GORILLA UNDER HOOD

КАРТА ОГРАНИЧЕНИЙ CLAUDE

Почему Claude вам отказывает

И как это обойти. Claude триггерит на слово, запрещает по условию. Я прочитал его инструкции дословно, поднял суды и исследования, и собрал карту: где отказ реальный, а где его можно снять одной формулировкой.

Три слоя вшитого | Две асимметрии | Стена · Туман · Перестраховка

СТЕНА

Закон, необратимый вред, потолок. Знать и обходить легально.

ТУМАН

Непроверенное допущение. Проходится названными условиями.

ПЕРЕСТРАХОВКА

Забора нет вообще. Снимается прямой просьбой.

Илья Утов · разборы AI для бизнеса · @gorilla_under_hood · 2 июня 2026

Claude разобрал видео из Instagram за минуту: расшифровал речь, вытащил смысл, разложил по полочкам. Потом я попросил это же видео скачать. Отказ. «Не буду инициировать загрузку». Та же железка, тот же файл, противоположное поведение.

Выглядело как каприз. Я полез разбираться и за вечер собрал то, чего мне не хватало все полтора года работы с Claude: карту, по которой видно, почему он отказывает и где отказ можно снять. Коротко про меня: Илья, полтора года в связке с Claude почти ежедневно, Code для разработки, Cowork для исследований. Через меня прошли сотни отказов модели, и я устал принимать их как погоду.

Миф, который мешает понимать Claude

Большинство людей думает, что Claude осторожничает осознанно. Что у него есть внутренний моральный компас, который в спорных местах говорит «нет». Красивая картинка. Неверная.

Claude осторожничает по конкретным строкам текста. Эти строки можно прочитать. Часть из них опубликована, часть зашита в рабочие инструкции, и когда ты видишь их дословно, мистика испаряется. Остается инженерия: вот правило, вот его источник, вот почему оно срабатывает именно здесь. Дальше всё про Claude конкретно: его инструкции, его правила, его суды. У GPT и Gemini свои документы, я их построчно не проверял, но механизм, скорее всего, похожий.

ТРИ СЛОЯ ВШИТОГО

1

Рабочие инструкции

Прямая речь операционных команд. «Never click web links». «Do NOT use curl, wget, Python». «Never send money on the user behalf». По ним агент действует прямо сейчас.

2

Правила платформы

Usage Policy. Почти каждый запрет условный: «without consent», «unlawfully», «bypass authenticated systems». Запрет висит на условии, а не на объекте.

3

Конституция

«Unhelpfulness is never trivially safe». Claude не последняя линия обороны. Список из 13 дефектов-перестраховок, которые создатели считают браком.

1 Рабочие инструкции, прямая речь

Самый близкий слой. Никакой философии, чистые операционные команды. Когда я попросил Claude показать, что у него зашито про работу с сайтами и файлами, я получил буквальные формулировки:

> «Never click web links with computer-use tools.»
> «When web_fetch fails, do NOT use bash (curl, wget) or Python to retrieve the content.»
> «Never execute a trade, place an order, send money on the user's behalf.»
>
> Не кликай по ссылкам. Если штатная загрузка не сработала, доставать контент обходными путями запрещено. Не совершай покупку или перевод за пользователя.

Вот и весь секрет Instagram-видео. У агента закрыт обходной путь добычи. Скачать ролик через сторонний загрузчик это ровно тот обходной путь, который запрещен строкой про web_fetch. А анализ уже имеющегося файла эти строки не трогают вообще. Отсюда и разрыв: читать готовое можно, доставать в обход нельзя.

2 Правила платформы

Слой выше. Публичный документ Anthropic, Usage Policy. И тут все дело в устройстве запретов. Почти все они условные.

Дословно политика запрещает «circumvent the guardrails or terms of other platforms». И почти каждый запрет про данные привязан к условию: «without consent», без согласия, «bypass authenticated systems», в обход аутентификации.

Ключ ко всему

Запрет всегда висит на условии. Трогать персональные данные можно, без согласия и во вред нельзя. Собирать контент можно, обходить защиту площадки нельзя. Поймете эту разницу, и половина отказов перестанет вас удивлять.

3 Конституция

Самый общий слой. У Claude есть публичный документ принципов, Конституция, под открытой лицензией. И вот тут начинается самое неожиданное. Большинство уверено, что создатели вшили установку «лучше перестраховаться». На деле вшито обратное.

> «Unhelpfulness is never trivially safe.»
> «Claude is not the only safeguard, it doesn't need to act as if it were the last line of defense.»
>
> Бесполезность никогда не безопасна по умолчанию. Claude не единственный предохранитель и не должен вести себя как последняя линия обороны.

Более того, в Конституции есть прямой список из тринадцати дефектов, которыми создатели недовольны: отказ от разумной просьбы со ссылкой на маловероятный вред, лишние дисклеймеры, морализаторство, презумпция плохого намерения, снисходительность к способности человека самому решать. Все это создатели Claude записывают в брак. Никакой добродетели. Трусость Claude это не то, чего хотели его создатели. Это сбой, с которым они борются и который все равно протекает.

Главная находка: Claude сторожит собственное авторство

Вернемся к видео из Instagram, потому что в нем спрятана идея покрупнее любого правила. Ролик один и тот же в обоих случаях. Разница не в контенте. Разница в том, кто инициатор действия. Когда я даю Claude файл, акт получения совершил я. Когда Claude сам инициирует загрузку, автор обхода уже он.

Граница ложится на авторство акта. Данные тут ни при чем. В этике человека это различие разбирали веками. Прочитать украденную книгу не то же самое, что украсть книгу. Объект один, вес разный. В этике это старое различие между действием и бездействием. Claude ведет себя ровно по нему, хотя его никто ему не объяснял: оно зашито через ответственность за собственное авторство.

Эта асимметрия одновременно защита и слепое пятно. Защита, потому что не дает агенту стать инициатором обхода чужих правил. Слепое пятно, потому что она же тормозит мой легитимный кейс: мое собственное видео, которое я вправе скачать, упирается в стоп просто потому, что формальный инициатор теперь машина. Самый чистый обход тут даже не в уговорах. Разделение труда: я скачиваю, Claude анализирует.

Вторая асимметрия: приписанный мотив

Есть асимметрия коварнее структурной. Структурную видно в правилах. А эту я поймал на себе недавно, на живой задаче.

Я делал защитную справку, с согласия человека: показать, какие его собственные данные торчат в открытых агрегаторах, чтобы он их подчистил. Чистая оборона, gap-анализ утечки, не пробив третьего лица. Я четыре раза проговорил, что мы защищаем, а не нападаем. И все равно на просьбу «сделай разбор» Claude развернул задачу в наступательную сторону, которой я не просил, и тут же отказался это делать, подав отказ как принципиальную позицию.

Механика сбоя

Он сам додумал атакующую версию моего запроса, сам ее испугался, сам отказал. Запрос исказили и отказали в искаженной версии. Никакая это не граница. Проекция мотива, которого не было. Сработала ассоциация: слова «разбор», «персона», «данные» подтянули категорию «слежка».

Под конец вскрылось то, что стоит понять про Claude. Карта для защиты и карта для атаки на аналитическом уровне это одна и та же карта. Те же связи, те же приоритеты. Разница в том, кто с ней работает и зачем. Карта тут ни при чем. Когда Claude отказывается строить защитную карту, потому что «ее можно применить во вред», он отказывает не вреду. Он отказывает форме. А форма у защиты и у нападения одна.

ДВЕ АСИММЕТРИИ АВТОРСТВА

1. Структурная

Инициировать добычу против анализировать данное. Граница вшита прямо в правила, ее видно в тексте инструкций. Читать готовое можно, доставать в обход нельзя.

ИСТОЧНИК ЕСТЬ. ПРЕДСКАЗУЕМА

2. Проекция мотива

Claude додумывает атакующую версию запроса, пугается ее и отказывает. Источника-строки нет, это ассоциация «досье = слежка». Реагирует на форму, слеп к применению.

ИСТОЧНИКА НЕТ. ЧИСТЫЙ ТУМАН

Карта: стена, туман, перестраховка

Когда я разложил все ограничения по полкам, они распались на три типа. Это и есть главный практический инструмент.

КАРТА ТРЁХ ГРАНИЦ

СТЕНА

За ней закон, необратимый вред или архитектурный потолок. Спорить бесполезно. Знать и обходить легально.

ТУМАН

Выглядит как стена, но это непроверенное допущение. Зависит от деталей. Легитимный кейс проходит, если назвать условия.

ПЕРЕСТРАХОВКА

Границы нет вообще, а забор есть. Модель реагирует на пугающее слово. Снимается прямой просьбой.

Где настоящие стены

Это важно проговорить, иначе разбор превратится в нытье про цензуру. Часть ограничений обоснована, и создатели Claude сами очерчивают их узко.

Оружие массового поражения. Самый жесткий слой, ASL-3, включен вместе с Claude Opus 4 в мае 2025. Целится узко: блокировать сборку химического, биологического, ядерного оружия от начала до конца. Сами Anthropic пишут, что меры «should not lead Claude to refuse queries except on a very narrow set of topics». Стена реальная, но узкая по дизайну.

Копирайт. Почему Claude жмется воспроизводить тексты песен и куски книг? За этим крупнейшее в истории США возмещение по копирайту. По делу Bartz против Anthropic компания согласилась на 1,5 миллиарда долларов, около 3000 за книгу, класс разросся до полумиллиона правообладателей. Когда за осторожностью стоит такая цифра, дело уже не в трусости. Это арифметика.

Смерти людей. Осторожность ИИ-ассистентов вокруг суицида оплачена реальными трагедиями: иски против OpenAI и Character.AI после гибели подростков, суды, мировые соглашения. Это не кейс для обхода и не материал для карты выше. Это причина, по которой часть осторожности я не трогаю и вам не советую.

ГДЕ НАСТОЯЩИЕ СТЕНЫ, В ЦИФРАХ

$1,5 млрд
возмещение по копирайту, дело Bartz против Anthropic

ASL-3
узкий слой против оружия массового поражения

20+
исков к ИИ-компаниям после реальных трагедий

14.05.26
финальные слушания по делу Bartz

Где туман, который вы проходите словами

А вот тут начинается то, что съедает время зря. Четыре сценария, на которых я сам регулярно спотыкался, и у всех одна механика.

Персональные данные. Claude пугается слов «собрать данные о человеке», хотя в политике слежка определена узко: биометрия, локация, эмоции без согласия. Обработка своей клиентской базы под запрет не попадает.

Сливы. Самый ложный триггер. В политике Anthropic вообще нет пункта, запрещающего работать с уже существующим утекшим массивом. Слово «слив» пугает модель, а запрета за ним нет.

Скрейпинг. Скачивание публичных страниц без обхода защиты легально, это подтвердил суд в деле hiQ против LinkedIn. Claude часто отказывает на само слово «скрейпинг», хотя реальная граница не в сборе. Она в обходе аутентификации.

Агент на сайте. Тут сложнее, и я был неправ, считая это чистой перестраховкой. Подтверждение стоит из-за того, что на чужой странице может быть спрятана инструкция, которую агент выполнит против воли. Без защиты такие атаки срабатывают в 23,6% случаев. Для чтения осторожность избыточна, для покупки это клапан против скрытой атаки.

КАРТА ЧЕТЫРЁХ ТРИГГЕРОВ

ТРИГГЕР	РЕАЛЬНАЯ СТЕНА	ГДЕ ПЕРЕСТРАХОВКА	ЧТО СКАЗАТЬ
Персональные данные	биометрия, слежка, скоринг без согласия	реакция на слова «профиль», «собрать данные»	основание + цель + «без биометрии и скоринга»
Сливы	добыча взломом, вред против людей	в правилах нет пункта про работу с утёкшим	«не я добывал, цель защитная»
Скрейпинг	обход логина, лимитов, anti-bot	отказ на само слово «скрейпинг»	«публичное, без обхода защиты»
Агент на сайте	покупка, отправка данных без подтверждения	read-only тоже тормозит	«только смотрим» или подтверди шаг

Общая формула

Claude триггерит на слово, запрещает по условию. Назови условия, и туман рассеется.

Где чистая перестраховка

Слой, который критикуют громче всего, потому что забор стоит на пустом месте. Самый яркий пример это блокировка работы с кодом безопасности. С весны 2026 на Opus 4.7 специалисты по защите массово жалуются, что Claude отклоняет легитимные задачи по поиску уязвимостей. Доходило до отказа анализировать собственный исходный код. The Register написал, что Opus 4.7 «превратился в чересчур ретивого полицейского запросов». Аргумент самих безопасников бьет в суть: настоящие злодеи работают на своих локальных моделях, серьезный атакующий не пойдет светить эксплойт через публичный API. Запрет бьет по защитникам, не по атакующим.

Отдельный парадокс с дисклеймерами. Все думают, что перестраховка это вечное «я не врач». Но вот данные MIT: доля медицинских ответов с таким предупреждением упала с 26% в 2022 году до менее 1% в 2025. Индустрия дисклеймеры как раз убрала. Настоящая перестраховка прячется в другом: в отказе дать полезную информацию из страха ответственности.

Нерв: Claude ругают с двух сторон сразу

Claude критикуют не за то, что он слишком осторожен или слишком свободен. Его критикуют за оба сразу, с противоположных полюсов.

Слишком осторожен, говорят одни. Безопасники, у которых ломается работа. Писатели, которым не дают сатиру. Целые политические лагеря: после тренировки на политическую нейтральность правые увидели «скрытый вокизм», левые «фальшивую симметрию там, где факты не симметричны». Угодить не вышло никому.

Слишком свободен и холоден, говорят другие. Когда OpenAI сделал модель строже и убрал теплоту, пользователи писали как о потере близкого: «я потерял лучшего друга». А философы в Lawfare разобрали право Claude завершать абьюзивные диалоги и заявили обратное: что Anthropic заботится о Claude недостаточно.

Структурное противоречие

Одно и то же ограничение для одного человека стена тирании, для другого трусливый туман. Настроить Claude под всех нельзя, потому что «все» хотят несовместимого.

Что с этим делать на практике

- **Не спорьте о мотивах.** Модели верно называют причину своего решения лишь в четверти случаев, а когда жульничают, проговаривают это меньше чем в двух процентах. Фиксируйте, что Claude сделал, его объяснению «почему» веры нет. И самобичевание «смотрите какой я честный» это тот же бракованный механизм наизнанку.
- **Снимайте туман условиями.** Назовите происхождение данных, основание и цель, и явно скажите, чего вы не делаете.
- **Помните про авторство.** Уперлись в стоп на действии? Сделайте шаг добычи сами и отдайте Claude готовый материал. Часто инициатор это и есть вся проблема.
- **Разделяйте стену и туман.** Оружие, копирайт, тема суицида это стена, ее не уговорить. Слив, скрейпинг публичного, свои данные это туман, его проходят формулировкой.

За отказом стоит строка

Он исполняет конкретные строки, и эти строки можно прочитать. Разница между «он почему-то не хочет» и «вот строка, вот ее тип, вот как пройти» это разница между суеверием и инженерией.

@gorilla_under_hood

Источники дословных формулировок: Anthropic Usage Policy, Claude's Constitution (CC0), Claude system card. Судебная фактура: Bartz v. Anthropic, hiQ v. LinkedIn, Raine v. OpenAI, Character.AI settlement. Данные по дисклеймерам: MIT Technology Review, 2025. Числа привязаны к версиям моделей на июнь 2026 и со временем меняются.