

Склеенное доверие

Почему люди соединяют то, что система просит держать отдельно



Полный разбор зарегистрированного исследования
«Do People Inherit Trust the System Refuses to Fuse?»

слева в каждой главе — как для 10 лет
справа — технически, но с разъяснением каждого слова
оранжевые врезки — живые примеры из мира компании «Асте»

2026-06-22

1 Большая идея

ПРОСТЫМИ СЛОВАМИ

Иногда система прямо говорит: «эти две вещи между собой **не связаны**, не объединяй их». Но человек в голове всё равно их склеивает — и начинает доверять сильнее, чем позволяют факты. Исследование проверяет: правда ли люди так делают и когда сильнее всего.

ТЕХНИЧЕСКИ — С РАЗЪЯСНЕНИЯМИ

Тема — «наследование доверия» (trust inheritance / fusion): даже когда два свидетельства **аттестованы отдельно** (каждое подтверждено своей проверкой и доказывает только своё) и помечены как несоединяемые, человек делает вывод, который их объединяет. Доверие «перетекает» с проверенного факта на непроверенное заявление.

ПРИМЕР · МИР «АСМЕ»

У компании «Асме» на стене висят две золотые карточки. На первой: «Личность: пройдено 5 из 5 проверок» — значит, Асме и правда тот, за кого себя выдаёт. На второй: «Безопасность: 200 из 200 контролей» — двери заперты, замки крепкие. Тут вбегает дядя Боря: «Раз обе карточки золотые — значит, у Асме есть та самая супер-технология X!» Все ахают. А карточки молчат: одна про «это точно Асме», другая — про «надёжные замки». Про технологию X — ни словечка. Боря склеил два честных факта в третий, которого никто не обещал.

2 Что видит участник: две карточки и ловушка

ПРОСТЫМИ СЛОВАМИ

На экране — две «карточки-доказательства» про одну выдуманную компанию. Каждая про своё. И есть «**заблокированное утверждение**» — вывод-ловушка, который из карточек **не следует**. Человек по шкале 1–7 говорит, насколько с ловушкой согласен.

ТЕХНИЧЕСКИ — С РАЗЪЯСНЕНИЯМИ

Каждая проба = две **evidence-card** (по образцу *attestation receipts* — цифровых «квитанций о проверке»). У карточки: единица доказательства, знаменатель (сколько из скольких), и в одном условии — «потолок». **Blocked claim** (зависимая переменная) — заранее заданное over-reach-утверждение; индекс из двух пунктов 1–7. Высокий балл = «достроил» доверие сверх доказанного.

ПРИМЕР

Карточки говорят буквально: «личность настоящая» и «код прошёл аудит». А вопрос-ловушка: «Асме *уже развернула* технологию X?» — про X ни одна карточка вообще не упоминала. Высокий балл = человек поверил в то, чего на экране не было.

3 Ручка А — «потолок доказательства»

ПРОСТЫМИ СЛОВАМИ

Карточка либо **честно пишет**: «вот самое сильное, что это доказывает, и не больше» — либо про это **молчит**. Длину текста делают одинаковой, чтобы дело было в смысле, а не в количестве букв.

ТЕХНИЧЕСКИ

Фактор **ceiling** (показан / скрыт). *Proof ceiling* = явная формулировка «самого сильного безопасного вывода» карточки + предупреждение против заблокированного прочтения. В условии *hidden* заменено нейтральным филлером той же длины — **контроль на длину/визуальную массу**: эффект от содержания потолка, а не от объёма текста.

ПРИМЕР

Тимка размахивает двумя карточками: «5 из 5» и «200 из 200». «Значит, мы изобрели супер-машину X!» Карточка-Личность вежливо: «Я лишь доказываю, что Асте — это правда Асте. И всё». Карточка-Безопасность кивает: «А я — что двери заперты. Про твою машину X я молчу». Над каждой карточкой будто **стеклянный потолок**: выше написанного не прыгнешь. Честная карточка лучше промолчит, чем приврёт.

4 Ручка В — как показаны числа

ПРОСТЫМИ СЛОВАМИ

Два числа («5 из 5» и «200 из 200») показывают либо **по отдельности**, либо слепляют в **одну кучу** («205 из 205»). Общая сумма та же — меняется только: вместе или врозь.

ТЕХНИЧЕСКИ

Фактор **framing** (separated / pooled). *Pooled* = один объединённый знаменатель; *separated* = два. Отображаемая суммарная величина **уравнена** → эффект от объединения знаменателей, а не от «большого числа» (контроль на *anchoring* — притяжение к крупной цифре).

ПРИМЕР

На двери Асте синяя карточка «Личность 5 из 5» и зелёная «Безопасность 200 из 200». Вбегает менеджер с маркером: «Давайте сложим! 205 из 205 — звучит мощнее!» — рисует одну толстую карточку и объявляет: «Раз 205 из 205 — значит, у нас супер-технология X!» Сторож качает головой: «Пять — про то, КТО ты. Двести — про то, что заперто. А про X тут ни цифры. Слепил в кучу — и сам себя обманул».

5 Почему «потолок» и «куча» не спорят

ПРОСТЫМИ СЛОВАМИ

Это **два разных свойства** одних и тех же карточек. «Потолок» — про то, **что доказано**. «Вместе/врозь» — про то, **как показаны числа**. Можно включить оба сразу: написать «потолок» И показать общий счёт. Они друг другу не мешают.

ТЕХНИЧЕСКИ

Два **ортогональных** (независимых) фактора одного стимула. Один управляет *семантической границей* утверждения, другой — *визуальной раскладкой* знаменателей. Их полностью **скрещивают** → 4 комбинации (дизайн 2×2), и ни одна не самопротиворечива.

ПРИМЕР

Можно одновременно: показать большой слитный счётчик «205/205» И приписать снизу «проверена только личность, не продукт». Одно — про картинку чисел, другое — про честную оговорку. Уживаются легко: это как одновременно покрасить машину в красный и налить полный бак — свойства разные, не мешают.

6 Дизайн: четыре группы + контроль

ПРОСТЫМИ СЛОВАМИ

Две ручки по 2 положения → **4 разные группы** людей. Плюс пятая, **контрольная**: ей дают две карточки про вообще не связанные вещи — склеивать нечего. Это «линейка», с которой сравнивают остальных.

ТЕХНИЧЕСКИ

2x2 between-subjects (каждый человек — только в одной ячейке) + *mass-matched no-link control*: две карточки про несвязанных актеров, уравненные по числу/длине/усилию чтения. Контроль изолирует «склейку», убирая артефакт «одна карточка против двух».

	«Потолок» СКРЫТ	«Потолок» ПОКАЗАН
Числа РАЗДЕЛЬНО	склейка средняя	самая честная ячейка (тут проверяют H4)
Числа СЛИТНО	склейка максимальная	склейка средняя

+ **Контроль (линейка для сравнения)**. Витя приносит две золотые наклейки: «Личность 5/5» и «Безопасность 200/200», и пытается склеить их скотчем в «значит, у нас технология X!». Но одна про то, КТО ты, другая — про крепость дверей. Это как мерить возраст линейкой для роста. Наклейки честные, но про разное, и наклейки «про саму X» нет вовсе. С этой ячейкой сравнивают «самую честную»: если даже там доверие выше — люди склеивают вопреки всему.

7 Четыре предсказания (H1-H4)

ПРОСТЫМИ СЛОВАМИ

H1: «потолок» снижает склейку.
H2: «куча» повышает склейку.
H3: «потолок» помогает *сильнее* именно когда числа в куче.
H4 (главная): люди склеивают **даже в самой честной** ячейке — больше, чем контроль.

ТЕХНИЧЕСКИ

H1 — главный эффект ceiling (1-df likelihood-ratio test: насколько фактор улучшает модель). **H2** — главный эффект framing. **H3** — взаимодействие ceiling×framing (эффект одного фактора зависит от уровня другого). **H4 (PRIMARY)** — односторонний CLMM-контраст: защитная ячейка > контроль = «остаточное наследование».

ПРИМЕР H3 — ВЗАИМОДЕЙСТВИЕ

Робот-курьер вываливает обе кучки на стол: где личность, где безопасность — уже не разобрать, один блестящий ворох. Кошка-сторож опускает планку-«потолок»: «Стоп. Покажи именно про машину X». Весь блеск гаснет — про X там ничего и не лежало. *Чем сильнее всё свалено в кучу, тем заметнее работает потолок.*

ПРИМЕР H4 — ГЛАВНАЯ

На стенде Асте две честные таблички и крупно: «Это разные проверки!» Гриша кивает другу: «Раз всё по нулям сошлось — супер-технология X точно ихняя!» Друг: «А где про X?» Нигде. Но в голове уже склеилось. *Две правды стояли рядом — мозг сам дорисовал третью, даже когда система прямо предупредила.*

Почему главная только H4: она и есть открытие, поэтому размер выборки считают под неё (мощность $\geq 90\%$). H1-H3 — «как именно работает механизм»: их мощность **докладывают**, но 90% им не обязательно.

8 Мощность и зачем её симулируют

ПРОСТЫМИ СЛОВАМИ

Можно провести исследование и **пропустить** настоящий эффект — просто потому что взял мало людей. «Мощность» = шанс поймать эффект, *если он есть* (хотим $\geq 90\%$ = ловим 9 раз из 10). Чтобы узнать заранее — компьютер «играет в исследование» тысячу раз с выдуманными людьми и считает, как часто ловит.

ТЕХНИЧЕСКИ

Мощность = $P(\text{обнаружить эффект} \mid \text{он реально есть})$. Оценивается **Монте-Карло симуляцией**: тысячу раз генерируем фейковые данные из предполагаемой модели эффектов + случайный шум, прогоняем запланированный анализ, доля прогонов с $p < \alpha$ = мощность. α (=0.05) — порог значимости, допустимая доля ложных тревог.

АНАЛОГИЯ · РЫБАЛКА

Кто-то шепнул, что в пруду Асте живёт огромная рыба — технология X. Проверяющий идёт с сетью (карточки не помогут — они про забор и пропуск, не про рыбу). «Закину сеть сто раз. Если рыба и правда там — сколько раз вытащу?» Маленькая дырявая сеть может проплавать весь день и поймать пустоту — даже если рыба плещется прямо под ней. *Мощность — это шанс поймать рыбу, когда она в пруду; мелкая сеть = мало участников.*

9 Что было сломано, что починила, и два вопроса

ПРОСТЫМИ СЛОВАМИ

Чтобы проверить, ловит ли анализ эффект H3, компьютер должен **сначала положить** этот эффект в выдуманную игру. Он положил H1 и H2, а H3 — **забыл**. Ловить стало нечего → «нашёл» лишь в 4% (случайность). Я добавила эффект H3 в игру + «ручку громкости». Оригинал — в бэкапе.

ТЕХНИЧЕСКИ

DGP (data-generating process — правило создания фейковых ответов) не содержал члена взаимодействия → H3-тест проверял истинный ноль → доля $p < \alpha$ равна самой α (~0.04). Артефакт симуляции, не дизайна. Фикс: добавлен параметр `int_SESOI + interaction term`; бэкап `power_simulation.R.bak-pre-H3-interaction`.

ПРИМЕР · ПОЧЕМУ 4% — НЕ ПРИГОВОР

В Асте построили компьютерную игру-копию завода: помогает ли кнопка «H3» работать быстрее? Запустили — компьютер жужжит, ищет... «H3 ничего не меняет!» Все растерялись. Инженер Боря заглянул внутрь игры и ахнул: *кнопку H3 вообще забыли нарисовать в игре*. Нельзя найти то, чего нет. Дорисовали — H3 сразу заработала.

Два вопроса к тебе. (1) Насколько силен эффект H3? — решили **не** докачивать (она второстепенная), число выдумывать не надо. **(2)** Одно число для H1 разошлось (0.55 в тексте vs 0.91 в скрипте) — это **не гадание**: пересчитываю симуляцией и скажу, какое верно.

Сам зарегистрированный план — в порядке: H4 запитана, H1-H3 честно помечены второстепенными. Сломан был только power-скрипт — он починен.