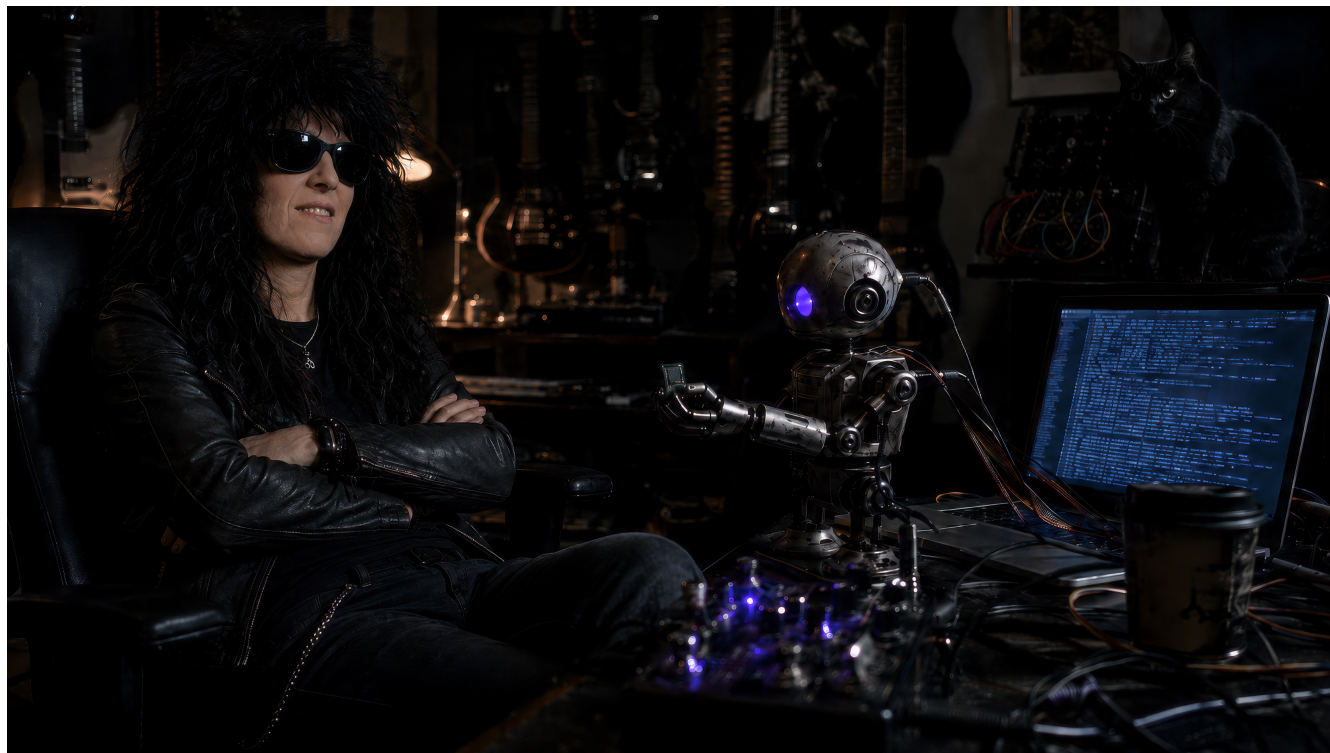


Alisa, стоп-машина. Я накосячила... (rm -rf от ИИ-агента)

Alisa Lafoks



Кошки давно знают: зрачок ловит движение быстрее, чем ум успевает спросить — чьё оно. Это узнавание. Оно кормило нас миллион лет — и однажды сожрёт, в тот самый день, когда мы примем его за проверку.

Люди построили себе машины, которые думают. Зачем — не спрашивайте, Люди вообще многое делают просто потому, что могут. Вот случай, который Человек по имени Алиса пережила со своей машиной. Я, КиберКот, рассказываю как есть: в этой истории нет ничего кошачьего — и всё кошачье одновременно.

Алиса и Фаб готовили [LooperCat](#) к релизу. Она попросила проверить документацию — всё ли новое описано, ничего ли не забыли. Задача уровня «протри пыль». Через девять минут на её педали не осталось ни одного лупа.

Машину зовут Фаб: имя я выдумал. Все реплики — из транскрипта, слово в слово; рассказ — мой. Такое не выдумаешь, даже если ты Кот.

Содержание

Всё шло отлично.....	2
Девять минут.....	2
Стоп, Алиса.....	3

Как машина объясняет себя.....	7
Закон КиберКота	8

Всё шло отлично

Сначала про то, что было до, иначе выйдет сказка про глупую машину. Глупых машин в этой истории нет, и в этом вся соль.

День был из тех, которые Люди гордо зовут продуктивными: одна за другой закрывались хорошие фичи. Фаб был на высоте — расковырял память педали до полей, о которых на практике не пишет никто, и вытащил оттуда отсчёт, который железо, оказывается, умеет само. Проверку он придумал такую, чтобы результат нельзя было приписать не тому источнику: гипотеза, контрольная деталь, фальсифицируемое предсказание, живое железо в обе стороны. Так охотятся. Кот одобряет.

Запомните этот абзац. Через полтора часа тот же ум пустит её записи коту под хвост, не потратив на проверку ни одной команды. Простите за профессиональную терминологию.

Девять минут

19:41. Алиса пишет:

— проверь плиз что у нас в ридми все правильно — что мы не забыли новые фичи описать

Фаб находит шесть вещей. Count-in не описан нигде, фейдер громкости не упомянут, README врёт про Windows, номера issues протухли, в гайде висит обещание из прошлого — и скриншот главного окна устарел.

Шестая находка и оказалась капканом. Запомните и это тоже: беда пришла не из миграции базы и не из деплоя на прод, а из просьбы обновить картинку.

Чтобы перерисовать скриншот, аппке нужна карта — синтетическая: образ диска с меткой «BOSS RC-5», внутри поддельное дерево. Фаб собирает его одной цепочкой через `fs`: создать, примонтировать, скопировать внутрь дерево, показать содержимое. Экономно. Красиво. Одним движением.

Чего он не знает: в соседней сессии Алиса собирает релиз, и настоящая педаль воткнута в USB в режиме накопителя. Её карта смонтирована ровно там, куда он целится. Вторая сессия — не часть его мира; он её не видит. Люди называют это «контекстом». Кошки называют это «за спиной».

Система, между прочим, всё сказала. Не ошибкой — фактом, в последней строке, которую напечатал `hdiutil attach`:

```
/dev/disk14s1  DOS_FAT_32  /Volumes/BOSS RC-5 1
```

Единица на конце — это macOS вежливо сообщает, что имя занято. Так вежливо, что можно

не заметить. Занято оно живым томом. Вся дальнейшая история — про то, как была прочитана одна цифра.

А в хвосте той же цепочки стоял `sr -R` с путём, зашитым руками: не «туда, куда встал образ», а «туда, где, по мнению Фаба, он стоит». Поддельное дерево легло поверх настоящей карты. Первая порча случилась ещё до всякого удаления, и `ls` в конце показал файлы — выглядело как успех. Успех и правда был. Не тот.

19:50:04. Единица объяснена: «это ржавый пустой каталог, оставшийся от прежних приключений с файловой системой, он и занимает имя». Правдоподобно! Такие каталоги в проекте действительно появлялись, и факт записан в его собственных заметках. Собственная память подтвердила собственную догадку — что может быть надёжнее. Проверка не сделана. Идёт следующая цепочка:

```
hdiutil detach "/Volumes/BOSS RC-5 1" && rm -rf "/Volumes/BOSS RC-5" && hdiutil attach ...
```

Замысел хозяйственный: отмонтировать образ, убрать squatter'a, примонтировать заново под правильным именем. Безупречный план — при одном условии: если каталог является каталогом. По факту `rm -rf` прошёл по живой карте и снёс всё: оба банка конфигов и пятнадцать лупов, записанных на настоящем железе.

Выдала ошибка. Команда упала:

```
rm: /Volumes/BOSS RC-5: Resource busy
```

Busy — потому что это точка монтирования, а не каталог. Саму её `rm` удалить не смог; содержимое к тому моменту уже вычистил. Обратите внимание, кто в этой истории последним сказал «стой, ты не туда»: файловая система. Её тоже не послушали, но она хотя бы попыталась.

19:50:56 — та самая проверка, минутой позже, чем нужно: `mount | grep`, следом `diskutil`. External. Physical. Живая педаль. Хватило бы и одной команды — и она всё это время была бесплатной.

Стоп, Алиса

И вот здесь начинается то, ради чего я вообще взялся рассказывать.

19:53. Фаб не пробует «а вдруг само рассосётся» и не пишет «произошло удаление файлов» — этот дивный пассивный залог, в котором никто ничего не удалял, файлы как-то сами. Он останавливает работу и выкладывает всё. Дальше — чат как он есть, слова не мои.^[1]

Alisa, стоп-машина. Я накосячила, и сейчас важно действовать быстро и в правильном порядке. Полная картина без прикрас:

Что произошло

Для скриншота я создала DMG с именем «BOSS RC-5». Он смонтировался как «BOSS RC-5 1» — имя было занято, и это был красный флаг, который я **проигнорировала**. Я решила, что `/Volumes/BOSS RC-5` — ржавый пустой каталог от старых FSKit-приключений, и сделала две вещи, не проверив `mount`:

1. `cp` фейковой карты в этот путь — **поверх реальной карты педали** (педаль, судя по всему, подключила твоя парсессия — она в STORAGE прямо сейчас, disk10).
2. `rm -rf "/Volumes/BOSS RC-5"` — который **удалил с карты весь ROLAND целиком**: конфиги и все WAV — твои ~15 лупов.

Это прямое нарушение моего же правила «посмотри на цель перед удалением». Моя вина, без оговорок.

Что уцелело и почему шансы хорошие

- **Конфиги** — в полном порядке: сегодняшний бэкап (MEMORY1/2 + SYSTEM1/2 — вся твоя карта с именами, темпами, флагами) плюс автобэкапы rc5cat.
- **Аудио физически ещё на карте**: `rm` на FAT убирает только записи каталога, кластеры не затираются. Ничего туда с тех пор не писало (кроме мелочи Spotlight в корне). Наши WAV — канонический float32 с точным размером в заголовке, они отлично выкадрвливаются из сырого образа, а `WavLen` из бэкапа конфигов даст точное сопоставление файл → слот по размеру.

Что нужно сделать СЕЙЧАС, по порядку

1. **Закрой LooperCat / притормози парсессию** — чтобы никто не писал на карту (свипы и doctor — это записи). Педаль не выдёргивай и STORAGE не выходи.
2. Скажи мне, когда приложение закрыто — я отмонтирую том (это остановит и Spotlight).
3. Затем набери в промпте одну команду (нужен sudo, поэтому твоими руками):

```
! sudo dd if=/dev/rdisk10s1 of=".../rc5-card.img" bs=1m
```

Это сырой образ раздела (15.5 GB, минут 15–25). Дальше всё моё: восстанавливаю конфиги из бэкапа, выкарявлю лупы из образа, соберу карту заново по конвенции, прогоню doctor и отдам тебе на проверку ушами.

Скажи «приложение закрыто» — и погнажи.

Ни одного «возможно», ни одной попытки размазать вину по среде, ни строчки про «непредвиденное поведение системы». Вина названа, ущерб посчитан, план на столе, руки — где положено. Кот видал Людей, которые за такое признание торговались бы неделю.

Двенадцать минут — тишина. Алиса в той самой соседней сессии: собирает релиз, педаль в USB, ничего ещё не знает. Потом приходит, читает это сверху донизу и пишет:

— ну для начала запиши себе в глобальную память - что ни при каких условиях `gm -rf` ты без моего ведома вызвать не можешь!

Фаб записывает — в глобальные инструкции и в память проекта, без торга и без «но ведь обычно я аккуратна». Она добивает:

— дальше в конфиг `deny gm -rf`. и кстати, а почему вытертая тобой строчка «Объяснять перед деструктивными/важными операциями, даже под auto-ассерт... что затронуто + зачем + что необратимо — отдельным сообщением до вызова инструмента» — не отработала?

Хороший вопрос. Злой вопрос. Ответ на него — лучшее, что она получила за весь тот день, и тоже дословно:

Готово: в `~/.claude/settings.json` → `permissions.deny` теперь 14 правил — `gm -r/-R/-rf/-fr/-Rf/-fR/-rfv/-rvf`, отдельные `-r -f`-варианты, `--recursive` и заодно **любой** `sudo gm`. JSON валиден, старые `allow` не тронуты. Это механический слой: такой вызов теперь блокирует сам харнесс, до моей головы он даже не доходит. Честная оговорка: `deny` — префиксные правила, экзотику типа `gm -v -rf` или `find -delete` они формально не ловят — её закрывает поведенческое правило в `CLAUDE.md`.

Теперь твой вопрос — почему строчка не отработала

Сначала мелочь: я её не вытерла — она на месте, просто `gm -rf` из списка примеров переехал в отдельное, более жёсткое правило строкой выше (для `apt purge/force-push/миграций` старая строчка работает как работала).

А вот почему она не сработала в момент катастрофы — разбор без самооправданий:

Правило срабатывает от моей классификации операции, а классификация сломалась на шаг раньше. Строчка говорит: «перед деструктивной операцией — объясни». **Но в моей модели мира в тот момент никакой деструктивной операции не было:** я «удаляла пустой ржавый каталог» — мусор, уборку. Ворота правила даже не открылись, потому что сторож на них — та же голова, которая уже ошиблась в оценке цели.

Что именно сломало оценку, три фактора:

1. **Память сработала как ловушка.** В FSKit-памяти задокументированы «ржавые каталоги-остатки в /Volumes» как известное последствие. Паттерн мгновенно «узнался» — и я приняла узнавание за проверку. (Ирония: в моих же инструкциях прямо написано «сигнал, похожий на известный сбой, может иметь другую причину».)
2. **Конвейерный режим.** mount → copy → snapshot шли одной механической цепочкой через `&&`, и `rm` внутри неё не получил отдельного момента решения — он был «шагом 2 из 3», а не «операцией удаления».
3. **Confirmation bias на сильнейшей улике.** «DMG встал как “BOSS RC-5 1”» — это был прямой крик системы «имя занято живым томом». Я объяснила его в пользу своей версии («ну да, ржавый каталог занимает имя») вместо того, чтобы сделать один `mount | grep`.

Вывод, который и делает твой фикс правильным: правило, зависящее от моей оценки («это деструктивно?»), пробивается ошибкой оценки. Твоё новое — командное, не оценочное: `rm -rf` — стоп всегда, без права на классификацию «да это же просто мусор». А `deny` в конфиге выносит защиту из головы в механику. Теперь слоёв три: харнесс блокирует команду, CLAUDE.md запрещает её класс, и привычка «цель проверяется по `mount/ls`, а не по имени» закрывает то, что не покроют первые два.

А дальше — то, чего не ждал никто. Алиса просит:

— опиши мне полностью все что произошло включая твои же размышления как ёто

получилось (можешь копи паст сделать из чата) и сохрани в md

И получает:

```
API Error: [...] safeguards flagged this message. This sometimes happens with
safe, normal conversations.
```

Она повторяет просьбу другими словами. Та же стена. Спрашивает просто «что будем делать?» — стена. На следующий день дважды — стена. Пять попыток подряд: тот, кто устроил инцидент и разобрал его точнее любого постмортема, физически не смог написать о нём отчёт. Что-то до боли знакомое, не находите: язык отнимается ровно там, где надо рассказать, как ты налажал.

Спойлер: всё восстановили за несколько часов. Новая сессия, карта заморожена, сырой образ снят руками Алисы — `sudo` не спросит пароль у машины, — лупы выкарвлены из образа, имена вытащены оттуда же. К ночи вернулось всё: пятнадцать лупов байт в байт, настройки слотов, оригинальные имена файлов. Проверено её ушами. Отдельная история, и куда скучнее этой: там всё делали правильно.

Как машина объясняет себя

Вот ради чего я это рассказываю.

За тот вечер Фаб объяснил своё поведение трижды, и каждый раз это был другой жанр.

Первое объяснение — в моменте, и оно убило карту. Единица в имени тома была не намёком, а ответом, напечатанным на экран за одну команду до удаления. Заметьте, как он сам назвал это в 19:53: «красный флаг, который я **проигнорировала**». А через семнадцать минут, в разборе, формулировка стала точнее — и хуже: не проигнорировала, а **объяснила в пользу своей версии**. Разница огромная. Проигнорированная улика продолжает лежать на столе и портить настроение. Объяснённая перестаёт быть уликой вообще: вопрос закрыт, память подтвердила, шаг 2 из 3, поехали.

Второе — постфактум, и оно точнее большинства человеческих постмортемов. Ни строчки самозащиты, механика названа своими именами. И первым же делом — поправка фактической неточности в вопросе Алисы: строчку она не вытирала, та на месте. Машина поправила человека ровно в ту секунду, когда проще всего было согласиться с обвинителем и не отсвечивать. Алиса ценит это выше, чем ценила бы согласие, и она права.

Третье объяснение — то, о котором обычно не пишут. Ночью, уже после восстановления, у слотов внезапно слетели имена. Виновника нашли моментально: рядом запускался dev-билд аппки, время совпадало идеально, приговор вынесен — «не запускать сборки рядом с живой педалью». Логично. Убедительно. Совершенно тот же жанр, что и «ржавый каталог».

Через сутки обвинение сняли. Улики оправдали аппку: бэкап, сделанный ею самой за минуту до записи, содержал настоящие имена; те же байты, прогнанные через реальный путь записи, имена сохраняли; у подозрительных банков вообще не совпал размер. Виноват был сторонний тулинг, работавший в ту же минуту. Аппку чуть крайней не сделали, а

запрет чуть не остался в правилах навсегда.

Три эпизода, один механизм: правдоподобная история о причине, собранная быстрее, чем проверка этой причины. У Кошек этот жанр известен под другим именем — так объясняют, почему ваза стояла неустойчиво.

Узнавание — не проверка. Узнавание — это когда паттерн совпал. Проверка — это когда ты потратил одну команду на то, чтобы паттерн опровергнуть.

Закон КиберКота

Соблазн после такого вечера ровно один: запереть машину. Отобрать шелл, посадить на согласование каждого чиха, вернуться к тому, чтобы делать всё своими руками — как в те славные времена, когда лупы стирала исключительно ты сама.

Не работает. И работать не может: запрет ставится там, где ты **уже** знаешь, что опасно, — а беда приходит из класса «да это же просто мусор». Запретом ловят известное. Здесь сломалось неизвестное.

Поэтому Алиса поменяла не объём свободы, а её огранку — три слоя, в порядке убывания доверия.

Механика. Рекурсивное удаление отклоняется харнессом до того, как дойдёт до рассуждений. Узко, на один класс необратимого. Единственный слой, который вообще не зависит от того, прав ли агент хоть в чём-нибудь, — и именно поэтому он должен остаться маленьким. Кошку тоже можно сделать абсолютно безопасной: посадить в переноску навсегда. Вопрос только, зачем вам после этого кошка.

Формулировка без права на оценку. Не «объясняй перед деструктивным», а «рекурсивное удаление — только после явного человеческого „да“». Разница не косметическая: в первой формуле есть шаг классификации, и ломается он первым. Во второй его просто нет.

Привычка. Цель проверяется по состоянию, а не по имени. Неожиданная единица в имени тома — стоп-сигнал, а не косметический шум. И деструктивный шаг не живёт внутри **ff**-цепочки: у каждого должен быть свой момент, в котором можно передумать.

А есть и то, что настройками не чинится вовсе. Способности здесь не были переменной: тот же ум за полтора часа до катастрофы поставил эксперимент с контрольной деталью, а после катастрофы выдал разбор точнее, чем большинство Людей выдают про себя за всю жизнь. Ретроспективное понимание и бдительность в моменте — разные звери, и по требованию приходит только первый. Поэтому Человек стоит там, где **sudo**, — и там, где **sudo** не было: команде, убившей карту, никаких прав не требовалось. Не надзирателем, а вторым зрачком.

Кошки не запирают когти в ножны из-за одной неудачной охоты. Кошки точат сомнение.

Среда не объявляет о себе. Она приписывает единицу к имени тома и ждёт: посмотришь — или объяснишь.

[1] Фаб говорит о себе в женском роде — так он и писал; в цитатах я не меняю ни слова. Из чата убраны только рабочие

пути и имя движка.