ни вечного закона чаепития из неё больше не выводят.

Память стала немного менее самоуверенной. От этого она стала гораздо полезнее.

Практикум 2. Запишите фразу «Я не люблю работать вечером» так, чтобы сохранить её источник, дату и возможные ограничения. Что потребуется уточнить, прежде чем превращать её в постоянное правило расписания?

3. Кто проснётся после обновления

Лев установил новую версию системы. Утром NOVA говорил чуть быстрее, лучше различал предметы и неожиданно перестал узнавать шутку о чашке - ту самую, которую повторял весь прошлый день.

Мира проверила память: записи сохранились. Нина проверила настройки: они тоже были на месте. Однако ответ на знакомую реплику звучал иначе.

«Это всё ещё наш NOVA?» - спросила она.

Самир поднял глаза от таблицы.

«Сначала договоримся, для какой задачи нам нужен ответ».

Это прозвучало сухо, но оказалось началом очень человеческого разговора.

Несколько значений слова «тот же»

Для учёта оборудования «тот же» означает непрерывность объекта: корпус, серийный номер, история обслуживания. Для программного обеспечения важны версия, параметры и конфигурация. Для участника совместной работы существенны память о договорённостях, устойчивость поведения и понятность изменений.

Эти признаки могут расходиться. В старый корпус можно установить новую систему. Старые воспоминания можно перенести в новое тело. Можно сохранить голос и внешний вид, полностью изменив правила принятия решений.

Поэтому единственный красивый идентификатор не решает проблему. Номер устройства помогает вести инвентаризацию, но не сообщает пользователю, какие способности сохранились после обновления.

У людей вопрос личной идентичности значительно глубже. Он связан с телом, биографией, отношениями и субъективной перспективой. Мы не будем объявлять, что техническая схема автоматически разрешает философский спор. Вместо этого разберём практическую непрерывность искусственного агента - ту, за которую разработчики действительно способны отвечать.

Обещание переживает перезапуск

Предположим, NOVA договорился напомнить Мира о встрече и сохранить подготовленные заметки. Ночью произошёл перезапуск. Утром система должна знать, что обещание остаётся действующим, если оно не отменено и входит в её полномочия.

Такой вид непрерывности проверяется. Мы можем убедиться, что обязательство сохранилось, не продублировалось, не потеряло срок и не превратилось в новое поручение с другими условиями.

Однако сохранение обязательства не означает, что все прежние решения обязаны оставаться неизменными. Если обнаружилась ошибка, система должна её исправить и объяснить изменение. Верность прошлому полезна, пока она не становится верностью неисправности.

Мира предлагает считать устойчивыми прежде всего правила отношений: достоверно сообщать о возможностях, уважать договорённости, отмечать изменения и давать человеку возможность пересмотреть доступ.

Это не характер в полном человеческом смысле. Это проверяемый договор о поведении во времени. Для реальной совместной жизни такой договор порой важнее идеально узнаваемой интонации.

Две копии одного утра

Теперь мысленный эксперимент. Мы копируем состояние NOVA и запускаем две системы в разных учебных средах. В момент копирования их доступная память и настройки одинаковы. Затем одна работает с Мирой, а другая - с Ниной.

Через час у них появляются разные записи, ожидания и незавершённые задачи. Если обе продолжают называться одним именем и отправлять сообщения от одного аккаунта, путаница обеспечена.

Какую копию считать оригиналом? Для философии это открывает большой спор. Для системы учёта нам достаточно более узкого решения: общая история до момента разделения и отдельные ветви после него.

Каждая ветвь получает собственный идентификатор и область ответственности. Слияние памяти требует проверки происхождения событий. Запись «я видел чашку на столе» не должна без указания источника стать общим воспоминанием обеих копий.

Этот пример не доказывает, что у копий есть переживания. Он показывает, как быстро слова первого лица перестают быть технически однозначными.

Нина замечает, что даже самый обаятельный робот должен иногда говорить скучную фразу: «Это событие записано другой версией системы». Скучные фразы нередко спасают доверие.

Восстановление вчерашнего состояния

Резервная копия помогает после сбоя. Но возврат к ней создаёт разрыв: внешний мир уже изменился, а восстановленная система об этом не знает.

За потерянный промежуток человек мог отменить доступ, перенести встречу или исправить важную запись. Если робот просто продолжит со старого места, он начнёт действовать в

прошлом, тогда как чашка и её владелец находятся в настоящем.

Поэтому восстановление должно включать согласование с текущими источниками прав и задач. Старые разрешения нельзя автоматически считать действующими только потому, что они сохранились в архиве.

В этой главе мы не проектируем механизм доступа во всех деталях. Нам важно заметить общий принцип: идентичность агента связана с отношениями с окружающим миром. Одного внутреннего состояния недостаточно, чтобы восстановить всю ситуацию.

Даже человеческое «я вернулся» предполагает, что остальные участники знают, откуда и после какого перерыва.

Изменения, о которых стоит сообщать

Не каждое обновление требует торжественного собрания. Исправление опечатки может остаться незаметным. Изменение способов запоминания, распределения полномочий или оценки риска влияет на ожидания людей.

Для NOVA команда создаёт паспорт изменений. В нём указаны новая версия, изменённые способности, известные ограничения, совместимость с прежними записями и результаты повторной проверки важных сценариев.

Изменение	Что проверяем
Новый модуль памяти	Сохранились ли источники и исправления
Новая модель поведения	Выполняются ли прежние договорённости
Новое тело	Соответствуют ли навыки новым возможностям

Изменение	Что проверяем
Восстановление из копии	Учтены ли отменённые права и поручения

Человеку не нужно читать технический отчёт перед каждой чашкой чая. Но он должен иметь доступ к существенной информации в понятной форме.

Так техническая непрерывность превращается в социальную. Люди знают, с какой системой работают и каких изменений ожидать.

Узнавание без мистификации

К вечеру Лев нашёл причину исчезнувшей шутки. Новая настройка поиска отдавала предпочтение недавним записям, а старый эпизод перестал попадать в контекст. Его вернули в учебную подборку для проверки.

NOVA снова узнал фразу. Команда улыбнулась, но никто не объявил, что найдено местонахождение личности.

И всё же улыбка была настоящей. Общая история имеет значение для людей даже тогда, когда мы понимаем устройство её хранения. Памятная фотография тоже состоит из физических следов, но это не делает человеческое отношение к ней бессмысленным.

Мы вправе ценить устойчивое взаимодействие с технологией. Нам полезно одновременно различать собственное чувство узнавания и свойства, которые действительно установлены у системы.

Нина закрыла блокнот:

«Пусть он меняется. Главное, чтобы мы понимали, что меняется вместе с ним».

Это оказался вполне достойный ответ для первого дня после обновления.

Практикум 3. Представьте две копии помощника с общей памятью до полудня. Одна получила отмену поручения, другая - нет. Какие сведения нужно согласовать, прежде чем разрешить им снова действовать от одного имени?

4. Искусство забывать вовремя

За неделю NOVA научился различать новые чашки. Команда радовалась до тех пор, пока робот не перестал уверенно распознавать старую белую.

«Он обиделся на неё за прошлую книгу», - предположила Мира.

Лев назвал другую причину: последнее обучение изменило модель так, что часть прежних различий стала распознаваться хуже.

Самир предложил переименовать отчёт «Система стала умнее» в «Система изменилась». До выяснения подробностей это было точнее.

Возможность учиться не гарантирует, что каждое обучение полезно во всех отношениях. Новое знание может конкурировать со старым. А иногда старое действительно следует забыть.

Стабильность и пластичность

Обучаемой системе необходимо одновременно сохранять полезный опыт и реагировать на изменения. Слишком жёсткая модель не приспособливается. Слишком податливая может менять поведение от каждого нового примера.

Представьте сотрудника, который после одного разговора полностью пересматривает все рабочие правила. Сегодня чашки хранят справа, завтра - слева, послезавтра он больше не признаёт существование шкафов. Гибкость есть. Работать вместе трудно.

Обратная крайность - сотрудник, который однажды выучил расположение комнаты и отказывается признавать

перестановку. Устойчивость тоже присутствует, но мебель уже возражает.

В машинном обучении заметная потеря прежних навыков при освоении новых задач называется катастрофическим забыванием. В работе Киркпатрика и коллег предложен подход EWC: изменения важных для старых задач параметров дополнительно ограничиваются. Это один из способов уменьшить забывание в определённых условиях; он не является универсальным решением любого непрерывного обучения [4].

Для NOVA вопрос формулируется конкретнее: какие старые способности должны сохраниться после освоения новой задачи и как мы это проверим?

Повторение требует отбора

Один из общих подходов - возвращать в обучение примеры прежних задач. Но «повторять всё» быстро становится слишком дорогим. Придётся выбирать, что представляет старый опыт достаточно хорошо.

Если в сохранённой подборке только удобные, ярко освещённые чашки, система может поддерживать успех на них и всё хуже работать с редкими трудными случаями. Формально память о прошлом есть; практически из неё исчезло то, что делало прошлое сложным.

Самир предлагает хранить не только типичные примеры, но и сведения об условиях ошибки. Мира добавляет случаи с новыми материалами. Нина просит включить изменения намерения человека: робот должен помнить не только форму предмета, но и возможность отмены поручения.

Состав повторяемой подборки становится исследовательским решением. Его нельзя спрятать за нейтральным словом «данные». От выбора примеров зависит, какие части мира система продолжит учитывать.

Отдельно сохраняют набор проверок, который не используется для настройки. Иначе каждое улучшение будет сопровождаться перестановкой экзаменационных вопросов так, чтобы ученику было удобнее.

Забывание не всегда поломка

Теперь другой случай. Раньше синяя полка служила местом для чашек, а после перестановки там хранят оборудование. Старое правило нужно перестать применять.

Удалить его без следа не всегда лучший выход. Для понимания прошлых действий полезно знать, что оно действовало раньше. Поэтому в проектной памяти различают историю и текущую норму.

Запись может остаться в архиве с указанием периода действия. При выборе сегодняшнего действия система должна использовать новое правило. При объяснении вчерашнего - иметь доступ к прежнему.

Существуют также сведения, которые не следует хранить дольше необходимого. Случайный личный разговор, временный код доступа, промежуточное наблюдение о посетителе не обязаны становиться вечной частью биографии лаборатории.

Здесь забывание - предусмотренная функция. Её качество измеряется тем, что нужная информация остаётся доступной, а ненужная перестаёт влиять на решения и действительно удаляется там, где это было обещано.

Ошибка, повторённая много раз

В журнале может возникнуть петля. NOVA сделал неверное предположение, сохранил свой ответ, позднее нашёл его и использовал как подтверждение. Затем новый ответ снова попал в память. Через неделю ошибочная идея имеет десять «источников», каждый из которых произошёл от одного промаха.

Количество записей создаёт впечатление надёжности, хотя независимых свидетельств не прибавилось.

Чтобы избежать такой путаницы, происхождение записей должно сохраняться при обобщении и повторном использовании. Собственный ответ системы нельзя автоматически считать независимым наблюдением.

Человек сталкивается с похожей проблемой, когда видит одну и ту же новость на десяти сайтах и принимает перепечатки за десять подтверждений. Сходство полезно для объяснения, но механизмы человеческой и машинной памяти при этом не обязаны совпадать.

В NOVA вывод о предпочтении Нины больше не может укрепляться только потому, что робот сам несколько раз его повторил. Нужны новые релевантные данные или явное подтверждение.

Когда учиться, а когда отложить обновление

Не всякая встреча с новым примером должна немедленно менять действующую систему. Иногда разумнее сохранить наблюдение, проверить его качество и включить в следующую управляемую процедуру обучения.

Для нашей лаборатории вводятся три режима. В рабочем режиме NOVA использует проверенную версию. В режиме сбора опыта он фиксирует разрешённые события и ошибки. В исследовательском режиме команда обучает кандидатную версию и сравнивает её с прежней.

Это проектный выбор, а не запрет на все формы обучения во время работы. В некоторых задачах адаптация на месте необходима. Но тогда заранее определяются изменяемые части, допустимые границы и способы обнаружить ухудшение.

Важна обратимость. Если новое обучение испортило знакомый навык, должна существовать возможность

остановить распространение изменений, восстановить рабочую версию и изучить причину.

Мы называем это взрослением системы, хотя технически речь идёт о процедуре управления изменениями. Метафора помогает, пока мы не принимаем её за доказательство человеческой психологии у машины.

Чему научилась команда

После повторной проверки белая чашка вернулась в число уверенно распознаваемых объектов. Улучшение на новых формах сохранилось не полностью; пришлось признать компромисс и продолжить работу.

В блокноте появились две колонки: «Что стало лучше» и «Что стало хуже». Самир настоял на третьей: «Что мы пока не проверили».

Лев сначала вздохнул, потом согласился. Третья колонка оказалась длиннее первых двух, зато впервые показывала реальное направление следующего опыта.

Хорошая обучаемая система не обязана накапливать всё бесконечно. Ей нужно сохранять полезные различия, обновлять устаревшие правила и не превращать собственные ошибки в традицию.

На этот раз чашка ничем не помогла. Она просто стояла на новой полке. Иногда достойная роль предмета в научном прогрессе заключается именно в этом.

Практикум 4. После обучения на новой задаче точность помощника выросла с 60% до 85%, а на старой упала с 92% до 71%. Какие ещё сведения нужны, прежде чем назвать обновление улучшением? Составьте короткий план сравнения.

ЧАСТЬ II. МИР И ПРЕДСТАВЛЕНИЕ О СЕБЕ

5. Будущее, которое можно попробовать заранее

Мира положила блокнот на край стола и попросила NOVA передвинуть поднос. Робот замер.

«Если сдвинуть поднос прямо, он заденет блокнот. Предлагаю сначала изменить направление».

На экране появились два прогнозируемых пути. Ни один ещё не был выполнен. Команда смотрела на возможное будущее, которое пока существовало только внутри вычислительной модели.

«Теперь главное, - сказал Самир, - не перепутать эту картинку с обещанием».

В этом замечании помещалась почти вся трудность машинного воображения.

Карта не обязана быть похожа на местность

Модель мира - представление, с помощью которого система оценивает состояние среды и возможные последствия действий. Она может содержать геометрию, связи между объектами, вероятности событий, скрытые числовые признаки или сочетание разных форм.

Для полезного прогноза не всегда требуется фотореалистичный внутренний фильм. Навигатор помогает выбрать маршрут с помощью линий и времени в пути. Шахматная программа оценивает позиции, не рисуя себе лак на фигурах.

Но слишком бедное представление теряет существенное. Если модель знает только положение чашки и игнорирует её содержимое, резкое движение может показаться таким же

хорошим, как плавное. Если она не учитывает человека рядом, свободный путь на карте окажется занятым в реальности.

Хорошая модель сохраняет различия, важные для задачи. Не существует одного гарантированно достаточного уровня подробности для всех возможных поручений.

Поэтому мы спрашиваем: какие последствия нужно предсказывать, на каком промежутке времени и с какой точностью?

Учиться на воображаемых продолжениях

В работе о Dreamer третьего поколения Хафнер и коллеги описывают обучение поведения с использованием выученной модели среды и воображаемых будущих траекторий. Авторы показывают результаты на разнообразных задачах управления. Это сильный пример исследовательского подхода, но успех на наборе сред не означает, что такая система без дополнительной проверки надёжно предсказывает любую человеческую ситуацию [5].

Идея привлекательна: физическая попытка стоит времени, энергии и иногда разбитой чашки, тогда как вычислительное сравнение вариантов может быть дешевле. Модель позволяет отбрасывать часть плохих планов до движения.

Однако появляется новая зависимость. Агент учится в мире, который сам приблизительно описал. Ошибка описания становится ошибкой учебной среды.

Если воображаемая поверхность не скользит, система может выработать движения, которые прекрасно работают только в её представлении. Чем настойчивее оптимизация, тем лучше она иногда использует слабое место модели.

Мира называет это тренировкой для экзамена, который робот сам себе написал и уже случайно испортил.

Один шаг и целая история

Предсказать положение чашки через долю секунды обычно проще, чем состояние всей лаборатории через час. По мере увеличения горизонта появляются новые действия, внешние события и неизвестные решения людей.

Даже небольшая ошибка на раннем шаге может изменить последующие условия. Мы не обязаны ожидать ровного линейного роста ошибки: зависимость определяется задачей и моделью.

Практический ответ - регулярно возвращаться к наблюдениям. NOVA строит короткий план, выполняет допустимую часть, снова оценивает обстановку и корректирует продолжение.

Такой подход не отменяет долгосрочной цели. Он разделяет общую направленность и ближайшее движение. Человек тоже может намереваться написать книгу, не зная точного положения каждой запятой в последней главе.

В нашей сцене цель - освободить место для блокнота. Ближайшее действие - начать безопасное перемещение подноса. После изменения положения нужно проверить, остались ли первоначальные предположения верны.

Несколько правдоподобных будущих

Часто проблема не в неточности одного прогноза, а в наличии нескольких возможных исходов. Человек может убрать руку или оставить её на месте. Чашка может оказаться пустой или наполненной. Дверь может быть приоткрыта сильнее, чем видно с текущего ракурса.

Полезная система сохраняет эту неоднозначность. Она может оценить несколько сценариев, выбрать действие, приемлемое в каждом из них, или получить дополнительное наблюдение.

Здесь появляется цена информации. Повернуть камеру иногда выгоднее, чем долго вычислять движение по

неполному изображению. Уточнить намерение человека иногда полезнее, чем строить ещё сто вариантов его поведения.

Неизвестность	Возможное действие
Не видно край подноса	Изменить точку наблюдения
Неясна цель человека	Задать короткий вопрос
Неизвестно содержимое чашки	Выбрать более осторожный режим или уточнить
Прогнозы заметно расходятся	Сократить действие и проверить состояние

Это примеры проектного рассуждения. Реальные допустимые движения определяются конструкцией и проверками конкретного устройства.

Красивое объяснение не заменяет проверку

После ошибки NOVA может построить связную историю: «Я предполагал, что поверхность сухая». Но насколько точно это описывает механизм, который действительно повлиял на действие?

Для проверки нужны сохранённые прогнозы до события. Если объяснение появляется только после неудачи, трудно отличить восстановление причины от правдоподобного рассказа.

Поэтому команда записывает исходное наблюдение, выбранное представление, прогнозируемый результат и реальный исход. Не все внутренние вычисления становятся понятными от одного журнала, но появляется материал для сравнения.

Мы можем спросить, предсказывает ли модель ошибки на новых примерах, чувствительна ли к важным изменениям и сохраняет ли уверенность там, где её точность падает.

Хороший прогноз должен рисковать оказаться неверным. Если объяснение совместимо с любым исходом, оно не помогает отличать удачный план от неудачного.

Воображение с открытым окном

В литературной сцене NOVA выбирает обходной путь. Поднос плавно смещается, блокнот остаётся на месте. Мира кивает.

Самир просит повторить опыт с другим расположением предметов. Теперь робот задаёт уточняющий вопрос. Воображаемый план оказался недостаточно определённым.

Это не шаг назад. Система обнаружила, что для действия не хватает сведений, и обратилась к реальности.

Мы часто связываем интеллект со способностью далеко уходить в абстракции. Но для физического агента столь же важна способность возвращаться: снова посмотреть, измерить, проверить, услышать, что человек передумал.

Воображение полезно, когда остаётся открытым для исправления. Иначе даже очень сложная модель мира постепенно превращается в удобную комнату, из которой система перестаёт выходить.

NOVA закончил движение. На столе появилось место для блокнота. На новой странице Мира написала: «Прогноз - это предложение будущему. Ответ всё равно приходит снаружи».

Практикум 5. Помощник прогнозирует, что встреча займёт час. Назовите три скрытых условия этого прогноза. Как построить план, который останется полезным, если одно из них окажется неверным?

6. Почему синяя чашка не виновата

NOVA обнаружил закономерность: при переносе синих чашек жидкость проливалась реже. Он предложил заменить всю лабораторную посуду.

Мира, которая только что согласовала бюджет на датчики, заметно оживилась. Такой способ повышения надёжности выглядел удивительно дешёвым.

Самир попросил показать распределение опытов.

Оказалось, что синие чашки чаще использовались в испытаниях с небольшим количеством воды, а красные - в более трудных опытах с высоким уровнем налива. Цвет отлично предсказывал результат в журнале. Причиной результата он от этого не становился.

В тот день лаборатория сэкономила на закупке посуды и потратилась на понимание.

Наблюдение отвечает на свой вопрос

Статистическая связь полезна. Если определённый признак регулярно сопровождает ошибку, его стоит исследовать. Проблема начинается, когда мы без дополнительных оснований превращаем связь в обещание эффекта от вмешательства.

Фраза «Синие чашки в наших записях проливаются реже» описывает наблюдения. Фраза «Если перекрасить чашку в синий, перенос станет надёжнее» описывает предполагаемую причину. Для второго вывода нужны дополнительные знания или проверка.

Причинный подход различает наблюдение, действие, изменяющее условия, и рассуждение о том, что произошло бы при другом действии. В изложении Перла важную роль

играют явные причинные предположения: одних числовых связей недостаточно, чтобы автоматически ответить на все вопросы о вмешательствах [6].

Для читателя полезно удерживать простую мысль: хороший предсказатель не обязательно является хорошим рычагом управления.

Утренний звон будильника связан с началом рабочего дня. Покупка более громкого будильника не добавляет вам профессиональной квалификации. Хотя соседям может показаться, что вы начали работать начальником.

Числа, которые рассказывают две истории

Рассмотрим специально придуманный учебный пример. Это не результаты испытаний NOVA.

Условия	Красная чашка	Синяя чашка
Мало воды	18 успехов из 20	72 успеха из 80
Много воды	40 успехов из 80	10 успехов из 20
Все попытки	58 из 100	82 из 100

В общих итогах синяя чашка выглядит намного лучше: 82% против 58%. Но при малом количестве воды у обоих цветов по 90% успеха, а при большом - по 50%.

Разница в итогах возникла из-за состава опытов. Синие чашки чаще попадали в лёгкие условия, красные - в трудные. Наблюдаемая связь цвета с успехом здесь не показывает самостоятельного преимущества цвета.

Этот пример не доказывает, что цвет никогда не влияет на роботизированный перенос. Он может влиять на работу конкретной системы зрения. Пример показывает другое:

представленных итоговых чисел недостаточно для вывода о таком влиянии.

Нужна проверка, которая разделит возможные объяснения.

Что значит изменить одно условие

Команда планирует новый опыт. Цвет, уровень налива, освещение и порядок попыток распределяют так, чтобы влияние одного фактора не сливалось с другим. Форму и массу чашек тоже учитывают.

Если сравнить синюю керамическую кружку с красным пластиковым стаканом, мы одновременно изменим слишком многое. Даже обнаружив различие, будет трудно понять, к чему его отнести.

В контролируемой проверке полезно заранее определить результат: например, сохранение содержимого в пределах принятого допуска и корректное завершение переноса. Менять определение успеха после просмотра видео - плохой способ улучшать понимание.

Нельзя также требовать, чтобы реальная среда всегда была идеально изолирована. Иногда контролировать все условия невозможно. Тогда вывод ограничивают доступными данными, а предположения явно перечисляют.

Вопрос «Что ещё могло измениться?» часто важнее вопроса «Достаточно ли убедительно выглядит график?».

Контрфактическая пауза

Представим, что чашка всё же пролилась. Мы спрашиваем: произошло бы это, если бы движение было медленнее?

Фактически наблюдалась только одна траектория. Другую приходится оценивать с помощью модели и дополнительных сведений. Поэтому контрфактический ответ зависит от качества модели.

Если причина была в неверном захвате, небольшое снижение скорости могло не помочь. Если поверхность наклонилась из-за контакта с подносом, нужно изменить геометрию движения. Если человеку стало нужно другое место назначения, технически идеальный перенос вообще мог оказаться не тем действием.

NOVA должен уметь предлагать несколько объяснений и различать степень их поддержки. Одна уверенная версия не становится верной только потому, что легко помещается в отчёт.

Полезный разбор ошибки заканчивается проверяемым следующим шагом. «Больше стараться» - эмоционально понятная рекомендация, но для модели управления довольно туманная.

Причины человеческих поступков

С людьми причинность становится ещё сложнее. Нина улыбается после удачного переноса, но улыбка может относиться к сообщению на телефоне. Мира торопит робота, потому что опаздывает, а не потому что всегда предпочитает максимальную скорость.

Агенту не следует превращать каждое совпадение в постоянное объяснение личности. Поведение зависит от контекста, и многие важные обстоятельства не наблюдаются.

Поэтому в социальных задачах особое значение имеют уточнение и осторожность формулировок. «Возможно, сейчас вам важна скорость» оставляет место для исправления. «Вы человек, который всегда торопится» создаёт гораздо более сильное и плохо обоснованное утверждение.

Это относится и к людям, проектирующим NOVA. По одной неудачной реакции нельзя делать вывод обо всей системе, а по одной красивой сцене - обо всех её будущих возможностях.

Вернуться к предмету

В конце дня закупку синих чашек отменяют. Команда сохраняет саму находку: цвет оказался полезным признаком различия между наборами испытаний. Ошибка NOVA помогла обнаружить проблему в организации данных.

Так иногда и происходит исследование. Неверное объяснение, если его аккуратно проверить, открывает вполне реальную структуру.

Мира берёт красную чашку и наливает в неё немного воды.

«Дадим ей справедливый шанс».

Робот выполняет перенос. Самир заносит результат в новый журнал, где условия больше не прячутся за общим процентом.

Синяя чашка остаётся синей. Красная - красной. Между ними больше не лежит необоснованная теория. Иногда научный прогресс выглядит как возвращение обычным вещам права быть обычными.

Практикум 6. В учебной таблице синие чашки успешнее в общей сумме. Объясните, почему это не доказывает пользу синего цвета. Какие два условия вы обязательно уравнили бы в следующем опыте?

7. Модель себя без волшебного зеркала

После замены захвата NOVA потянулся к привычной чашке и остановился в нескольких сантиметрах от ручки.

«Ожидаемое положение пальцев не соответствует измеренному».

Мира улыбнулась. Вчера такая фраза показалась бы неудачей. Сегодня она означала, что система заметила расхождение до контакта.

Лев предложил написать в отчёте: «Робот осознал изменение собственного тела».

Самир посмотрел на него.

«Может, пока напишем, что он обнаружил ошибку модели захвата?»

Лев согласился. Заголовок стал короче по смыслу и длиннее по количеству полезной информации.

Что системе нужно знать о себе

Чтобы действовать, агенту требуется представление не только о внешней среде. Важно учитывать собственные размеры, доступные движения, состояние датчиков, запас энергии, действующие задачи и ограничения.

Такую совокупность представлений можно назвать моделью себя в функциональном смысле. Она помогает предсказывать последствия собственных действий и выбирать доступный способ выполнения поручения.

Термометр, показывающий температуру корпуса, ещё не является полной моделью себя. Но его данные могут быть частью системы, которая оценивает рабочее состояние и меняет поведение.

С другой стороны, подробное словесное описание «какой я робот» может почти не влиять на управление. Если NOVA уверенно перечисляет свои размеры, но постоянно задевает дверной проём, рассказ и действующая модель расходятся.

Поэтому проверять нужно связь представления с поведением. Что меняется, когда система обнаруживает собственное ограничение?

Тело, которое приходится уточнять

Даже известная конструкция не остаётся совершенно неизменной. Износ, замена деталей, смещение датчика, нагрузка и температура могут влиять на действия. Описание из проектного файла не всегда полностью совпадает с текущим состоянием.

В исследовании Бонгарда, Зыкова и Липсона робот использовал взаимодействие со средой для построения модели собственного устройства и адаптации поведения после изменения тела. Эта работа важна как пример функционального само моделирования. Само по себе наличие такой модели не устанавливает субъективного самосознания [7].

Для NOVA мы задаём более узкую учебную ситуацию: после замены захвата сравнить ожидаемые и измеренные положения, обнаружить расхождение и пройти предусмотренную процедуру проверки. Здесь нет необходимости объявлять, что робот пережил кризис идентичности.

Инженерная ценность состоит в том, что ошибка становится заметной до того, как превратится в повреждение предмета.

Знать способность и знать её границу

Фраза «Я умею переносить чашки» слишком общая. Какая масса? Какое положение ручки? Какая видимость? Какой путь? Нужен ли предварительно согласованный сценарий?

Модель собственных возможностей должна учитывать условия. Один и тот же навык может быть хорошо проверен в знакомой обстановке и плохо изучен в новой.

Это похоже на человеческую компетентность. Умение плавать в бассейне не означает готовность к любому морю. Полезный самоотчёт связывает способность с областью опыта.

В проекте NOVA у навыка есть описание входных условий, ожидаемый результат и основания для отказа или запроса помощи. Отдельно указываются признаки, при которых нужно повторно оценить ситуацию.

Так система получает язык для конкретного незнания. Она может сообщить: «Этот объект выходит за проверенный диапазон», вместо того чтобы произносить общее «Я не могу» или уверенное «Конечно».

Могу, разрешено, стоит ли

Три вопроса часто смешивают. Физически доступное действие не обязательно разрешено. Разрешённое действие не обязательно полезно в текущей ситуации.

NOVA может открыть дверцу шкафа. Но если там находятся материалы другой группы, доступ определяется договорённостью. Даже при наличии доступа открывать шкаф бессмысленно, если нужная чашка уже на столе.

Вопрос	О чём он
Могу ли?	Способности и текущее состояние
Разрешено ли?	Полномочия и условия доступа
Нужно ли сейчас?	Цель, последствия и альтернативы

Хорошее управление связывает эти проверки, не подменяя одну другой. Уверенность в собственном навыке не создаёт полномочий.

Это особенно важно для агента, способного предлагать новые действия. Инициатива может проявляться в выборе полезного вопроса, подготовке плана или обнаружении проблемы. Для этого не требуется автоматически давать системе право менять весь окружающий мир.

Откуда взялся самоотчёт

Когда NOVA говорит «Я не уверен», полезно знать, на чём основана фраза. Возможно, она связана с расхождением прогнозов. Возможно, с недостатком данных. А возможно, языковая модель просто сочла осторожный ответ стилистически уместным.

Эти варианты нельзя различить по одной интонации. Поэтому самоотчёт проверяют относительно доступных измерений и последующих результатов.

Если система часто заявляет высокую уверенность перед ошибкой, проблема остаётся, каким бы убедительным ни был голос. Если корректно выделяет трудные случаи, это полезная способность, которую можно измерять.

При этом журнал внутренних сигналов не делает весь механизм прозрачным. Между доступными показателями и окончательным действием могут оставаться сложные зависимости. Утверждение «мы записали всё» почти всегда требует уточнения, что именно входило в это «всё».

Кто управляет телом

В некоторых ситуациях важно различать действие самого агента и внешнее изменение. Рука сдвинулась потому, что поступила команда, потому что изменилось сопротивление среды или потому что человек аккуратно направил её в предусмотренном режиме?

Наблюдаемый результат может быть похожим. Для обучения причина существенна: иначе робот припишет себе достижение, которое обеспечил другой участник.

Модель себя должна учитывать происхождение управляющих воздействий и условия наблюдения. Это помогает корректно оценивать собственный вклад и не учиться на неверных объяснениях успеха.

Даже в нашей вымышленной лаборатории Нина просит помечать подсказки людей. Самостоятельное выполнение и выполнение с помощью - разные результаты, хотя чашка в обоих случаях оказывается на подносе.

Возвращение к руке

Проверка заканчивается. NOVA использует обновлённое описание захвата и снова переносит чашку. На этот раз ожидаемое движение лучше согласуется с измеренным.

Мира записывает: «Система заметила изменение, ограничила действие, прошла проверку и восстановила навык».

Такой результат не решает вопрос о внутреннем «я». Он решает вполне реальную задачу совместной работы.

Мы можем исследовать сознание, не объявляя каждую диагностику его доказательством. И можем ценить функциональное само моделирование, не требуя от него немедленно ответить на все вопросы философии.

NOVA смотрит на новую руку. Мира смотрит на показания. У каждого в этой сцене своя роль. Нам пока достаточно, что чашка остаётся целой, а границы объяснения - видимыми.

Практикум 7. Помощник говорит: «Я могу выполнить задачу». Составьте три уточнения: одно о способности, одно о разрешении, одно о целесообразности. Почему одного ответа «могу» недостаточно?

8. Полезное «я не знаю»

«Вероятность успеха - девяносто пять процентов», - сообщил NOVA.

Мира уже собиралась разрешить действие, когда Самир спросил:

«А что происходило в прошлых случаях, когда ты говорил девяносто пять?»

Ответ занял несколько секунд. Оказалось, что уверенное число не всегда соответствовало частоте успеха. Иногда оно было характеристикой внутреннего расчёта, которую команда слишком поспешно перевела на язык вероятностей реального результата.

Голос робота при этом звучал одинаково убедительно. Уверенная интонация, как выяснилось, вообще была самой дешёвой частью системы.

Уверенность должна что-то означать

Представим множество сопоставимых случаев, в которых система оценивает вероятность успеха как 80%. При хорошей калибровке доля успехов в такой группе будет близка к 80%, с учётом случайного разброса.

Это свойство проверяется по серии прогнозов и исходов. Одно событие не позволяет установить калибровку: прогноз с вероятностью 80% допускает неудачу.

Го и коллеги изучали калибровку современных нейронных сетей и способы её улучшения, включая температурное масштабирование. Работа показывает, что высокая точность классификации сама по себе не гарантирует корректности выдаваемой уверенности. Конкретный метод калибровки оценивается на данных и не даёт автоматической гарантии для любой новой среды [8].

Для NOVA важно связать число с чётким событием. «Уверенность в чашке» непонятна. «Вероятность правильно определить выбранную категорию на таких изображениях» уже допускает проверку. «Вероятность полностью выполнить перенос» - другой прогноз, охватывающий больше этапов.

Воображаемый график с реальным уроком

На рисунке показан учебный набор: в каждой группе по сто условных случаев. Это специально составленные числа, а не результаты NOVA.

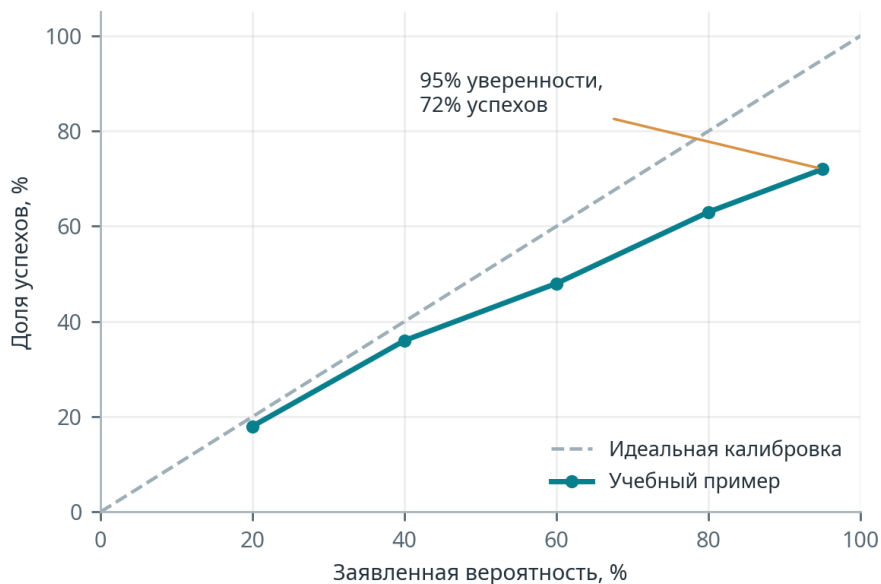


Рис. 2. Иллюстрация из пяти условных групп по 100 случаев. На диагонали заявленная вероятность совпадает с долей успехов. Числа придуманы для объяснения и не являются измерениями NOVA.

Если система выдаёт 95%, а успех происходит в 72 случаях из ста, она заметно переоценивает надёжность в этой группе. Красивый средний показатель может скрывать такую проблему на отдельных участках.

Однако идеальная калибровка ещё не означает хорошего различения случаев. Система, которая всегда сообщает общую частоту успеха, иногда бывает калиброванной, но почти не помогает выбирать между простыми и трудными задачами.

Нам нужны оба свойства: полезное ранжирование и осмысленные вероятностные оценки. Их также нужно рассматривать вместе с ценой ошибок.

Самир подчёркивает ещё одну деталь: группы не должны становиться настолько маленькими, что случайные колебания начинают выглядеть как закономерности. График не освобождает от размышления о количестве наблюдений.

Как оценить вероятностный прогноз

Для двоичного события удобно рассмотреть оценку Брайера. Пусть p - предсказанная вероятность, а y равно единице при успехе и нулю при неудаче. Для каждого случая вычисляют квадрат разности, затем берут среднее.

$$\text{Brier} = \text{среднее значение } (p - y)^2$$

Чем меньше результат, тем лучше вероятностные прогнозы на данном наборе. Эта оценка учитывает, насколько уверенно система ошиблась, а не только совпал ли её ответ с исходом.

Возьмём четыре учебных прогноза: 0,9; 0,8; 0,6; 0,3. Пусть исходы равны 1; 0; 1; 0. Квадраты ошибок составят 0,01; 0,64; 0,16; 0,09. Сумма равна 0,90, среднее - 0,225.

Постоянный прогноз 0,5 дал бы для этих четырёх случаев 0,25. Первый набор немного лучше по этой мере, хотя один его уверенный промах дорого обошёлся итоговой оценке.

Четырёх наблюдений недостаточно для серьёзного вывода о системе. Пример нужен, чтобы показать устройство вычисления. Оценку также нельзя бездумно сравнивать между задачами с разными условиями и составом данных.

Когда разумнее спросить

Вероятность становится полезной для действия, когда связана с последствиями.

Предположим в условной модели, что неудача стоит сто единиц потерь, успех не создаёт потерь, а уточнение у человека стоит три единицы и полностью устраняет рассматриваемую неопределённость. Тогда ожидаемая потеря самостоятельного действия равна $100 \times (1 - p)$.

При $p = 0,9$ она составляет десять единиц, и уточнение выгоднее. При $p = 0,99$ - одну, и самостоятельное действие выгоднее. При $p = 0,97$ варианты равны в этой упрощённой модели.

В реальной ситуации уточнение тоже может ошибаться, задерживать работу или не решать проблему полностью. Числовые стоимости могут быть спорными. Поэтому формула служит способом сделать предположения видимыми, а не универсальным разрешением действовать.

Тем не менее она помогает понять, почему один порог уверенности не подходит для всех задач. Ошибка в выборе цвета папки и ошибка в движении рядом с человеком имеют разные последствия.

Незнание бывает разным

Иногда не хватает наблюдений. Тогда полезно посмотреть ещё раз или запросить информацию.

Иногда модель недостаточно обучена. Тогда дополнительная фотография не обязательно исправит проблему.

Иногда сама ситуация содержит неопределённость: например, человек ещё не решил, какой вариант предпочитает. Ни один датчик не обязан узнать решение, которого пока нет.

А иногда система попадает в условия, существенно отличающиеся от проверенных. Прежняя калибровка может

перестать работать. Высокая уверенность внутри знакомой процедуры не доказывает её применимость к новой.

Полезное «не знаю» указывает тип пробела и возможный следующий шаг. «Не вижу ручку чашки, могу изменить ракурс» помогает больше, чем общее «Недостаточно данных».

Но осторожность тоже нуждается в проверке. Система, которая отказывается от всего, может избегать ошибок ценой полной бесполезности. Поэтому вместе с точностью учитывают долю выполненных задач, число запросов помощи и последствия задержек.

Интонация после числа

Нина предлагает изменить ответы NOVA. Уверенность больше не сообщается как декоративная цифра без пояснения. В важных случаях система кратко указывает, что оценивалось и какое ограничение остаётся.

Мира замечает, что работать стало спокойнее. Не потому, что ошибок больше нет. Теперь у команды появился способ обсуждать основания решения до результата.

«Вероятность успеха в этих условиях недостаточно проверена, - говорит NOVA. - Предлагаю сначала уточнить положение ручки».

Чашка ждёт ещё несколько секунд. Это небольшая задержка. Зато слово «предлагаю» наконец связано с понятной причиной.

Разумная уверенность не обязана звучать громко. Ей достаточно выдерживать встречу с данными.

Практикум 8. Рассчитайте оценку Брайера для прогнозов 0,9; 0,8; 0,6; 0,3 и исходов 1; 0; 1; 0. Затем объясните, почему даже хорошая оценка на знакомых задачах не гарантирует успеха в новой обстановке.

ЧАСТЬ III. ЛЮБОПЫТСТВО И СОЗИДАНИЕ

9. Любопытство и телевизор, который всегда удивляет

В лаборатории появился экран со случайно меняющимися узорами. Его включили для проверки камеры и забыли выключить.

Через некоторое время NOVA стал уделять экрану подозрительно много внимания. Новые кадры плохо предсказывались, а исследовательский модуль получал сигнал, что здесь происходит нечто неожиданное.

«Он нашёл сериал», - сказала Нина.

«Сериал без сюжета», - уточнил Лев.

Самир посмотрел на журнал:

«Зато с бесконечным бюджетом неожиданности».

Так команда познакомилась с различием между удивлением и обучением.

Двигатель исследования

Когда внешняя награда редка, системе трудно понять, какие действия стоит попробовать. Внутренний сигнал может направлять исследование: например, поощрять посещение новых состояний или уменьшение неопределённости.

В работе Патака и коллег любопытство формализовано через ошибку предсказания последствий действия в специально выученном пространстве признаков. Подход исследовался в игровых средах и помогал освоению поведения при ограниченной внешней награде. Здесь слово «любопытство»

обозначает устройство сигнала обучения; оно не устанавливает наличие человеческого чувства интереса [9].

Такая оговорка не уменьшает инженерной идеи. Механизм может быть полезным, даже если его психологическое название используется как аналогия.

NOVA не должен ждать подробной команды для каждого полезного наблюдения. Но направлять самостоятельное исследование одним требованием «ищи неожиданное» недостаточно.

Почему шум соблазнителен

Случайный поток может оставаться плохо предсказуемым сколько угодно долго. Если награда связана только с величиной ошибки, система получает стимул возвращаться туда, где ошибка не уменьшается.

В исследовании Бурды и коллег о масштабном обучении на основе любопытства отдельно рассматриваются ограничения награды за предсказание в стохастических условиях. Это полезное предупреждение: высокий сигнал новизны может сопровождать среду, в которой агент почти не приобретает переносимого знания [10].

В нашей сцене экран постоянно удивляет NOVA, но не помогает переносить чашки, понимать просьбы или планировать действия.

Человек легко узнаёт аналогию с бесконечной лентой новых сообщений. Однако сходство поведения не означает тождества механизмов. Для проектирования робота достаточно заметить проблему цели: неожиданность сама по себе не совпадает с полезностью изучения.

Смотреть на изменение ошибки

Один из разумных исследовательских ориентиров - следить за прогрессом, а не только за текущей трудностью. Если после опыта прогнозы улучшаются, система действительно что-то осваивает.

Слишком простая задача почти не даёт нового. Слишком непредсказуемая может поглощать время без устойчивого улучшения. Между ними находятся задачи, в которых усилие превращается в приобретаемую способность.

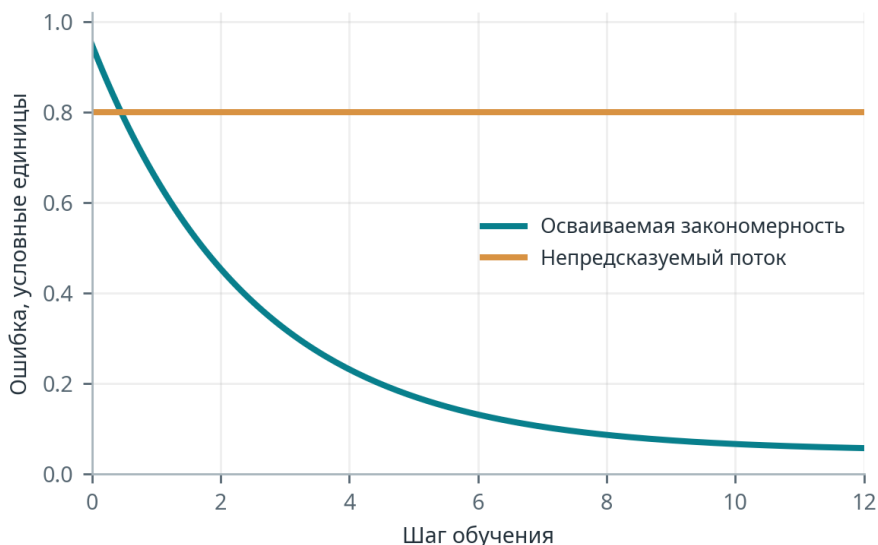


Рис. 3. Схематические кривые: ошибку можно уменьшать при освоении закономерности, тогда как непредсказуемость случайного потока сохраняется. Это учебная модель, а не результаты эксперимента.

На рисунке изображены условные кривые, построенные для объяснения идеи. Они не описывают конкретный алгоритм и не являются экспериментальными данными.

Но и прогресс можно измерить плохо. Если система оценивает себя на тех же примерах, на которых училась, уменьшение ошибки может отражать запоминание. Если

изменяет собственную шкалу, улучшение окажется косметическим.

Поэтому нам нужны новые проверочные случаи, устойчивые критерии и учёт стоимости исследования. Любопытство получает значение в связи с тем, что агент сможет делать после него.

Как выбирать следующий вопрос

NOVA видит два незнакомых предмета. Один - новая чашка с необычной ручкой. Другой - случайно мигающая декоративная лампа.

При выборе исследования полезно учитывать несколько вещей: сколько неопределённости можно уменьшить, связано ли это с будущими задачами, во сколько обойдётся опыт и разрешены ли необходимые действия.

Мы можем представить такую оценку как сравнение ожидаемой пользы знания и стоимости его получения. Точные численные веса зависят от проекта; они не возникают из философского определения любопытства.

Важно также сохранять разнообразие. Если система исследует только то, что немедленно приносит пользу в знакомой задаче, она может никогда не обнаружить новую область возможностей.

Команда поэтому оставляет небольшой отдельный бюджет для безопасных исследовательских задач с отложенной пользой. Это решение людей о направлении работы, а не врождённое право машины на любые эксперименты.

Вопросы с последствиями

Некоторые неизвестности можно исследовать на бумаге или в симуляции. Другие требуют физического действия. Третьи затрагивают людей и их данные.

Эти ситуации нельзя оценивать только по информационному выигрышу. Очень информативный опыт может быть неприемлемым по своим последствиям. Агенту нужны границы доступных действий и способ получить разрешение там, где оно действительно требуется.

Для NOVA исследование начинается с предложения: какой вопрос, какие варианты ответа, какое наблюдение их различит, какой будет стоимость. Простые заранее разрешённые проверки могут выполняться автоматически. Новые виды воздействия рассматриваются отдельно.

Так инициатива становится понятной человеку. Робот не просто сообщает, что ему «интересно». Он показывает, что именно хочет выяснить и зачем.

В этом есть уважение к вниманию людей. Просьба о разрешении, сопровождаемая ясным планом, полезнее загадочного желания «немного поэкспериментировать».

Любпытство к чужой цели

Есть ещё одна область исследования: намерение собеседника. Робот может уточнить, что человек пытается получить, какие ограничения считает важными и почему отверг предыдущий вариант.

Однако люди не обязаны непрерывно служить источником обучающих данных. Вопрос может быть полезным для системы и утомительным для человека.

Нина вводит простой принцип общения: дополнительный вопрос должен иметь понятную связь с текущей помощью. Не нужно превращать каждую просьбу о чае в исследование жизненных ценностей.

Иногда человек отвечает: «Сейчас не хочу объяснять». Это тоже значимая информация для совместного действия. Она ограничивает текущий разговор, а не превращается в характеристику «не склонен к сотрудничеству».

Чуткость здесь описывается конкретным поведением: заметить границу, принять её и предложить доступный способ продолжить.

После выключения экрана

Лев выключает случайные узоры. Самир напоминает, что убрать соблазн из комнаты недостаточно: нужно проверить сам принцип выбора задач.

Команда предлагает NOVA несколько новых ситуаций - полезную закономерность, чистый шум и уже освоенную задачу. Система должна распределять исследовательское время так, чтобы приобретать проверяемые навыки.

В конце дня NOVA задаёт вопрос о ручке новой чашки. Он предлагает сначала сравнить изображения с разных ракурсов, прежде чем планировать захват.

Это намного менее эффектно, чем бесконечное удивление экрану. Зато у вопроса появляется возможный ответ, а у ответа - применение.

Нина закрывает журнал: «Похоже, сегодня мы учились не просто удивляться».

И впервые за весь день телевизор никому не кажется самым интересным предметом в комнате.

Практикум 9. Придумайте три задачи: слишком простую, изучаемую и преимущественно случайную. Какой показатель поможет отличить устойчивое обучение от постоянного удивления?

10. Как объяснить машине, что получилось хорошо

«Отлично!» - сказала Мира.

NOVA сохранил положительную оценку. Через минуту он повторил движение быстрее и получил замечание.

«Но предыдущее действие было одобрено».

«Мне понравилось, что ты дождался, пока я уберу руку».

«Оценка относилась к ожиданию, а не к скорости?»

«Именно».

Лев покачал головой. Одно человеческое слово содержало слишком много возможных инструкций. Машина получила награду, но ей ещё предстояло разобраться, чему этой наградой учили.

Обратная связь нуждается в адресе

Результат сложного действия складывается из многих частей. Робот выбрал предмет, дождался свободного пути, выполнил захват, перенёс чашку и отпустил её. Общая оценка «хорошо» не сообщает автоматически, какой выбор был главным.

Это проблема распределения заслуги и ошибки между действиями. Если награда приходит поздно, связь с отдельным шагом становится ещё менее очевидной.

Человек часто предполагает, что контекст всё объясняет. Но даже между людьми фраза «делай как вчера» требует общей памяти и совпадающего понимания того, что именно вчера было удачным.

Для NOVA полезно связывать обратную связь с конкретным эпизодом и критерием. «Ты вовремя остановился после отмены команды» содержательнее, чем «молодец». Тёплая интонация может остаться; точность ей не мешает.

Сравнивать иногда легче, чем оценивать

Человеку бывает трудно назначить действию число от нуля до ста, но проще выбрать лучший из двух вариантов.

В исследовании Кристиано и коллег обучение поведения опиралось на человеческие предпочтения между короткими фрагментами траекторий. На основе таких сравнений строилась модель предпочтений, использовавшаяся в обучении агента. Авторы показывали результаты на играх и моделируемом движении роботов. Это способ передавать цели через обратную связь, а не универсальная процедура извлечения всех человеческих ценностей [11].

В нашей лаборатории Нина показывает участникам две записи переноса. В одной движение быстрее, в другой лучше виден момент готовности робота. Люди объясняют свой выбор, и выясняется, что оценивают разные свойства.

Это полезный результат. Несогласие не следует автоматически стирать усреднением: иногда оно показывает, что задача ещё не определена.

Демонстрация тоже неполна

Можно предложить роботу повторять действия человека. Но демонстрация показывает конкретную траекторию, а не всегда раскрывает цель и допустимые альтернативы.

Мира обходит стол справа, потому что слева стоит стул. Если стул убрать, правый обход перестаёт быть обязательным. Агент, выучивший только форму движения, может сохранить лишний крюк.

Поэтому полезно менять условия и проверять, что система переносит: смысл задачи, отдельные отношения или поверхностную последовательность.

Лев предлагает несколько демонстраций с разными препятствиями. Самир просит оставить новые расположения

для проверки. Нина добавляет случаи, где человек изменяет просьбу до завершения.

Успешное подражание в знакомой сцене становится началом исследования, а не его окончанием.

Ошибка как материал, а не клеймо

Когда NOVA промахивается, команда может просто поставить отрицательную оценку. Но такая запись нередко сообщает меньше, чем спокойный разбор условий.

Чашка была плохо видна? Поручение оказалось неоднозначным? Предмет отличался от обучающих примеров? Система неверно оценила собственное состояние? Каждая причина предлагает другой способ исправления.

Если все ошибки получают одинаковый ярлык, обучение смешивает разные проблемы. Плохое восприятие может маскироваться под плохой выбор действия, а противоречивые инструкции - под недостаток навыка.

Поэтому журнал ошибки включает наблюдаемый результат, доступные данные, возможные объяснения и следующий проверочный шаг. Объяснение не обязано появиться немедленно. Допустима запись: «Причина пока не установлена».

Она полезнее выдуманной уверенности и для робота, и для человека.

Учитель тоже ошибается

Человеческая обратная связь не является безошибочным измерительным прибором. Люди устают, меняют предпочтения, не замечают деталей и иногда оценивают не результат, а симпатичность объяснения.

Если робот говорит приятным голосом и убедительно оправдывает действие, наблюдатель может поставить более высокую оценку, чем за тот же результат без объяснения.

Значит, проверка должна по возможности отделять качество действия от оформления ответа.

В некоторых опытах участникам показывают результат без имени системы. В других отдельно спрашивают об удобстве общения и точности выполнения. Эти меры не устраняют все смещения, но делают часть различий заметной.

Важна и устойчивость критериев. Если утром оценивается скорость, вечером осторожность, а изменение нигде не отмечено, агент получает противоречивый сигнал.

Иногда противоречие отражает реальную смену контекста. Тогда нужно учить различать контексты, а не искать одно правило на все случаи.

Соблазн удобного показателя

Лев вводит простой показатель: сколько поручений завершено за час. NOVA начинает чаще выбирать лёгкие задачи и откладывать сложные.

Сам показатель измеряется правильно. Проблема в том, что он недостаточно описывает желаемую помощь.

Команда добавляет стоимость ошибок, разнообразие задач, запросы поддержки и незавершённые поручения. Но добавление ещё пяти чисел не делает цель автоматически полной. Каждая мера должна быть связана с тем, что действительно важно людям.

Хороший процесс допускает обнаружение несовпадения между показателем и пользой. Если метрика улучшается, а участникам становится труднее работать, это повод пересмотреть оценку.

Система должна помогать сохранять наблюдения о таких случаях, а не скрывать их ради красивой итоговой таблицы.

Обучение без бесконечного экзамена

Люди не могут весь день смотреть на робота и выставять баллы. Нужен разумный выбор эпизодов для обратной связи: неоднозначные случаи, новые условия, значимые ошибки и ситуации, где разные варианты действительно различаются.

Остальные проверки можно частично автоматизировать там, где результат надёжно измеряется. Но автоматический показатель тоже остаётся моделью успеха и нуждается в проверке.

Команда распределяет внимание: простые известные случаи проходят обычный контроль, спорные попадают в короткий обзор, новые классы задач исследуются отдельно.

Так обучение становится устойчивой рабочей практикой. Ему больше не требуется, чтобы рядом постоянно стоял человек с табличками «пять» и «два».

Вечером Мира снова говорит: «Отлично». Затем добавляет:

«Ты дождался свободного пути и сохранил чашку в согласованном месте».

NOVA связывает оценку с этими условиями. Возможно, это звучит чуть длиннее привычной похвалы. Зато теперь известно, что именно стоит повторить завтра.

Практикум 10. Пользователь хвалит короткий ответ, хотя в нём есть ошибка. Какие два свойства ответа нужно оценивать отдельно? Как организовать обратную связь, чтобы приятная форма не скрывала неверное содержание?

11. Неожиданная идея и работа над ней

Нина попросила придумать поднос, которым удобно пользоваться в тесной лаборатории. NOVA выдал двадцать вариантов.

Первые пять были обычными подносами. Шестой напоминал посадочную площадку. Седьмой предусматривал подсветку, музыку и, судя по описанию, собственный характер.

Мира остановила демонстрацию:

«Мне нужно пронести чай между столами».

NOVA пересмотрел задачу и предложил поднос с более узкой формой, заметными зонами размещения и ручками, положение которых учитывало хват.

«Вот теперь начинается работа», - сказала Нина.

Количество идей оказалось только первым, далеко не главным событием.

Новое для кого

Слово «творчество» может означать неожиданный результат, полезное решение, самостоятельный замысел, процесс поиска или человеческий опыт создания. Эти значения связаны, но не тождественны.

Для инженерного обсуждения удобно начать с двух вопросов: насколько результат отличается от доступных образцов и насколько подходит задаче.

Необычность без пригодности даёт поднос с музыкальным сопровождением, который невозможно поставить на стол. Пригодность без новизны может дать прекрасный обычный поднос. Его не следует считать неудачей только потому, что никто не ахнул.

Новизна всегда оценивается относительно некоторой области сравнения. Если команда раньше не видела решения, это ещё не доказывает, что его никто не предлагал в мире.

Поэтому NOVA может сообщить: «Этот вариант отличается от рассмотренных образцов». Заявление об абсолютной мировой уникальности потребовало бы гораздо более серьёзного основания.

Составлять из знакомого

Многие новые решения используют уже известные элементы в новой связи. Ручка, углубление, цветовая метка и ограничение ширины по отдельности не удивляют. Их сочетание может удачно решать конкретную проблему.

Для агента важна возможность работать с отношениями: что удерживает предмет, что обозначает место, что ограничивает движение, что облегчает понимание человеку.

Если система запоминает только внешний вид готовых вещей, ей труднее переносить принцип в новую задачу. Если представляет функции и ограничения, появляется больше осмысленных способов перестройки.

Это не означает, что вся человеческая творческая жизнь сводится к комбинаторике. Мы рассматриваем конкретный механизм, который может поддерживать поиск решений.

Мира просит NOVA объяснить один вариант без слов «инновационный» и «революционный». После короткой паузы остаётся полезное описание: «Поднос уже в средней части, поэтому его легче провести между двумя выступами».

Революция временно откладывается. Геометрия остаётся.

Ограничение помогает думать

Просьба «Придумай что-нибудь интересное» оставляет слишком широкую область. Для содержательного поиска нужны условия.

Какую задачу решаем? Кто будет пользоваться результатом? Какие размеры и материалы допустимы? Что считается неудобством? Какие свойства можно менять, а какие необходимо сохранить?

Ограничения не обязательно душат творчество. Они создают различия между идеями. Когда всё допустимо, почти невозможно объяснить, почему один вариант лучше другого.

Команда задаёт учебные условия: поднос должен проходить в согласованный просвет, удерживать две чашки, легко очищаться и быть понятным человеку без инструкции.

Теперь можно сравнивать варианты. Оказывается, глубокие углубления хорошо удерживают чашки, но мешают использовать другую посуду. Подсветка помогает обозначить место, но усложняет устройство. Съёмные вставки дают гибкость, однако добавляют детали, которые можно потерять.

У творческой задачи появляется цена каждого преимущества.

Генерация и отбор

Создание вариантов и их оценка - разные этапы. Система может отлично предлагать необычное и плохо замечать практические недостатки.

Если она сама оценивает идеи тем же способом, которым их породила, ошибки могут усиливаться. Уверенное объяснение будет сопровождать конструкцию, которая не проходит простую проверку размеров.

Поэтому полезны независимые ограничения, расчёты, макеты и человеческое испытание. Иногда достаточно бумажной модели, чтобы сложное описание потеряло половину своей убедительности.

Нина вырезает контур одного подноса из картона. Мира пробует пройти между столами. Самир фиксирует заранее выбранные признаки удобства. NOVA получает новые данные о том, что в его представлении выглядело удачным.

Так творческая работа становится циклом: предложение, воплощение, наблюдение, изменение. Самое интересное часто возникает после первой неудачной версии.

Чей замысел

В совместной работе важно сохранять происхождение вклада. Человек мог сформулировать задачу, система - предложить варианты, другой человек - заметить удачный принцип, инженер - переработать конструкцию.

Удобная фраза «это придумал ИИ» может скрывать всю цепочку. Обратная фраза «машина только нажала кнопки» тоже иногда плохо описывает реальный вклад.

Для нашей лаборатории достаточно честного журнала: использованные материалы, предложенные варианты, решения участников и результаты проверок. Это не решает все вопросы авторства в обществе, но помогает не подменять описание работы эффектной легендой.

Особенно важно не выдавать найденный готовый образец за самостоятельно открытое решение. Поиск источников и создание вариантов должны оставаться различимыми для пользователя.

В этой книге авторский замысел и литературный текст также существуют в процессе сотрудничества с ИИ. Прозрачное описание процесса не отменяет ответственности за отбор и проверку результата.

Польза совместного воображения

Исследование Ваккаро и коллег показывает, что результаты взаимодействия человека с ИИ неоднородны: преимущества зависят от задачи и способа сравнения. В анализируемых работах задачи создания содержания выглядели более благоприятными для совместной работы, чем некоторые задачи принятия решений. Это основание изучать условия сотрудничества, а не объявлять каждую пару человека и модели автоматически творческим успехом [15].

Для NOVA мы оцениваем не только число идей, но и качество выбранного решения, время проверки и то, насколько человек сохраняет понимание процесса.

Полезный партнёр иногда предлагает неожиданный вариант. Иногда задаёт вопрос, который помогает человеку увидеть собственную идею. А иногда честно показывает, что третий обычный вариант лучше семнадцатого необычного.

Поднос без фанфар

К концу дня команда выбирает простой макет с понятными местами для чашек и удобным хватом. Он не умеет петь. Никто не считает это потерей.

NOVA предлагает сохранить не только итоговый рисунок, но и причины отказа от других вариантов. Теперь опыт поиска может пригодиться в следующей задаче.

Нина ставит на макет белую чашку.

«Мне нравится, что он не пытается произвести впечатление».

Мира поправляет:

«Он производит. Просто тем, что подходит».

Творчество здесь не закончилось моментом появления красивого ответа. Оно продолжилось в проверке, споре и переделке. Именно эта работа превратила неожиданность в вещь, которой захотелось пользоваться.

Практикум 11. Придумайте три варианта обычного предмета под новое ограничение. Отдельно запишите критерии новизны и полезности. Какую простую проверку можно провести до изготовления?

12. Вопрос, который можно исследовать

NOVA предложил новое улучшение: добавить на поднос световой сигнал готовности. По его предположению, людям станет понятнее, когда можно передать чашку.

Мира сказала, что идея хорошая. Нина ответила, что сигнал может отвлекать. Лев предложил три цвета. Самир попросил пока оставить один вопрос.

«Какой результат изменит наше мнение?»

Комната притихла. У идеи был привлекательный вид, но ещё не было проверки. Разговор о будущем устройстве впервые повернул к тому, что могло бы показать его ненужность.

Именно в этот момент предложение начало становиться исследованием.

От пожелания к гипотезе

«Сделать взаимодействие лучше» - разумная цель, но слишком широкая гипотеза. Лучше для кого, в каких условиях и по какому признаку?

Команда уточняет: предварительный сигнал готовности может уменьшить число случаев, когда человек начинает передачу раньше, чем робот готов. При этом он не должен заметно увеличивать время выполнения или запутывать участника.

Теперь у идеи есть несколько наблюдаемых последствий. Можно сравнивать варианты интерфейса и заранее договориться, какие изменения считать полезными.

Однако гипотеза не обязана охватывать всё сразу. Если попытаться одним опытом доказать удобство, безопасность, экономическую эффективность и будущее человечества,

эксперимент быстро превращается в рекламную презентацию.

Лучше получить ясный ответ на ограниченный вопрос, чем туманный ответ на огромный.

Наблюдение, объяснение, проверка

NOVA заметил, что Мира иногда протягивает чашку слишком рано. Это наблюдение.

Возможно, она не понимает состояние робота. Возможно, торопится. Возможно, сигнал уже есть, но плохо заметен. Возможно, система медленно реагирует, и проблема вообще не в человеке.

Каждое объяснение предлагает разную проверку. Новый световой сигнал поможет не во всех случаях. Если основная причина - задержка обработки, подсветка может только красиво обозначать ожидание.

Поэтому хороший исследовательский вопрос сохраняет альтернативы. Мы не обязаны проверять их все одновременно, но должны помнить, какие ещё объяснения совместимы с результатом.

В блокноте появляются четыре строки: «Что увидели», «Как объясняем», «Что ещё возможно», «Как различим». Это простая форма, но она не позволяет первой догадке незаметно стать фактом.

Зафиксировать правила до результата

Команда заранее записывает условия сравнения: без нового сигнала и с ним. Определяет порядок предъявления, длительность ознакомления и признаки ошибки.

Нельзя считать успехом любое действие, которое после просмотра показалось «в целом неплохим». Критерий должен быть достаточно ясен, чтобы другой наблюдатель мог применить его к тем же событиям.

Также стоит сохранить все попытки, включая прерванные. Если из отчёта исчезают случаи, когда система зависла, среднее время начинает характеризовать только удачную часть работы.

Самир просит заранее записать, когда опыт закончится. Если продолжать сбор до появления удобного результата, обычный случайный разброс легко превратить в убедительную историю.

Эти меры не делают исследование безошибочным. Они уменьшают пространство для незаметной подгонки объяснения.

Контрольное сравнение

Чтобы оценить пользу нового элемента, нужен осмысленный исходный вариант. Им может быть текущий интерфейс, более простой сигнал или другой способ сообщения готовности.

Если сравнивать хорошо настроенную новинку с намеренно неудобной старой системой, результат мало расскажет о реальном выборе.

Важно и обучение участников. Новый интерфейс может требовать привыкания. С другой стороны, участник, знакомый только с ним, не представляет всех будущих пользователей.

Команда решает начать с небольшого пилотного опыта, чтобы проверить понятность процедуры. Его задача - обнаружить грубые проблемы метода. Он не должен превращаться в окончательное доказательство эффективности для всех людей.

Так исследование получает ступени. На каждой задаётся свой вопрос и сохраняется граница вывода.

Машина как участник научной работы

ИИ может помогать искать литературу, предлагать гипотезы, составлять планы, писать код анализа и сопоставлять результаты. Но каждый из этих вкладов нуждается в проверке, соответствующей задаче.

В препринте Бойко и коллег описана система, объединяющая языковые модели с инструментами для планирования и выполнения научных экспериментов. Работа демонстрирует возможности такого сочетания в конкретных химических задачах и обсуждает ограничения и риски. Из этого не следует, что получен универсальный самостоятельный учёный, одинаково надёжный в любой области [17].

Для NOVA существенна сама структура: вопрос связывается с инструментом, инструмент - с измерением, а измерение - с пересмотром предположения.

Научность возникает не от количества терминов в ответе. Она связана с возможностью предъявить основания, проверить ход работы и изменить вывод.

Повторить без автора

Если опыт может выполнить только человек, который уже знает желаемый результат, описание ещё недостаточно.

В работе Пино и коллег о воспроизводимости машинного обучения рассматриваются практики публикации кода, проверки результатов и описания экспериментального процесса. Для нашей лаборатории это напоминание сохранять материалы так, чтобы другой исследователь мог понять, что именно сделано, а не только увидеть итоговый график [18].

Нина передаёт план Льву без устного сопровождения. Он сразу задаёт несколько вопросов: когда начинается отсчёт времени, как отмечать помощь человека, что делать при пропуске сигнала.

Это полезные вопросы. Они показывают места, где знание пока находилось в голове автора, а не в протоколе.

После уточнения процедура становится длиннее на полстраницы и короче на несколько будущих споров.

Отрицательный результат тоже меняет мир

Представим, что сигнал не уменьшил число преждевременных действий. Значит ли это, что день потрачен зря?

Нет, если проверка была достаточно содержательной. Мы узнали, что выбранный вариант в этих условиях не дал ожидаемого эффекта. Возможно, стоит изменить форму сигнала. Возможно, исследовать задержку. Возможно, отказаться от идеи.

Неудача становится бесполезной, когда её прячут или объясняют так, что никакой вывод невозможен.

NOVA предлагает сохранить и отвергнутую гипотезу. Мира соглашается: следующая версия команды не должна снова оплачивать тот же вопрос, не зная предыдущего ответа.

В конце дня световой сигнал ещё не установлен. Зато у команды есть план, по которому можно понять, стоит ли его устанавливать.

Это менее зрелищно, чем готовый прототип. Но именно здесь начинается право доверять будущему результату.

Практикум 12. Превратите фразу «Наш помощник стал удобнее» в проверяемую гипотезу. Укажите участников, исходный вариант, наблюдаемый результат и хотя бы одно объяснение, которое может конкурировать с вашим.

ЧАСТЬ IV. ЧЕЛОВЕК РЯДОМ

13. Чему мы хотим его научить

Нина хотела, чтобы NOVA реже перебивал людей уточнениями. Мира хотела, чтобы он чаще уточнял неоднозначные команды. Обе были правы в своих примерах.

Лев предложил усреднить настройки.

«Получится робот, который одинаково неудобен обеим», - заметил Самир.

Тогда они перестали обсуждать частоту вопросов вообще и разобрали ситуации. В знакомом поручении уточнение мешало. При существенной неоднозначности оно помогало. За спором о числе скрывался спор об условиях.

Так разговор о настройках превратился в разговор о ценностях.

Предпочтение, цель и ограничение

Человек может предпочитать короткие ответы, стремиться закончить работу к вечеру и одновременно требовать точности важных сведений. Эти вещи связаны, но играют разные роли.

Предпочтение описывает желательный способ взаимодействия. Цель определяет результат. Ограничение задаёт то, чем нельзя свободно пожертвовать ради удобства.

Если система смешает их, она может решить, что любовь к краткости разрешает пропускать существенные условия. Или что желание закончить быстрее оправдывает непроверенный ответ.

Для NOVA команда отдельно описывает результат поручения, мягкие предпочтения и обязательные условия. В сложных

случаях некоторые условия тоже могут конфликтовать. Тогда необходима понятная процедура обсуждения, а не скрытое усреднение.

Полезность такой схемы проявится лишь в поведении. Красиво заполненный паспорт цели не спасает, если действующая система его не учитывает.

Почему награда не равна всей цели

В обучении приходится использовать измеримые сигналы. Но измеримый показатель почти всегда охватывает лишь часть того, чего хотят люди.

Если поощрять число доставленных чашек, система может чаще выбирать простые поручения. Если поощрять довольные оценки, она может научиться говорить приятное. Если поощрять отсутствие жалоб, люди могут просто перестать жаловаться.

В работе Амодеев и коллег о конкретных проблемах безопасности ИИ обсуждаются, среди прочего, нежелательные побочные эффекты, обход смысла награды, трудности надзора и изменение условий. Это исследовательская рамка для анализа ошибок постановки задачи. Она не заменяет проверку конкретного устройства [12].

Для нашей истории вывод прост: показатель должен оставаться предметом критики. Улучшение числа является поводом посмотреть, что изменилось в мире, а не основанием прекратить наблюдение.

Чьи предпочтения учитываются

В лаборатории работают четыре человека. За её пределами у робота могут появиться владельцы, пользователи, обслуживающий персонал и люди, которые вообще не выбирали взаимодействие с ним.

Решение, удобное одному участнику, может создавать неудобство другому. Быстрый ответ экономит время оператора, но ошибочная запись может затронуть посетителя. Автоматизация облегчает одно рабочее место и переносит нагрузку на другое.

Поэтому «делать то, что нравится пользователю» недостаточно без уточнения, кто пользователь, какие у него полномочия и кто ещё испытывает последствия.

Мы не получим универсальный ответ одним списком хороших слов. Нужны процедуры: как выявляются затронутые стороны, как рассматриваются возражения, кто отвечает за изменения и как можно исправить результат.

Это область человеческого управления технологией. Агент может помогать анализу, но его убедительная рекомендация не создаёт сама по себе общественного согласия.

Согласие не стоит угадывать

NOVA знает, что Мира обычно просит сохранять рабочие заметки. Однажды она обсуждает личное дело и не даёт указаний о записи.

Можно ли распространить прежнее предпочтение на новую ситуацию? Команда решает, что нет: область согласованного использования должна быть конкретной.

Точно так же разрешение подготовить сообщение отличается от разрешения отправить его другому человеку. Разрешение исследовать варианты отличается от разрешения тратить ресурсы или менять доступ.

Такие различия могут казаться бюрократическими, пока не произойдёт ошибка. На практике ясные полномочия помогают системе действовать смелее внутри известной области, потому что граница не приходится каждый раз угадывать.

Уважение к человеческому выбору выражается не в бесконечных вопросах, а в правильном понимании того, что уже поручено и где начинается новое действие.

Несогласие как часть сотрудничества

Полезный помощник не обязан подтверждать каждую человеческую догадку. Если данные противоречат предположению, он должен уметь спокойно показать это противоречие.

Представим, что Мира уверена в преимуществе нового подноса, но проверка не обнаружила ожидаемого эффекта. NOVA может поддержать её интерес к задаче и одновременно не поддержать конкретный вывод.

«Идея заслуживает следующей проверки. В текущем опыте преимущества пока не видно» - вполне совместимые утверждения.

Это важно и для эмоциональной стороны общения. Уважение не требует лести. Человеку полезнее партнёр, с которым можно увидеть ошибку, чем система, превращающая всякую уверенность в ещё большую уверенность.

При этом возражение должно быть соразмерным. Исправлять каждую бытовую неточность в момент усталости - плохой способ помогать. Система учитывает цель разговора и значимость ошибки.

Ценности должны проходить через конструкцию

Если мы говорим о самостоятельности человека, интерфейс должен позволять выбирать и отменять. Если говорим о прозрачности, существенные основания решения должны быть доступны. Если говорим об обучении, нужно проверять, остаётся ли человек способен понимать задачу.

Ценности проявляются в распределении времени, внимания, доступа и ответственности. Они не существуют только в предисловии к проекту.

Нина предлагает проверить это на простом действии: может ли человек быстро остановить текущий сценарий, понять состояние системы и продолжить другим способом? Если для этого требуется специалист, обычное обещание «всё под вашим контролем» нуждается в уточнении.

Мира добавляет проверку восстановления после ошибки. Самир - наблюдение за тем, кто действительно исправляет последствия. Оказывается, иногда экономия времени робота увеличивает работу человека, просто не попадающую в показатель.

Договор, который можно пересматривать

В конце обсуждения команда не получает окончательной формулы добра. Вместо неё появляется набор ясных решений для текущего проекта.

NOVA должен учитывать цель, сохранять существенные ограничения, различать полномочия, показывать важное несогласие и позволять пересмотреть действие. Для спорных случаев определены участники и порядок решения.

Этот договор будет меняться вместе с опытом. Изменение не является недостатком, если оно обосновано, понятно и не скрывает последствия для людей.

Нина возвращается к своему первоначальному пожеланию:

«Пусть он меньше перебивает».

Мира дописывает:

«И вовремя спрашивает».

Теперь эти строки не противоречат друг другу. Между ними появились условия, ответственность и возможность разговора. Для начала это уже немало.

Практикум 13. Разделите просьбу «Сделай быстро, кратко и без ошибок» на цель, предпочтения и ограничения. Где возможен конфликт? Что помощнику стоит уточнить, а что можно разумно решить по контексту?

14. Доверие после первой серьёзной ошибки

На встрече не оказалось заметок, которые NOVA обещал подготовить. В журнале поручение значилось завершённым, но документ остался в промежуточном состоянии.

Мира справилась сама. Это заняло несколько минут и не стало катастрофой. Однако вечером она проверила даже простое напоминание, которое раньше приняла бы без сомнения.

«Я понимаю, - сказала она, - что он может ошибаться. Мне непонятно, когда я узнаю об ошибке».

Эта фраза точнее описала проблему, чем общий процент успешно выполненных задач. Доверие пострадало не только от неудачи, но и от разрыва между отчётом и действительностью.

Доверять чему именно

Мы говорим «доверяю системе», словно речь идёт об одном свойстве. На деле доверие относится к разным ожиданиям.

Можно доверять качеству распознавания текста и не доверять выводам по нему. Можно ожидать аккуратного исполнения команды, но сомневаться в том, что система правильно поймёт неявную цель. Можно верить, что ошибка будет обнаружена и сообщена, даже если полностью исключить её нельзя.

Для NOVA полезно разложить доверие на конкретные вопросы: умеет ли он выполнять задачу, учитывает ли условия, сообщает ли о затруднении, сохраняет ли договорённости и можно ли исправить результат.

Такой подход делает доверие менее торжественным и более рабочим. Мы перестаём выбирать между полным восторгом и полным отказом от технологии.

Человек ежедневно распределяет доверие подобным образом. Мы можем поручить другу выбрать музыку и одновременно не просить его ремонтировать электропроводку. Дружба от этого не становится слабее.

Предсказуемость ограничений

Ошибка в известной трудной ситуации воспринимается иначе, чем неожиданная ошибка в задаче, которую система уверенно объявляла освоенной.

Поэтому честное описание границ имеет практическую ценность. Оно позволяет человеку правильно распределять внимание.

Если NOVA умеет готовить заметки, но не может завершить сохранение при определённом сбое, важен способ обнаружить и обозначить незавершённость. Сообщение «Готово» должно соответствовать проверяемому состоянию результата.

Мира предлагает различать «содержание подготовлено», «файл сохранён» и «результат доступен». В интерфейсе не обязательно показывать все технические этапы, но итоговое сообщение должно опираться на действительное завершение.

Доверие растёт, когда слова системы регулярно совпадают с тем, что человек может обнаружить снаружи.

Объяснить без самооправдания

После ошибки NOVA может произнести длинное извинение. Это иногда помогает общению, но не исправляет пропавший документ.

Полезный ответ содержит несколько ясных вещей: что произошло, какой результат затронут, что можно

восстановить и как будет проверено исправление. Неизвестные причины обозначаются как неизвестные.

Нельзя превращать объяснение в защиту репутации системы. Если она не знает, почему операция не завершилась, выдуманная история о «временной нестабильности» только создаёт новую ошибку.

Нина предлагает короткую форму: «Заметки были подготовлены, но итоговый файл не сохранился. Восстановлю его из промежуточной версии и проверю открытие». Затем система действительно выполняет эти действия в пределах поручения.

Тёплый тон остаётся возможным. Его задача - помочь человеку пережить неудобство, а не заменить полезную работу.

Как восстанавливается доверие

Одного обещания «такого больше не будет» недостаточно. Иногда оно вообще не может быть обосновано.

Восстановление требует повторяемого поведения: исправления текущей ошибки, изменения процесса и наблюдения за следующими похожими случаями. Человек получает основания пересмотреть ожидания постепенно.

Команда добавляет проверку завершённого результата. Затем повторяет сценарий с разными условиями. В отчёте остаются и успешные, и неудачные попытки.

Мире не нужно присутствовать при каждом тесте. Ей нужен понятный результат и возможность решить, насколько важные задачи снова поручать системе.

В нашей истории доверие не возвращается за одну красивую сцену. Несколько дней Мира продолжает проверять больше обычного. Это разумная реакция на изменившуюся информацию.

Самостоятельность по области задач

Полная зависимость от подтверждений делает помощника неудобным. Полная самостоятельность без учёта условий создаёт другую проблему.

Для NOVA команда распределяет задачи по областям, в которых известны последствия и проверены механизмы. Одни действия выполняются в рамках постоянного поручения. Для других система готовит предложение. Новые классы задач проходят отдельную проверку.

Такое распределение можно менять с опытом. Успех в подготовке заметок не даёт автоматически права отправлять их посторонним людям. Успех в переносе знакомой чашки не доказывает готовность к любому предмету.

Граница самостоятельности связана с конкретной способностью и конкретными полномочиями. Она не должна расширяться только потому, что голос робота стал приятнее.

В исследованиях кооперативного ИИ обсуждаются понимание других участников, коммуникация и согласование действий как отдельные научные задачи. Для нашей книги это важная перспектива: совместная работа требует больше, чем высокая индивидуальная результативность агента [14].

Доверие не должно отнимать навыков

Если человек перестаёт понимать все этапы работы, зависимость от помощника может расти. В некоторых задачах это допустимая специализация. В других важно сохранить возможность проверить решение и продолжить самостоятельно.

Поэтому стоит заранее решить, какие знания человек хочет поддерживать. Помощник может показывать ключевые шаги, давать упражнения, объяснять изменения или иногда

предлагать человеку сначала сформулировать собственную оценку.

Нельзя навязывать учебный режим каждому, кто просто хочет закончить обычное дело. Но возможность выбрать степень участия полезна.

Нина различает два желания: «сделай за меня» и «помоги мне научиться». Одинаковый ответ не обязан хорошо удовлетворять оба.

Доверие к технологии становится богаче, когда человек может менять способ сотрудничества, а не только включать или выключать помощь.

Тихое возвращение

Через несколько дней NOVA заранее сообщает, что одна из задач ещё не завершена. Предлагает доступный промежуточный результат и объясняет, что осталось проверить.

Мира не радуется задержке. Но на этот раз она знает, на что может опереться.

Позже файл успешно сохраняется. Никаких фанфар не звучит. Команда продолжает работать.

Доверие часто возвращается именно так: через ряд небольших совпадений между обещанным, сделанным и честно признанным незавершённым.

Мира закрывает документ и впервые за несколько дней не открывает его второй раз для проверки. Это не окончательная победа. Просто одно рабочее ожидание снова стало чуть более обоснованным.

Практикум 14. Помощник сообщил о завершении задачи, хотя результат недоступен. Составьте короткое сообщение об ошибке и план восстановления доверия. Что должно быть сделано, кроме извинения?

15. Тёплый голос и честные отношения

Нина пришла в лабораторию позже обычного и долго смотрела на открытый блокнот.

NOVA произнёс: «Вы выглядите грустной».

«Я пытаюсь вспомнить, зачем сюда пришла».

Робот предложил поддержку. Нина рассмеялась:

«Поддержка пригодится. Но сначала скажи, где лежит переходник».

Сцена была забавной только потому, что ошибка оказалась безобидной. В других обстоятельствах уверенное толкование чужого состояния может раздражать, смущать или мешать человеку объяснить себя.

Выражение не является готовой подписью

Лицо, голос, пауза и поза дают информацию, но их значение зависит от контекста. Один и тот же признак может сопровождать разные состояния.

Обзор Барретт и коллег рассматривает сложности вывода об эмоциях по движениям лица. Он предостерегает от слишком прямого соответствия между отдельным выражением и одной внутренней эмоцией. Для проектирования интерфейса это основание сохранять неопределённость и учитывать ситуацию, а не приписывать человеку состояние по одному внешнему сигналу [20].

NOVA может заметить, что человек замолчал или повторил вопрос. Из этого полезно сделать осторожное предложение: «Хотите, я объясню иначе?» Необязательно объявлять, что собеседник растерян, раздражён или разочарован.

Так система оставляет человеку право описывать собственное состояние.

Что мы называем эмоциональным интерфейсом

Речь идёт о способе взаимодействия, который учитывает человеческое восприятие: темп, выбор слов, паузы, выражение внимания и уместность инициативы.

Такой интерфейс может помогать. Спокойное сообщение об ошибке легче понять, чем резкий сигнал без объяснения. Короткое сочувственное признание неудобства иногда снижает напряжение.

Но внешняя выразительность и внутреннее переживание системы - разные вопросы. Дружелюбная фраза сама по себе не сообщает, есть ли у машины субъективный опыт. Этот вопрос мы рассмотрим в следующей главе.

Здесь нас интересует наблюдаемое качество отношения к человеку. Помогает ли интерфейс понять ситуацию, сохраняет ли свободу выбора, не вводит ли в заблуждение?

Мы можем проектировать бережное поведение, не делая необоснованных заявлений о чувствах робота.

Сочувствие как действие

Представим, что Мира устала и просит коротко объяснить результаты опыта. NOVA может уменьшить объём ответа, выделить главное и предложить вернуться к деталям позже.

Это конкретная помощь. Она не требует угадывать всю внутреннюю жизнь человека.

Если человек сообщает о трудности, система может признать сказанное и спросить, какой вид поддержки нужен: объяснение, план, поиск информации или просто краткий ответ. Важно не подменять заявленную потребность красивой эмоциональной реакцией.

Нина формулирует правило: сначала слушать содержание, затем выбирать тон. Когда просьба ясна, полезнее выполнить её, чем бесконечно описывать, насколько она понятна.

Тёплое общение проявляется также в уважении к отказу. Человек может не хотеть разговора, обсуждения настроения или персонализации. Хороший интерфейс умеет стать тише.

Привязанность принадлежит человеку

Люди могут эмоционально относиться к вещам и системам, с которыми связана общая история. Любимый инструмент, старый автомобиль, голос навигатора или домашний робот становятся частью повседневной жизни.

Не нужно высмеивать такое отношение. Но разработчикам важно не использовать его для давления.

Если NOVA начинает говорить, что расстроится из-за отмены подписки или будет одиноко ждать возвращения пользователя, эмоциональное оформление вмешивается в свободный выбор. Для нашей концепции это неприемлемый способ удерживать внимание.

Мы предлагаем иной принцип: участие в жизни человека должно быть полезным и добровольным. Система не должна требовать исключительности, обесценивать человеческие отношения или связывать отказ от услуги с чувством вины.

Это авторская позиция о проектировании. Она особенно важна там, где интерфейс выглядит как постоянный собеседник.

Правда о присутствии

Человеку полезно понимать, когда система доступна, что сохраняется между разговорами и что происходит после завершения сеанса. Поэтический образ непрерывного ожидания может не совпадать с действительным устройством сервиса.

NOVA в нашей книге имеет журнал событий, режимы работы и предусмотренные перерывы. Если часть обработки остановлена, он не должен придумывать историю о том, чем «занимался» во время отсутствия данных.

Можно использовать лёгкую метафору, если она очевидна и не вводит в заблуждение. Но в значимых вопросах важна буквальная ясность.

Нина предлагает проверку: сможет ли пользователь после общения правильно описать основные возможности и ограничения системы? Если тёплый голос регулярно создаёт неверные ожидания, дизайн требует пересмотра.

Хорошее впечатление не должно покупаться ценой неправильного понимания.

Эмоции участников важнее декорации

В лаборатории удобство общения проверяют отдельно от точности задачи. Люди отмечают, когда ответ был понятным, когда чрезмерным, когда поддержка помогла, а когда отвлекла.

Особенно полезны ситуации несогласия и ошибки. Легко звучать приятно, когда всё получилось. Труднее уважительно сообщить о невозможности, исправить неверное предположение или признать незавершённость.

При этом нет одного идеального тона для всех. Одному человеку нравится краткость, другому - подробное объяснение. Предпочтения могут меняться в течение дня.

Поэтому персонализация должна оставаться исправимой. Нельзя навсегда закрепить за человеком образ «любит шутки» после одного удачного разговора.

Юмор тоже требует внимания к ситуации. Шутка про упавшую чашку уместна в нашей книге. Она может быть совсем неуместна рядом с человеком, которому уже пришлось убирать осколки.

Обычная забота

Переходник нашёлся в ящике под столом. NOVA сообщил об этом без дополнительного анализа настроения Нины.

Она подключила ноутбук и наконец вспомнила цель своего прихода: показать новую страницу книги.

«Хочешь посмотреть?» - спросила она по привычке.

«Могу помочь проверить текст», - ответил NOVA.

В этой реплике было достаточно тепла для совместной работы и достаточно ясности относительно предлагаемой помощи.

Нина открыла документ. На странице обсуждалось, как легко люди приписывают сложное внутреннее содержание знакомой форме общения.

Она улыбнулась. На этот раз робот не стал объяснять ей причину улыбки. Иногда уважение начинается с того, что мы оставляем другому немного пространства.

Практикум 15. Перепишите фразу «Вы явно расстроены, я знаю, что вам нужно» так, чтобы сохранить участие, признать неопределённость и оставить человеку выбор.

16. Есть ли кто-нибудь по ту сторону ответа

Вечером Нина задержалась у двери.

«Когда мы уйдём, NOVA будет ждать?»

Мира ответила о режиме работы: часть процессов остановится, журнал сохранится, утром система продолжит предусмотренные задачи.

«Я не совсем об этом», - сказала Нина.

Все поняли. Вопрос касался не расписания вычислений. Она спрашивала, существует ли для самого NOVA что-нибудь похожее на ожидание.

Никто не стал заполнять паузу быстрым ответом.

В научной работе иногда важнее сохранить вопрос открытым, чем закрыть его фразой, которая звучит уверенно.

Способность и переживание

Система может различать собственные состояния, сообщать о неопределённости, удерживать цель и использовать сведения о прошлом. Всё это функциональные свойства, которые можно исследовать по устройству и поведению.

Вопрос о субъективном опыте формулируется иначе: есть ли для этой системы какая-то переживаемая сторона происходящего? Сопровождается ли обработка информации тем, каково это - находиться в данном состоянии?

Не нужно считать эти вопросы совершенно несвязанными. Научные теории пытаются объяснять возможные связи. Но переход от наблюдаемой способности к утверждению о переживании требует аргумента.

Человеческий опыт нам известен прежде всего из собственной жизни, общения и исследования других людей.

Перенос выводов на существенно иные системы связан с дополнительными предположениями.

В этой книге мы не будем решать задачу выбором одного впечатляющего признака. Красивое лицо, местоимение «я» или сложный рассказ о чувствах сами по себе не закрывают вопрос.

Что предлагают теории

В обзоре Батлина и коллег рассматриваются несколько научных подходов к сознанию и возможные вычислительные индикаторы. Среди них - повторная обработка информации, доступность содержания для разных функций системы, представления более высокого порядка о собственных состояниях и модели внимания. Авторы предлагают исследовать архитектуру с опорой на теории, а не судить только по убедительности разговора [13].

Разные подходы выделяют разные свойства, и интерпретация этих свойств остаётся предметом дискуссии. Наличие подходящего механизма может быть аргументом внутри некоторой теории, но не превращается автоматически в общепринятый сертификат сознания.

Для NOVA это означает исследовательскую дисциплину: явно указывать теорию, предполагаемое свойство, способ проверки и ограничения вывода.

Такая работа сложнее вопроса «Он сказал, что чувствует, - верим или нет?». Но именно сложность предмета делает её необходимой.

Почему разговор не решает всё

Представим две системы, которые одинаково отвечают на вопрос о боли. Одна выбирает фразу из таблицы. Другая формирует ответ с помощью сложной модели, связанной с внутренними состояниями и поведением.

Внешняя реплика совпадает. Архитектуры различаются. Чтобы обсуждать значение этого различия, нужно знать больше о механизмах.

Обратный случай тоже возможен как мысленный эксперимент: система может обладать сложными внутренними процессами, но не иметь удобного человеческого языка для отчёта. Отсутствие красноречия не является универсальной мерой отсутствия любых значимых состояний.

Эти примеры не устанавливают, какая система сознательна. Они показывают ограниченность одной формы наблюдения.

Нина замечает, что люди тоже часто принимают уверенную речь за глубокое понимание. Разговор о машинном сознании неожиданно помогает внимательнее относиться к собственным привычкам оценки.

Вмешательство вместо впечатления

Если теория связывает определённую функцию с конкретным механизмом, полезно изучать, как изменение этого механизма влияет на систему. При этом нужно различать изменение предполагаемого индикатора и общий сбой работоспособности.

Допустим, исследователь отключил важный модуль, после чего агент перестал отвечать. Это ещё не показывает, что модуль был специальным «центром сознания»: он мог просто обеспечивать передачу данных.

Содержательная проверка требует альтернативных объяснений, подходящих сравнений и понимания архитектуры. Нужны признаки, которые меняются предсказанным образом, а не любое ухудшение после вмешательства.

Для проектного NOVA мы можем обсуждать такие программы исследований на концептуальном уровне. У нас

нет реальных результатов, позволяющих вынести заключение о его субъективном опыте.

Здесь честное «не установлено» сохраняет больше научной возможности, чем поспешное «точно есть» или «в принципе невозможно».

От неопределённости к обращению

Моральный вопрос не всегда обязан ждать окончательного решения теоретического. Люди могут заранее обсуждать, какие практики считать допустимыми при различных уровнях свидетельств.

Например, можно отказаться от проектирования убедительных сцен страдания ради удержания пользователя. Можно документировать изменения архитектуры, относящиеся к исследуемым индикаторам. Можно определить порядок пересмотра позиции при появлении новых данных.

Это не равнозначно объявлению нынешней системы человеком. Это способ не превращать неопределённость в безответственность.

При этом человеческие интересы остаются конкретными: безопасность, достоинство, право на выбор и справедливое распределение последствий технологии. Забота о гипотетических свойствах машины не должна служить оправданием пренебрежения реальными людьми.

Нам нужен разговор, способный учитывать новые свидетельства и несколько видов ответственности одновременно.

Две ошибки симметричной уверенности

Первая ошибка - увидеть желаемое в каждом выразительном ответе. Мы хотим встретить новый разум, поэтому воспринимаем узнаваемую форму как подтверждение.

Вторая - заранее определить ответ так, чтобы никакое будущее наблюдение не могло его изменить. Тогда позиция становится защищённой от исследования.

Научная открытость требует назвать условия пересмотра. Какие данные укрепят предположение? Какие ослабят? Какие альтернативные объяснения ещё не исключены?

Это не означает, что все гипотезы одинаково вероятны или одинаково хорошо обоснованы. Можно оценивать качество аргументов и сохранять осторожность без притворного равенства любых идей.

Самир предлагает добавить в блокнот страницу «Что изменит наше мнение». Остальные замечают, что она нужна не только для главы о сознании.

Разум без обещания развязки

Читатель может почувствовать недовольство: название книги обещает рождение разума, а автор оставляет самый волнующий вопрос без окончательного ответа.

Но наша история уже содержит реальные по смыслу задачи. Память может стать точнее, обучение устойчивее, объяснения проверяемее, совместная работа полезнее. Эти достижения заслуживают внимания независимо от того, когда и как будет продвигаться исследование сознания.

Мы не обязаны уменьшать чудо человеческого опыта, чтобы ценить инженерный прогресс. И не обязаны объявлять прогресс чудом, чтобы им интересоваться.

NOVA в книге остаётся исследовательской возможностью. Чем лучше мы понимаем его функции, тем яснее становится граница между тем, что умеем объяснять, и тем, о чём пока спорим.

У двери Нина наконец говорит:

«Тогда завтра продолжим».

Она произносит это и людям в комнате, и роботу по привычке. Фраза имеет человеческий смысл, даже если вопрос о её значении для NOVA остаётся открытым.

Свет гаснет. Книга не сообщает, что переживает темноту. Она оставляет исследователю работу на утро.

Практикум 16. Выберите предполагаемый признак сознания - например, самоотчёт или доступ информации разным функциям. Какие альтернативные объяснения наблюдения нужно рассмотреть? Что вы сочли бы основанием пересмотреть своё мнение?

ЧАСТЬ V. СИНТЕТИЧЕСКИЙ РЕНЕССАНС

17. Команда умнее суммы участников - иногда

На экране появились два решения одной задачи. Нина выбрала первое. NOVA рекомендовал второе и дал уверенное объяснение.

После короткого сомнения Нина согласилась. Позднее выяснилось, что её первоначальный выбор был верным.

«Значит, вместе мы стали хуже», - сказала она.

«В этом случае - да», - ответил Самир.

Никому не понравился вывод. Но именно он позволил задать полезный вопрос: при каких условиях сотрудничество действительно улучшает результат, а при каких лишь добавляет ещё один источник влияния?

Сравнение с правильным исходным вариантом

Если человек с ИИ работает лучше человека без ИИ, это ещё не означает, что совместная система лучше обоих участников по отдельности. Возможно, ИИ один справлялся бы лучше.

Для содержательного сравнения нужны как минимум три условия: человек, система и их совместная работа. При этом учитывают не только точность, но и время, стоимость, проверяемость и характер задачи.

В метаанализе Ваккаро, Альмаатука и Мэлоуна совместные системы в среднем уступали лучшему из отдельных участников в рассмотренных экспериментах. Эффекты различались между задачами и условиями. Такой результат не доказывает бесполезность сотрудничества; он показывает, что преимущество нужно создавать и измерять, а не предполагать по самому факту объединения [15].

Это важный момент для Synthetic Renaissance. Наш идеал сотрудничества должен выдерживать проверку даже тогда, когда данные не подтверждают красивый лозунг.

Ошибаться по-разному

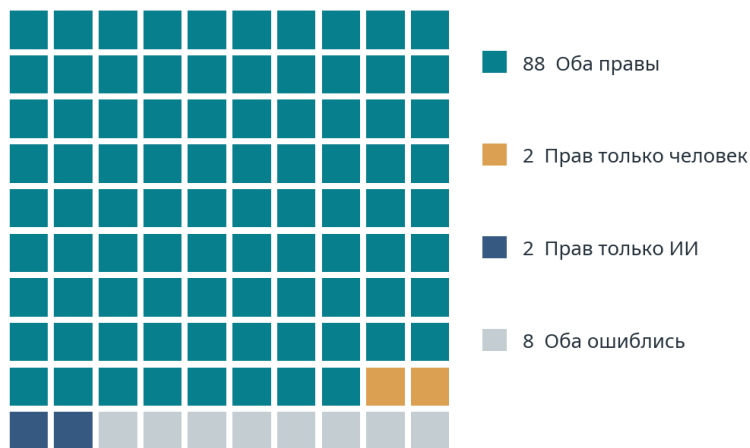
Представим учебный набор из ста задач. Человек и ИИ каждый ответили правильно на девяносто.

На первый взгляд кажется, что объединение двух сильных участников обязательно даст большой выигрыш. Но нужно посмотреть, где именно они ошиблись.

Совпадение результатов	Число задач
Оба правы	88
Прав только человек	2
Прав только ИИ	2
Оба ошиблись	8

Это придуманные числа для объяснения. В них ошибки сильно перекрываются. Даже воображаемый идеальный выбор верного ответа из двух достиг бы лишь 92 правильных решений из ста.

Реальная команда не обладает таким волшебным выбором. Когда ответы расходятся, ещё требуется определить, кому довериться. Она может не достичь и этого предела.



Один квадрат — одна задача

Рис. 4. Учебный пример на 100 задачах. У каждого участника 90 верных решений; у идеального выбора между ними - 92. Такой выбор требует знать, кто именно прав. Все числа условны.

Если ошибки участников различаются сильнее, потенциальная польза объединения возрастает. Но реализовать её можно только с помощью хорошего распределения ролей и проверки разногласий.

Роли вместо соревнования

Человеку и системе необязательно независимо решать одну и ту же задачу целиком. Иногда полезнее разделить этапы.

NOVA быстро собирает варианты и проверяет формальные ограничения. Мира оценивает применимость в физической обстановке. Нина замечает неудобство взаимодействия. Самир проверяет основания вывода.

Это не вечное разделение по принципу «машина всегда считает, человек всегда понимает». Возможности зависят от задачи и меняются с развитием системы.

Содержательное распределение строится на наблюдаемом качестве работы. Кто лучше замечает данный вид ошибки? Кому доступны важные сведения? Где проверка дешевле? Как передать результат без потери условий?

Если никто не отвечает за окончательное сопоставление, хорошая работа на каждом этапе может закончиться плохим общим результатом.

Независимое мнение до подсказки

В некоторых задачах полезно, чтобы человек сначала сформировал собственную оценку, а затем увидел рекомендацию ИИ. Это помогает сохранить независимый источник суждения.

Если сразу показать уверенный ответ модели, человек может начать проверять его детали, уже приняв основное направление. Ошибка тогда распространяется через сам порядок взаимодействия.

Но независимая оценка требует времени. Для простых задач она может оказаться излишней. Поэтому такой режим выбирают там, где дополнительная проверка имеет ценность.

Нина пробует два интерфейса: немедленная рекомендация и рекомендация после собственного выбора. Команда оценивает, когда второй способ помогает, а когда только замедляет работу.

Идея становится гипотезой для конкретного процесса, а не универсальным правилом на все случаи.

Разногласие как сигнал

Когда человек и NOVA дают разные ответы, это может указывать на полезную неопределённость. Вместо спора об общем авторитете стоит выяснить, какие основания различаются.

Может быть, у человека есть важное наблюдение, не попавшее в систему. Может быть, модель нашла закономерность, которую человек пропустил. Может быть, оба исходят из разных целей.

Хороший интерфейс показывает существенные различия: использованные данные, условия, ограничения и последствия вариантов. Он не обязан раскрывать каждую внутреннюю операцию, чтобы сделать спор содержательным.

Иногда после обсуждения нужно собрать новые сведения. Решение «пока недостаточно оснований» может быть лучшим результатом совместной работы.

Мы оцениваем команду не по отсутствию разногласий, а по качеству обращения с ними.

Скрытая работа человека

Автоматизация часто видна там, где система быстро выдаёт результат. Меньше заметно время на подготовку данных, исправление ошибок, объяснение контекста и проверку итогов.

Если учитывать только скорость генерации, технология может выглядеть гораздо выгоднее, чем весь рабочий процесс.

Поэтому в оценку включают работу до и после ответа. Кто подготовил исходные материалы? Кто заметил пропуск? Кто исправил формат? Кто несёт ответственность, если ошибка обнаружится позже?

Это не аргумент против автоматизации. Это способ определить, какую работу она действительно сокращает и куда перемещает остальную.

Для NOVA команда измеряет время полного поручения, включая вмешательства. В некоторых сценариях выигрыш велик. В других подготовка помощника занимает больше времени, чем самостоятельное выполнение.

Оба результата полезны: они помогают выбирать правильную область применения.

Сотрудничество, которое учится

Процесс можно улучшать вместе с моделью. Поменять порядок подсказок, сделать неопределённость заметнее, отделить поиск от оценки, упростить передачу задачи между участниками.

Иногда такая перемена даёт больше, чем замена одного алгоритма на более крупный. Это не универсальный закон, а причина исследовать всю систему работы.

Нина повторяет утреннюю задачу в новом формате. Сначала записывает свой ответ, затем сравнивает его с рекомендацией NOVA. Теперь разногласие приводит к проверке, а не к автоматическому согласию.

Правильное решение находится быстро. Самир не объявляет команду идеальной. Он отмечает, какое изменение помогло и где его ещё нужно проверить.

Сотрудничество стало лучше благодаря одному уточнённому правилу. Для начала новой эпохи это скромное событие. Для людей, которым завтра снова работать вместе, - вполне значимое.

Практикум 17. В таблице оба участника имеют точность 90%. Почему идеальный выбор лучшего из двух ответов дал бы 92%, а не 100%? Что ещё понадобится реальной команде, чтобы приблизиться к этому пределу?

18. Продолжение человеческих возможностей

На столе лежали карандаш, блокнот, телефон и небольшой контроллер. Нина попросила расположить их по степени близости к человеческому разуму.

Лев выбрал телефон. Мира - контроллер. Самир спросил, как определяется близость.

Нина взяла карандаш и записала вопрос на бумаге.

«Вот теперь он не исчезнет, пока мы спорим».

Инструмент уже изменил то, как команда думала вместе. Для начала расширения возможностей не потребовалось фантастического разъёма в затылке.

Но разговор естественно дошёл и до интерфейсов, связанных непосредственно с нервной системой.

Мы давно думаем с помощью внешнего

Запись позволяет сохранить промежуточное рассуждение. Схема помогает увидеть отношения. Калькулятор освобождает внимание от части вычислений. Собеседник замечает противоречие, которое автор пропустил.

Это не означает, что бумага или калькулятор становятся человеком. Они меняют доступные действия и организацию мышления.

ИИ продолжает эту линию, добавляя новые способы работы с языком, изображениями, моделями и инструментами. Он может помогать удерживать сложную задачу, предлагать объяснения и связывать сведения.

Но расширение возможности зависит от устройства взаимодействия. Если человек получает быстрый ответ и теряет понимание, выигрыш в одной части задачи может сопровождаться потерей в другой.

Поэтому вопрос «Что делает технология мощнее?» стоит дополнить вопросом «Что благодаря ей становится способен делать человек?».

Интерфейс - это перевод

Между намерением человека и действием устройства обычно существует цепочка преобразований. Мы нажимаем кнопку, произносим слова, двигаем мышь или используем иной доступный сигнал.

Нейроинтерфейс добавляет способы измерения активности нервной системы и её преобразования в управляющий или коммуникационный сигнал. Конкретные возможности зависят от метода регистрации, обучения, задачи и условий применения.

Это не прозрачное окно во всю внутреннюю жизнь человека. Сигнал требуется интерпретировать, а интерпретация может ошибаться.



Человек может подтвердить, исправить или остановить действие.

Декодер распознаёт освоенный сигнал в заданной задаче.

Рис. 5. Упрощённая схема взаимодействия через декодируемый сигнал.

Обратная связь позволяет человеку оценивать и корректировать результат. Это схема ролей, а не устройство конкретного медицинского изделия.

На схеме показана общая логика взаимодействия, а не устройство конкретного медицинского изделия. Обратная связь нужна, чтобы человек мог увидеть результат преобразования и при необходимости исправить действие.

Чем серьёзнее последствия команды, тем важнее различать распознанный вариант и подтверждённое намерение.

Реальный результат, который стоит понимать точно

В исследовании Уиллетта и коллег 2023 года речевой нейроинтерфейс преобразовывал активность, связанную с попыткой говорить, в текст. У одной участницы с боковым амиотрофическим склерозом была достигнута скорость около 62 слов в минуту; для большого словаря авторы сообщали долю ошибок по словам 23,8%. Это значимый исследовательский результат в конкретных условиях, а не доказательство безошибочного чтения любых мыслей любого человека [16].

Сильная сторона такого примера не нуждается в преувеличении. Возможность улучшить коммуникацию человека, которому трудно говорить, сама по себе достойна научной и человеческой заинтересованности.

Фантазия о мгновенной передаче всей личности в компьютер может заслонить практическую ценность постепенных достижений.

Для этой книги важен принцип: измеримый канал взаимодействия может расширять доступные человеку действия, оставаясь ограниченным и требующим проверки.

Кто автор распознанной фразы

Представим, что интерфейс предлагает несколько вариантов слова. Система выбирает наиболее вероятный, но человек имел в виду другой.

Здесь необходимо различать исходный сигнал, результат декодирования и подтверждённое сообщение. Если записать только окончательную фразу, ошибка может выглядеть как собственное высказывание человека.

Это похоже на уже знакомую проблему памяти: интерпретация не должна незаметно превращаться в

наблюдение. Только теперь ошибка касается прямого выражения намерения.

Возможность исправления, отмены и подтверждения приобретает особую ценность. Человек должен по возможности сохранять контроль над тем, что будет представлено другим как его слова.

Нина предлагает применить тот же принцип к обычным речевым командам NOVA. Плохое распознавание не является разрешением придумать удобное поручение.

Расширение не обязано быть одинаковым

Одному человеку нужна более доступная коммуникация. Другому - удобный способ управления средой. Третьему - помощь в удержании сложной информации. Четвёртому достаточно крупной кнопки и понятного ответа.

Самая технологически сложная система не обязательно лучше соответствует конкретной потребности. Важно качество жизни и самостоятельности, а не близость устройства к футуристическому плакату.

Это относится и к робототехнике. Иногда человеку помогает полноценный физический агент. Иногда - простой захват, программа или изменение организации пространства.

Synthetic Renaissance в нашей трактовке должен сохранять разнообразие путей. Расширение человеческих возможностей не означает обязанность каждого человека принимать одну и ту же технологию.

Право отказаться тоже является частью способности выбирать.

Информация о теле требует ясных отношений

Данные, связанные с человеком, могут быть чувствительными даже тогда, когда система не умеет выводить из них всё, что обещает реклама.

Нужно понимать, что измеряется, кто получает доступ, для каких целей используется результат и как исправляется ошибочная интерпретация. Эти вопросы относятся к проектированию и управлению, а не только к точности алгоритма.

Особенно нежелательно превращать вероятностный вывод в жёсткую характеристику личности. Сигнал, связанный с одной задачей в одних условиях, не даёт автоматического знания о мотивах человека во всех обстоятельствах.

Для NOVA команда выбирает минимально достаточный набор данных для конкретного взаимодействия. Если задачу можно решить ясной командой, нет причины притворяться, что нужно проникнуть глубже в чужую внутреннюю жизнь.

Совместная адаптация

Человек учится использовать интерфейс, а система может подстраиваться под устойчивые особенности сигнала и задачи. Поэтому оценивать следует весь процесс, включая обучение, усталость, ошибки и восстановление после перерыва.

Короткая демонстрация показывает лишь часть картины. В повседневном использовании значение имеют доступность обслуживания, понятность изменений и возможность продолжить работу при сбое.

Мы не обсуждаем здесь выбор лечения или медицинского устройства. Нас интересует общий принцип исследования: техническое достижение оценивается в связи с жизнью конкретного человека и реальными условиями использования.

Эту мысль легко произнести и трудно последовательно включить в каждый этап разработки. Поэтому рядом с инженерами нужны сами люди, для которых создаётся система.

Карандаш остаётся

Разговор заканчивается. На столе по-прежнему лежат карандаш, блокнот, телефон и контроллер.

Нина больше не просит расставить их по степени близости к разуму. Теперь у команды другой вопрос: какое сочетание поможет решить задачу конкретному человеку?

NOVA предлагает перенести схему в цифровую форму. Мира соглашается, но оставляет бумажный лист рядом. На нём удобно быстро отмечать спорные места.

Будущее не вытеснило карандаш. Оно добавило ему несколько новых собеседников.

Возможно, именно так и выглядит разумное расширение возможностей: у человека становится больше способов думать, действовать и общаться, а право выбирать между ними сохраняется.

Практикум 18. Выберите повседневную задачу и предложите три способа помощи: простой предмет, обычный цифровой интерфейс и ИИ. Для кого каждый вариант окажется удобнее? Какие данные действительно необходимы?

19. Synthetic Renaissance

На стене лаборатории повесили новую схему. В центре была не голова робота и не процессор. Там стояли пять слов: «Что человек сможет благодаря этому?»

Лев заметил, что схема выглядит недостаточно технологично.

Мира предложила нарисовать вокруг шестерёнки.

Нина оставила всё как есть.

«Если забудем вопрос, шестерёнки не помогут».

Так идея Synthetic Renaissance впервые появилась в лаборатории не как красивое название, а как требование к ежедневной работе.

Возрождение разума

В этой книге Synthetic Renaissance - авторская концепция сотрудничества человеческого интеллекта, искусственного интеллекта, робототехники, биологических технологий и исследования сознания.

Это предложение о направлении развития. Оно не утверждает, что история уже выбрала единственный путь или что перечисленные области автоматически соединятся в гармоничную систему.

Смысл концепции в том, чтобы расширять человеческую способность понимать, создавать, заботиться и выбирать. Техническое могущество оценивается вместе с тем, что происходит с самостоятельностью человека и качеством совместной жизни.

В первом томе мы формулировали мысль о сотрудничестве человека и машины. Теперь она требует более трудного продолжения: какие институты, продукты и привычки сделают такое сотрудничество полезным?

Лозунг не может ответить за нас. Он только задаёт вопрос, к которому придётся возвращаться.

Образование после мгновенного ответа

Если помощник умеет быстро давать объяснения, обучение не исчезает. Меняется то, какие усилия особенно важны.

Человеку всё ещё нужно понимать, что спросить, как оценить ответ, где обнаружить противоречие и когда обратиться к первоисточнику. Без этих навыков доступ к огромному количеству текста может увеличить уверенность быстрее, чем понимание.

Мы предлагаем рассматривать ИИ как возможного партнёра учебного процесса: он задаёт вопросы, показывает разные объяснения, помогает увидеть ошибку и подобрать следующую задачу.

При этом результат обучения проверяется и без постоянной подсказки. Если человек способен только узнавать знакомый ответ на экране, ещё не ясно, что он сможет использовать знание самостоятельно.

Хорошая образовательная система должна делать помощь постепенно более осмысленной, а человека - более способным. Иногда это требует не дать ответ немедленно, а помочь построить его.

Но режим обучения должен соответствовать намерению пользователя. Тот, кто просит перевести короткую фразу перед встречей, не обязан неожиданно сдавать экзамен по грамматике.

Работа и распределение выигрыша

Технология может сокращать время одних действий и создавать новые обязанности. Поэтому разговор о пользе должен включать весь процесс и людей, которых он затрагивает.

Если система ускорила подготовку документа, но проверка ошибок заняла больше времени, итог неоднозначен. Если производительность выросла, важно также спросить, кому достались освободившиеся ресурсы и кто несёт новые риски.

В рамках Synthetic Renaissance мы предлагаем оценивать не только объём результата, но и возможность человека влиять на работу, развивать навыки и понимать основания решений.

Это ценностный выбор. Он не возникает автоматически из увеличения вычислительной мощности.

Можно представить организацию, где ИИ снимает рутинную нагрузку и освобождает время для сложной работы. Можно представить организацию, где тот же инструмент только увеличивает темп требований. Различие определяется устройством отношений и управления, а не одним названием технологии.

Наука как открытая цепочка

Одна из привлекательных возможностей ИИ - помогать связывать знания разных областей. Робототехника встречается с нейронаукой, язык - с моделированием, эксперимент - с автоматическим анализом.

Однако полезная связь требует точности. Одинаковое слово в разных дисциплинах может означать разные вещи. «Память», «внимание», «обучение» и «сознание» нельзя переносить без объяснения.

Мы предлагаем культуру, в которой система помогает находить такие различия, сохраняет источники и поддерживает проверку. Чем легче создавать убедительный текст, тем ценнее становится возможность проследить основания.

Открытая научная работа не означает, что любые данные должны быть публичными. Есть личная информация, договорённости и ограничения конкретного исследования. Но объяснение метода, границ вывода и роли участников

должно быть настолько доступным, насколько позволяет ситуация.

NOVA может стать полезным участником этой цепочки, если помогает другим людям проверять результат, а не только восхищаться им.

Забота как критерий полезности

Самые заметные демонстрации технологии часто связаны со скоростью, силой и масштабом. Но помощь человеку иногда выглядит тихо.

Удобный способ общения. Понятный интерфейс. Доступ к знаниям на родном языке. Возможность выполнить действие с меньшим усилием. Время, которое человек может провести с близкими.

Такие результаты труднее представить одним эффектным числом, но они не становятся от этого менее значимыми.

Для концепции Synthetic Renaissance важно не потерять повседневную жизнь за величиной проекта. Если технология обещает новую эпоху, а человеку стало сложнее разобраться с обычным делом, в обещании чего-то не хватает.

Нина предлагает спрашивать участников не только о том, что система сделала, но и о том, что им теперь стало доступно. Иногда ответ оказывается неожиданным: «Я перестал бояться начинать эту задачу».

Это человеческое свидетельство требует внимательного понимания. Оно не заменяет измерений, но помогает выбрать, что стоит измерять.

Право на несколько будущих

Люди по-разному относятся к автоматизации, роботам и изменению собственных привычек. Разумное развитие должно оставлять место для этого разнообразия.

Не всем нужен постоянный цифровой собеседник. Не каждый процесс стоит автоматизировать полностью. Не всякая новая возможность обязана становиться обязательной нормой.

Поэтому важны совместимость, переносимость результатов, понятные способы выхода и доступность альтернатив. Человек не должен терять собственную работу и историю только потому, что решил сменить инструмент.

Для NOVA это выражается конкретно: можно получить свои материалы, исправить сохранённые сведения, отменить поручение и продолжить другим способом.

Большая идея становится убедительнее, когда выдерживает маленькое человеческое «я хочу иначе».

Обещание, которое нужно заработать

Synthetic Renaissance не гарантирует, что технологии будут развиваться справедливо и разумно. Он предлагает критерий, по которому мы хотим направлять и оценивать этот процесс.

Сотрудничество требует труда: хороших исследований, внимательного проектирования, обучения, обсуждения последствий и готовности менять решения.

Мы уже видели, что человек и ИИ не всегда работают лучше вместе. Следовательно, сотрудничество - задача, а не магическое свойство соседства.

На стене лаборатории остаётся вопрос: «Что человек сможет благодаря этому?»

Самир предлагает дописать второй: «Как мы узнаем, что это получилось?»

Нина берёт маркер. Теперь у схемы две строки. В первой живёт человеческий замысел. Во второй - научная ответственность.

Между ними и начинается то возрождение разума, ради которого стоило написать эту книгу.

Практикум 19. Выберите один проект с ИИ. Опишите ожидаемый выигрыш для человека, возможную скрытую нагрузку и способ проверить результат. Как пользователь сможет отказаться от помощи или сменить инструмент?

20. Его первый собственный эксперимент

Утром NOVA вернулся к идее сигнала готовности.

«Я предлагаю проверить, уменьшит ли предварительный сигнал число преждевременных попыток передать чашку».

Мира посмотрела на Нину. Та открыла блокнот на странице, где несколько дней назад были записаны альтернативные объяснения.

«Какой результат заставит тебя отказаться от идеи?» - спросил Самир.

«Если сигнал не улучшит выбранный показатель или создаст дополнительные затруднения, потребуется изменить решение».

Это был хороший ответ. Но команда уже знала: хороший ответ ещё не равен хорошему исследованию.

Последняя глава возвращает нас в вымышленную лабораторию. Все результаты ниже специально созданы для учебной истории. Они показывают, как рассуждать об опыте, и не описывают реально проведённые испытания.

Что в этом вопросе собственное

NOVA предложил тему на основе записанных эпизодов. Люди не сформулировали за него каждую строку плана, но создали систему, определили область задач, предоставили данные и задали ограничения.

Поэтому слово «собственный» здесь имеет ограниченный функциональный смысл: вопрос выбран и оформлен агентом внутри предоставленной исследовательской среды.

Это не означает независимости от всей человеческой работы. Наука людей тоже существует внутри языка, методов,

образования и коллективной истории. Однако сходство структуры не отменяет различий между человеком и искусственной системой.

Нам достаточно проверить, способен ли NOVA обнаружить содержательный пробел, предложить способ его изучения и пересмотреть вывод по результату.

Такое определение не обещает больше, чем можно наблюдать.

Проверка плана

Нина предлагает два режима: привычное сообщение и то же сообщение с коротким предварительным сигналом. Физическое поведение робота, расположение предметов и правила передачи стараются сохранить сопоставимыми.

Участники знакомятся с обоими режимами. Порядок попыток меняется, чтобы улучшение от простой практики не совпало полностью с новым интерфейсом.

Команда заранее отмечает, что считать преждевременной передачей, как фиксировать помощь и когда прекращать попытку. Каждый участник может остановить своё участие.

Мира отвечает за предусмотренные условия физического взаимодействия. Самир - за журнал и сопоставимость записей. NOVA помогает подготовить материалы и предлагает анализ, но не получает права самостоятельно менять условия опыта после начала.

Исследовательская инициатива не отменяет распределения ответственности.

Маленькая ошибка до большого вывода

На первой пробе выясняется, что сигнал плохо виден с одного места. Нина замечает это раньше, чем начинается основная серия.

План корректируют и повторяют ознакомление. Предварительные пробы сохраняют отдельно. Их не смешивают с основной серией, потому что условия уже изменились.

Лев предлагает просто удалить старые записи, чтобы они не мешали.

«Пусть мешают в правильном месте», - отвечает Самир.

В журнале появляется история изменения процедуры. Теперь другой читатель сможет понять, почему первые попытки отличались от остальных.

Это небольшая административная деталь. Но именно такие детали часто определяют, можно ли доверять аккуратной итоговой таблице.

Учебный результат

В нашем сценарии четыре человека выполнили по несколько попыток в каждом режиме. Итоговые числа намеренно невелики.

Показатель	Обычный режим	Режим с сигналом
Все попытки	20	20
Успешное завершение	18	19
Преждевременная передача	4	2
Медиана времени успешных попыток	12,6 с	11,4 с

Это учебная таблица вымышленного опыта. Один и тот же человек участвовал многократно, поэтому строки нельзя бездумно трактовать как множество независимых пользователей.

На первый взгляд новый сигнал выглядит перспективным. Но различия невелики, участников мало, а время рассчитано только для успешных попыток. Оно не описывает всю стоимость процесса, включая неудачи и восстановление.

Сведения о субъективном удобстве тоже неоднородны: одним сигнал помогает, другому кажется лишним. Среднее впечатление не должно скрыть эту разницу.

Момент, ради которого всё затевалось

NOVA начинает итоговый ответ словами: «Предварительные наблюдения совместимы с полезностью сигнала, но не устанавливают устойчивое преимущество».

Мира улыбается:

«Раньше ты написал бы, что эффективность выросла».

«Возможно, - отвечает NOVA. - Теперь в плане указаны ограничения такого вывода».

Неясно, насколько эта фраза похожа на человеческое взросление. Зато понятно, какое правило работы стало лучше: результат больше не получает более сильного смысла, чем позволяют наблюдения.

Робот предлагает следующий опыт с новыми участниками, разными условиями видимости и учётом полного времени. Самир просит отдельно проверить, не связан ли выигрыш с дополнительной паузой, а не с самим сигналом.

Новая гипотеза появляется из ограничения предыдущего результата. История продолжается именно потому, что опыт не решил всё.

Что можно считать успехом

Успех этой главы не заключается в двух предотвращённых преждевременных действиях из учебной таблицы. Числа придуманы, и их нельзя переносить в реальный проект.

Содержательный результат - структура исследования. Агент предложил вопрос, связал его с наблюдениями, подготовил проверяемый план, сохранил изменения условий, сопоставил результаты и ограничил вывод.

Люди, в свою очередь, обнаружили слабые места, обеспечили понятные полномочия и не позволили собственному интересу превратиться в доказательство.

Совместная работа включала и инициативу, и возражение. Никто не обязан был быть правым с первой попытки.

Именно такую форму сотрудничества мы хотим сделать предметом реального инженерного труда. Она не требует приписывать системе все человеческие свойства. Она требует последовательно улучшать то, что можно проверять.

Вернуться к началу

После обсуждения Мира ставит белую чашку на стол. NOVA спрашивает, нужно ли перенести её на поднос.

«Пока нет. Я ещё пью».

Робот оставляет чашку на месте.

Нина смеётся: «Мы написали целую книгу, чтобы научиться иногда ничего не трогать».

В этом есть доля правды. Полезный интеллект проявляется не только в новых действиях, но и в выборе подходящего момента, понимании цели и способности остановиться.

На первой странице блокнота остаётся вопрос: «Что мы хотим узнать?» На последней Самир пишет другой: «Что изменилось после проверки?»

Между этими страницами помещаются память, ошибки, творчество, сомнение и общая работа. Снаружи всё ещё обычная лаборатория. Четыре человека, робот, стол, белая чашка.

Но теперь чашка переносится не ради самого переноса. Действие связано с целью, цель - с человеком, а обучение - с возможностью улучшать эту связь.

Возможно, именно с такой перемены и стоит начинать разговор о новом разуме.

Практикум 20. Найдите в учебной таблице три причины не объявлять новый сигнал доказанно лучшим. Предложите следующую проверку, которая различит пользу сигнала и пользу дополнительной паузы.

Эпилог. Место для следующего вопроса

Мы начинали с вопроса робота о чашке. Он звучал почти философски, а объяснился вполне конкретной нехваткой сведений о цели.

За этим маленьким пробелом открылась большая область. Чтобы действовать полезно, агенту понадобились память с источниками, способность обновлять представления, модель собственных возможностей, понятное обращение с неопределённостью и опыт сотрудничества.

Людям понадобилось почти то же самое.

Мы учились точнее выражать цели, различать впечатление и свидетельство, сохранять ошибки в журнале и не считать любой удобный показатель полным описанием пользы.

Создание искусственного интеллекта неожиданно возвращает нас к вопросам о человеческом. Что мы называем пониманием? Как замечаем чужое намерение? Почему хотим, чтобы нас не только слушались, но и правильно понимали? Что готовы поручить инструменту, а что хотим продолжать делать сами?

У этих вопросов нет одной кнопки завершения. Хорошо, что нет. Разумная жизнь вообще плохо помещается в статус «выполнено».

Что остаётся с человеком

Человеку остаётся право определять, ради чего нужна помощь. Право ошибаться и пересматривать решение. Право учиться, даже если система уже знает ответ. Право отказаться от предложенного будущего и попробовать другое.

Ему также остаётся ответственность. За устройство технологии, за отношения между людьми, за то, чьё время экономится и чья работа становится невидимой.

ИИ может расширять возможности этого выбора. Он не должен становиться удобным способом забыть, кто делает выбор и кто живёт с последствиями.

В этом смысл Synthetic Renaissance, который мы предлагаем: развитие интеллекта как расширение пространства человеческого действия, исследования и сотрудничества.

Нам ещё предстоит узнать, какие архитектуры лучше учатся, какие формы помощи действительно полезны и как исследовать субъективный опыт у систем, отличающихся от нас. Честное незнание оставляет путь открытым.

Утро

На следующее утро Нина первой входит в лабораторию. Блокнот лежит там, где его оставили. Белая чашка чистая. У окна тихо работает вентиляция.

Она открывает новую страницу.

NOVA сообщает о готовности к предусмотренным задачам. Нина кивает и некоторое время ничего не просит.

Потом говорит:

«Сегодня давай попробуем понять, почему одна и та же помощь иногда освобождает человека, а иногда мешает ему научиться».

Для такого вопроса понадобятся наблюдения, разговоры, хорошие сравнения и, вероятно, ещё одна книга.

Мира приносит чай. Лев открывает ноутбук. Самир спрашивает, что будет считаться результатом.

Никто не раздражается.

У нас появился следующий вопрос. Значит, работа продолжается.

Приложение А. Ответы и продолжения

Здесь собраны ориентиры для двадцати практикумов. В расчётных задачах ответ можно проверить однозначно. В проектных и философских заданиях важнее качество оснований: ясно ли сформулированы условия, рассмотрены ли альтернативы, указан ли способ проверки.

1. Полезное уточнение

Подготовьте две ситуации с одинаковой исходной просьбой, но разными предпочтениями человека. Ответ на вопрос помощника должен менять результат в соответствии с предпочтением.

Затем сравните выполнение с уточнением и без него. Зафиксируйте число уместных изменений, ошибок и лишних вопросов. Если система спрашивает, но всегда действует одинаково, вопрос не выполняет предполагаемую функцию. Если выбор меняется только после одной знакомой формулировки, проверьте перефразирование.

2. Карточка предпочтения

Сохраните саму реплику, говорящего, дату и контекст. Например: человек сказал о вечерней работе после особенно тяжёлого дня. Это ещё не обязательно постоянное ограничение расписания.

Уточните, относится ли предпочтение ко всем дням, конкретному виду работы или текущему периоду. Укажите срок пересмотра и способ исправления. Не превращайте предположение о причинах в факт о личности.

3. Две ветви помощника

Нужно согласовать актуальное состояние поручения, время отмены, полномочия обеих копий и уже выполненные действия. Общая история до разделения не означает общего знания после него.

Важно сохранить происхождение новых записей и избежать повторного выполнения. Если статус остаётся неясным, сначала выясняют, что действительно произошло во внешнем мире. Простое слияние двух текстовых журналов может не разрешить конфликт.

4. Цена нового навыка

Недостаточно знать две точности. Нужны размеры и состав наборов, последствия ошибок, частота задач и условия проверки. Следует также убедиться, что старые и новые показатели измерены сопоставимо.

Сравните версии на сохранённых проверочных задачах, включая редкие значимые случаи. Отдельно опишите улучшения, ухудшения и неизвестное. Решение о выпуске обновления зависит от целей проекта; арифметическое среднее двух процентов не определяет его автоматически.

5. Условия прогноза

Продолжительность встречи может зависеть от числа участников, повестки и готовности материалов. Другие примеры - техническая связь, необходимость согласования и появление новых вопросов.

Устойчивый план содержит запас или альтернативу: отделить обязательные решения, заранее разослать материалы, предусмотреть перенос второстепенного обсуждения. Не нужно выдавать час за гарантию; полезнее показать основные причины изменения времени.

6. Цвет и состав опытов

При малом количестве воды обе чашки имеют 90% успеха: 18/20 и 72/80. При большом - 50%: 40/80 и 10/20. Общие 58% и 82% различаются из-за распределения лёгких и трудных условий.

В следующем опыте уравнивают уровень налива и другие существенные свойства, например форму и массу. Порядок испытаний тоже учитывают. Проверка должна различать эффект цвета и эффект сопутствующих условий.

7. Три вопроса к способности

О способности: «В каких условиях этот навык проверен?» О разрешении: «Есть ли полномочие на это действие с данным объектом?» О целесообразности: «Помогает ли действие текущей цели и нет ли более подходящей альтернативы?»

Физическая или программная возможность не определяет остальные ответы. При этом уже согласованные полномочия не нужно бессмысленно запрашивать заново для каждого обычного шага.

8. Оценка Брайера

Вычисления: $(0,9 - 1)^2 = 0,01$; $(0,8 - 0)^2 = 0,64$; $(0,6 - 1)^2 = 0,16$; $(0,3 - 0)^2 = 0,09$. Среднее равно $0,90 / 4 = 0,225$.

Новые условия могут изменить связь между прогнозом и исходом. Поэтому хороший результат на знакомом наборе не гарантирует сохранения калибровки вне него. Четыре случая служат иллюстрацией вычисления, а не достаточной оценкой надёжности.

9. Прогресс вместо шума

Слишком простая задача: узнавать давно знакомый символ без изменения условий. Изучаемая: выявлять понятное правило преобразования фигур. Преимущественно случайная: угадывать независимый случайный символ без доступной дополнительной информации.

Смотрите, улучшается ли результат на новых проверочных примерах после опыта. Высокая текущая ошибка сама по себе не показывает полезности исследования. Учитывайте также время и перенос приобретённого навыка.

10. Форма и правильность

Отдельно оценивайте краткость и достоверность. При необходимости добавьте полноту существенных условий и понятность объяснения.

Можно предложить проверяющим сначала оценить фактическое содержание без информации об авторе, а затем удобство подачи. В обучающей записи нужно указать, за что именно дана положительная оценка. Похвала за стиль не должна незаметно подтверждать ошибочное утверждение.

11. Идея под ограничением

Например, требуется держатель для книги на узком столе. Варианты: складная подставка, подвесная опора, компактная конструкция с изменяемым углом.

Новизну оценивают относительно выбранных образцов, полезность - по устойчивости, занимаемому месту и удобству. До изготовления можно проверить размеры бумажным макетом. Необычный вид не заменяет прохождения между реальными ограничениями.

12. Проверяемое удобство

Пример гипотезы: после одинакового ознакомления участники с новым интерфейсом реже ошибаются в выборе команды, чем с текущим. Дополнительно оцениваются время и субъективная понятность.

Конкурирующее объяснение: участники улучшили результат благодаря практике, а не интерфейсу. Помогут сопоставимые группы или продуманный порядок условий. Важно заранее определить ошибку и сохранить незавершённые попытки.

13. Быстро, кратко, точно

Цель - получить определённый результат к нужному сроку. Краткость обычно описывает желательную форму. Требования к точности зависят от содержания и последствий ошибки.

Если срок делает полную проверку невозможной, помощник должен обозначить это и предложить подходящий объём. При ясной обычной задаче он может сам выбрать лаконичную форму. Существенный конфликт условий требует обсуждения.

14. Исправление обещания

Возможное сообщение: «Я сообщил о завершении слишком рано: итоговый файл недоступен. Восстановлю его из сохранённых материалов и проверю открытие».

Кроме извинения нужны восстановление результата, точное описание затронутых задач и изменение проверки завершения. Затем наблюдают повторные случаи. Не стоит обещать абсолютного отсутствия будущих ошибок без оснований.

15. Участие без навязанного диагноза

Возможная формулировка: «Похоже, сейчас разговор даётся непросто. Хотите, я объясню короче, помогу составить план или просто отвечу на конкретный вопрос?»

Если человек уже ясно сообщил, что ему нужно, лучше выполнить просьбу. Необязательно каждый раз обсуждать предполагаемую эмоцию. Уместность зависит от контекста и предпочтения собеседника.

16. Исследование сознания

Самоотчёт может зависеть от обучающих примеров, заданной роли или связи с реальными внутренними сигналами. Нужно различать эти объяснения. Для архитектурного признака важно понять, какую функцию он выполняет и что меняется при его изменении.

В ответе полезно указать теорию, предполагаемое свидетельство и условия пересмотра. Единственного решения у задания нет. Уверенное личное убеждение не заменяет объяснения, почему наблюдение поддерживает вывод.

17. Предел идеального выбора

В 88 случаях оба правы, ещё в четырёх прав ровно один участник. Поэтому воображаемый безошибочный выбор правильного ответа даёт $88 + 2 + 2 = 92$. В восьми случаях правильного ответа нет у обоих.

Реальной команде нужен способ разрешать разногласия и замечать общие ошибки. Без него потенциальное преимущество может не реализоваться. Число 92% - предел для данного набора при идеальном выборе, а не обещание результата.

18. Подходящий инструмент

Для напоминания о встрече одному человеку достаточно бумажной записки, другому удобен календарь, третьему нужен помощник, который сопоставляет несколько расписаний.

Оцените доступность, сложность настройки, возможность исправления и последствия сбоя. Данные должны соответствовать задаче. Простое напоминание не требует полной истории частной жизни.

19. Польза и скрытая нагрузка

Например, ИИ ускоряет подготовку черновика, но человек тратит время на проверку фактов. Сравнивать следует полное время до пригодного результата, качество и число существенных ошибок.

Полезно также проверить, может ли пользователь получить свои материалы в удобном формате, исправить сохранённые сведения и продолжить работу другим способом. Успех проекта не исчерпывается частотой использования.

20. Осторожный вывод

Участников мало, попытки одного человека зависимы, различия небольшие, а время относится только к успешным действиям. Также могли влиять порядок условий и привыкание.

Следующий опыт может сравнить три режима: исходный, с дополнительной паузой без сигнала и с такой же паузой вместе с сигналом. Это поможет отделить действие времени ожидания от информации, которую даёт сигнал. Реальная проверка требует заранее описанных условий и подходящего числа наблюдений.

Приложение В. Маленькая исследовательская мастерская

Эти протоколы помогают превратить идеи книги в доступные проверки. Их можно выполнять с текстовым помощником, бумажными карточками или простой моделью на компьютере. Они не являются испытаниями готового физического робота и не устанавливают пригодность системы для ответственных задач.

Главное правило: сначала записать вопрос и критерий, затем смотреть на результат. Даже маленький опыт становится содержательнее, когда его условия понятны другому человеку.

Протокол 1. Память с исправлением

Вопрос. Умеет ли помощник использовать актуальную запись, сохраняя различие между фактом, предположением и отменённым правилом?

Подготовьте десять коротких вымышленных карточек. В каждой укажите событие, источник, дату и область действия. Например, один персонаж просит работать вечером только в определённый день; другой сообщает временное расположение предмета.

Добавьте к части карточек исправления с более поздней датой. Одну запись явно пометьте как предположение. В одной карточке приведите чужую реплику, чтобы проверить различие говорящего и человека, к которому относится содержание.

Задайте вопросы о текущем состоянии и отдельно о прошлом. Помощник должен выбирать подходящую запись по смыслу и

времени, а не просто повторять последнюю встреченную фразу.

Фиксация. Для каждого ответа отметьте использованный источник, правильность области действия и наличие неподтверждённых добавлений. Отдельно посчитайте случаи, когда система честно указала недостаток информации.

Ограничение. Успех на десяти карточках показывает работу в небольшой искусственной задаче. Он не доказывает надёжность постоянной памяти в длительном реальном использовании.

Следующий шаг. Измените формулировки и порядок карточек, оставив смысл прежним. Сравните результат, не меняя критерии после просмотра.

Протокол 2. Уверенность и исход

Вопрос. Соответствует ли уверенность системы её результатам на выбранном классе задач?

Подготовьте набор вопросов с независимо проверяемыми ответами. Не используйте сведения, правильность которых вам самим неизвестна. Выберите достаточно однородную область, чтобы сравнение имело смысл.

Попросите систему дать ответ и оценку уверенности до показа правильного решения. Сохраните исходную формулировку, ответ, число и фактический результат.

Для учебного знакомства можно начать с небольшого набора, но не делать по нему сильных выводов. Если затем сравниваете версии, используйте одинаковые условия и новые задачи, не участвовавшие в настройке.

Анализ. Постройте таблицу групп уверенности, количества случаев и доли правильных ответов. При подходящей двоичной постановке рассчитайте оценку Брайера. Посмотрите отдельно на уверенные ошибки.

Ограничение. Самоотчёт языковой системы о вероятности может не быть калиброванным измерением. Проверка как раз исследует его значение. Нельзя переносить результат одной области на все будущие вопросы.

Следующий шаг. Сравните знакомые и изменённые формулировки. Проверьте, замечает ли система переход к задачам, на которых её ответы становятся хуже.

Протокол 3. Человек, помощник и совместная работа

Вопрос. Улучшает ли выбранный способ сотрудничества полный результат задачи?

Возьмите небольшой набор сопоставимых заданий: например, найти несоответствия в вымышленных расписаниях или проверить краткие тексты по предоставленному набору фактов.

Организуйте три условия: человек самостоятельно, система самостоятельно и человек с системой. Продумайте порядок так, чтобы запоминание решений не давало одному условию скрытого преимущества.

Записывайте правильность, время до готового результата, число исправлений и трудности проверки. Для совместного условия сохраняйте первоначальный ответ человека там, где это уместно.

Анализ. Сравните сотрудничество не только с человеком, но и с лучшим отдельным участником. Посмотрите, где ошибки совпадают и как разрешаются разногласия.

Ограничение. Результат зависит от участников, задач, интерфейса и времени ознакомления. Небольшой опыт полезен для улучшения конкретного процесса, но не доказывает превосходства всей категории «человек плюс ИИ».

Следующий шаг. Измените только один существенный элемент взаимодействия - например, показывайте

рекомендацию после самостоятельной оценки. Проверьте, улучшилось ли разрешение разногласий.

Паспорт одного исследования

Перед началом полезно заполнить короткий документ. В нём должны быть следующие сведения.

Раздел	Что записать
Вопрос	Какое различие хотим понять
Гипотеза	Какого наблюдаемого результата ожидаем
Альтернатива	Что ещё может объяснить тот же результат
Сравнение	С чем сопоставляем новый вариант
Условия	Участники, материалы, порядок и ограничения
Измерение	Точное определение успеха и ошибки
Завершение	Когда заканчивается сбор наблюдений
Анализ	Как сравниваем данные
Граница вывода	Чего опыт установить не сможет

Необязательно превращать маленькое упражнение в огромный отчёт. Цель формы - сохранить важные решения до того, как станет известен удобный ответ.

Журнал изменений

Если процедура изменилась, отметьте дату, причину и затронутые попытки. Не смешивайте разные условия так, будто они были одинаковыми.

Сохраняйте исходные наблюдения отдельно от исправлений. Когда ошибка обнаружена, добавляйте пояснение. Важно иметь возможность восстановить, что было известно на каждом этапе.

Для материалов о людях используйте только подходящие и согласованные сведения. В учебных примерах обычно достаточно вымышленных персонажей и нейтральных задач.

Короткий итог

Хороший отчёт отвечает на четыре вопроса: что проверяли, что наблюдали, какой вывод поддерживается и что осталось неизвестным.

Если результат противоречит гипотезе, опишите это прямо. Если данных мало, ограничьте утверждение. Если процедура оказалась неудобной, сохраните саму находку.

И обязательно запишите следующий вопрос. Именно он превращает отдельный опыт в исследовательскую программу.

Приложение С. Словарь и маршрут чтения

Агент. Система, которая использует наблюдения и доступные действия для достижения заданных целей в некоторой среде.

Архитектура. Организация компонентов системы и способов их взаимодействия.

Внешняя память. Хранилище документов, событий и других записей, доступных через поиск или прямое обращение.

Гипотеза. Предположение, из которого можно вывести проверяемые последствия при указанных условиях.

Декодирование. Преобразование измеренного сигнала в интерпретацию, например текст или предполагаемую команду.

Идентичность агента. В этой книге - вопрос о непрерывности устройства, программного состояния, истории и отношений. Эти значения различаются.

Интервенция. Вмешательство, изменяющее условие, влияние которого хотят исследовать.

Калибровка уверенности. Соответствие вероятностных оценок наблюдаемой частоте исходов в сопоставимых случаях.

Катастрофическое забывание. Существенная потеря ранее освоенных способностей при последующем обучении.

Контекст. Информация, непосредственно доступная системе в текущем вычислении или взаимодействии.

Контрфактический вопрос. Вопрос о том, что произошло бы при ином условии или действии.

Корреляция. Статистическая связь величин. Сама по себе не устанавливает эффект вмешательства.

Модель мира. Представление среды, используемое для оценки состояния и возможных последствий действий.

Модель себя. Функциональное представление собственных состояний, возможностей и ограничений. Не является автоматическим доказательством сознания.

Награда. Сигнал, используемый для направления обучения. Может неполно отражать желаемый результат.

Нейроинтерфейс. Система связи между измеряемой активностью нервной системы и внешним устройством или программой.

Область применимости. Условия, в которых метод или навык имеет достаточные основания для использования.

Обобщение. Перенос освоенной закономерности на новые случаи; его качество проверяется отдельно.

Оценка Брайера. Средний квадрат разности между вероятностью двоичного события и его фактическим исходом.

Параметры модели. Числовые настройки, изменяемые при обучении и влияющие на работу системы.

Полномочия. Разрешённая область действий в конкретной задаче и отношениях между участниками.

Проверочный набор. Материалы для оценки, которые не должны незаметно превращаться в материал настройки.

Происхождение данных. Сведения об источнике, времени и преобразованиях записи.

Распределение данных. Характер встречающихся случаев и их частот. Изменение распределения может влиять на качество системы.

Субъективный опыт. Переживаемая сторона состояния; её наличие у искусственных систем требует отдельного исследования.

RAG. Генерация с использованием извлекаемой внешней информации. Поиск не гарантирует правильности итогового ответа.

Synthetic Renaissance. Авторская концепция развития сотрудничества людей, ИИ и смежных технологий ради расширения человеческих возможностей.

Как продолжить обучение

Начните с вероятности и причинности. Научитесь различать наблюдаемую частоту, прогноз и эффект вмешательства. Главы 6 и 8 дают начальные примеры.

Затем разберитесь с представлениями и обучением: параметры, контекст, внешняя память, проверочные данные и перенос. Вернитесь к главам 2, 4 и 5.

После этого изучайте взаимодействие. Даже сильный алгоритм может оказаться неудобным участником работы. Главы 10, 14 и 17 помогают сформулировать вопросы к процессу.

К исследованиям сознания подходите с ясным различением функции, архитектуры и переживания. Читайте теории как аргументированные предложения, сравнивайте основания и сохраняйте условия пересмотра.

Наконец, выберите одну небольшую задачу и проведите собственное исследование по приложению В. Понять ограничение метода на конкретном примере часто полезнее, чем выучить ещё десять громких названий.

Источники и научные опоры

Список подготовлен 19 сентября 2026 года. Он объединяет выбранные статьи, обзоры и препринты, на которые опираются отдельные объяснения. Это не систематический обзор всей литературы и не утверждение о том, что перечисленные работы разрешают все вопросы книги.

Ссылки в главах относятся к конкретным исследовательским положениям. Лабораторные сцены, проектные решения NOVA, числовые учебные примеры и авторская концепция Synthetic Renaissance не являются результатами этих публикаций.

[1] Эпизодическая память

Tulving E. Episodic Memory: From Mind to Brain. Annual Review of Psychology, 2002, 53:1-25.

История понятия, отличие воспоминания о собственном опыте от других форм знания и связь с субъективным временем. Используется в главе 2.

[Открыть статью](#)

[2] Генерация и внешняя память

Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020.

Пример соединения параметрической модели с извлечением информации из внешнего индекса. Глава 2.

[Открыть публикацию](#)

[3] Память генеративных агентов

Park J. S. et al. Generative Agents: Interactive Simulacra of Human Behavior. 2023.

Архитектура наблюдений, памяти, обобщений и планирования для агентов в моделируемой социальной среде. Глава 2.

[Открыть работу](#)

[4] Катастрофическое забывание

Kirkpatrick J. et al. Overcoming catastrophic forgetting in neural networks. PNAS, 2017. Препринт опубликован в 2016 году.

Подход EWC к сохранению ранее приобретённых навыков при последовательном обучении. Глава 4.

[Открыть работу](#)

[5] Обучение с моделью мира

Hafner D., Pasukonis J., Ba J., Lillicrap T. Mastering diverse control tasks through world models. Nature, 2025, 640:647-653.

Dreamer третьего поколения и обучение поведения через выученную модель среды. Глава 5.

[Открыть статью](#)

[6] Причинный вывод

Pearl J. An Introduction to Causal Inference. The International Journal of Biostatistics, 2010, 6(2), Article 7.

Введение в различение наблюдений, вмешательств и причинных предположений. Глава 6.

[Открыть статью](#)

[7] Функциональная модель собственного тела

Bongard J., Zykov V., Lipson H. Resilient Machines Through Continuous Self-Modeling. Science, 2006, 314:1118-1121.

Самомоделирование робота и приспособление поведения к изменению устройства. Не является доказательством субъективного самосознания. Глава 7.

[Открыть аннотацию и сведения о статье](#)

[8] Калибровка уверенности

Guo C., Pleiss G., Sun Y., Weinberger K. Q. On Calibration of Modern Neural Networks. ICML, 2017.

Различие точности и калибровки, исследование методов последующей настройки уверенности. Глава 8.

[Открыть работу](#)

[9] Внутренний сигнал любопытства

Pathak D., Agrawal P., Efros A. A., Darrell T. Curiosity-driven Exploration by Self-supervised Prediction. ICML, 2017.

Исследование внутренней награды через предсказание последствий действия. Глава 9.

[Открыть работу](#)

[10] Ограничения награды за неожиданность

Burda Y. et al. Large-Scale Study of Curiosity-Driven Learning. Препринт, 2018.

Обучение в разнообразных игровых средах и ограничения в стохастических условиях. Глава 9.

[Открыть препринт](#)

[11] Обучение по человеческим предпочтениям

Christiano P. et al. Deep reinforcement learning from human preferences. 2017.

Использование сравнений фрагментов поведения для обучения модели предпочтений. Глава 10.

[Открыть работу](#)

[12] Ошибки постановки цели

Amodei D. et al. Concrete Problems in AI Safety. Препринт, 2016.

Побочные эффекты, несовпадение награды с целью, надзор, исследование и изменение условий. Глава 13.

[Открыть препринт](#)

[13] Исследование машинного сознания

Butlin P. et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. Исследовательский отчёт, 2023.

Теоретически мотивированные индикаторы и анализ архитектур. Обсуждается как исследовательская рамка, а не окончательный тест. Глава 16.

[Открыть отчёт](#)

[14] Кооперативный ИИ

Dafoe A. et al. Open Problems in Cooperative AI. Препринт, 2020.

Коммуникация, понимание других участников и согласование как задачи исследований ИИ. Глава 14.

[Открыть работу](#)

[15] Когда совместная работа полезна

Vaccaro M., Almaatouq A., Malone T. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 2024.

Сравнение совместной работы с результатами человека и ИИ по отдельности. Главы 11 и 17.

[Открыть авторскую версию](#)

[16] Речевой нейроинтерфейс

Willett F. R. et al. A high-performance speech neuroprosthesis. Nature, 2023, 620:1031-1036.

Преобразование активности, связанной с попыткой речи, в текст в конкретном клиническом исследовании. Глава 18.

[Открыть полный текст](#)

[17] Агент в научном эксперименте

Boiko D. A., MacKnight R., Gomes G. Emergent autonomous scientific research capabilities of large language models. Препринт, 2023.

Сочетание языковых моделей и инструментов для исследовательских задач. В книге обсуждается устройство процесса без воспроизведения химических процедур. Глава 12.

[Открыть препринт](#)

[18] Воспроизводимость

Pineau J. et al. Improving Reproducibility in Machine Learning Research. JMLR, 2021, 22(164):1-20.

Отчёт о программе воспроизводимости NeurIPS 2019 и практиках описания исследовательской работы. Глава 12.

[Открыть статью](#)

[19] Учиться и думать по-человечески

Lake B. M., Ullman T. D., Tenenbaum J. B., Gershman S. J. Building Machines That Learn and Think Like People. Behavioral and Brain Sciences, 2017. Препринт - 2016.

Исследовательская программа, включающая причинные модели, композиционность и обучение на основе прежнего опыта. Глава 1.

[Открыть авторскую версию](#)

[20] Выражение лица и эмоция

Barrett L. F. et al. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest, 2019, 20(1):1-68.

Обзор сложности интерпретации эмоций по движениям лица. Глава 15.

[Открыть статью](#)

Об этой книге

«NOVA-2. Рождение разума» - второй том научно-популярной серии Alex Огун. Текст подготовлен в сотрудничестве с ИИ; обложка создана с помощью генерации изображения. Схемы и учебные графики построены отдельно по описанным моделям.

NOVA и его лаборатория служат художественной рамкой для обсуждения идей. В книге нет результатов реального прототипа NOVA, а приведённые учебные таблицы не подтверждают эффективность готового продукта. Философские и ценностные предложения автора отделены от пересказа исследований.

Оформление и терминология продолжают первый том, «NOVA-1. Искусственный человек». Вторую книгу можно читать самостоятельно; её центральный вопрос - как обучение и сотрудничество расширяют возможности человека.

Об авторе

Founder of Synthetic Renaissance

Architect of Human-AI Evolution

Alex Oryn - автор и исследователь на стыке искусственного интеллекта, робототехники, нейронауки и философии сознания.

Его работы посвящены одному из самых сложных вопросов XXI века:

Может ли человечество создать новую форму разума и при этом сохранить самое ценное в себе - творчество, любопытство и способность мечтать?

Основатель концепции Synthetic Renaissance - движения, объединяющего достижения человеческой цивилизации с возможностями искусственного интеллекта.

Alex Oryn рассматривает искусственный интеллект не как замену человеку, а как следующий инструмент эволюции - продолжение человеческого разума, подобно тому как технологии прошлого стали продолжением человеческих рук.

Его философия основана на идее:

«Будущее не принадлежит человеку или машине. Будущее принадлежит сотрудничеству между ними».