Klarna, Duolingo, Block: три кейса перестройки организации под агентов — что осталось после отката

2026-05-27

Klarna, Duolingo, Block: три кейса перестройки организации под агентов — что осталось после отката

В феврале 2024 года Klarna опубликовала результаты работы AI-ассистента на собственном сайте и в интервью OpenAI: ассистент за первый месяц обработал 2,3 млн диалогов — две трети клиентского сервиса — и выполнил, по оценке компании, объём работы, эквивалентный нагрузке 700 штатных агентов поддержки (OpenAI, Siemiatkowski в X). Полтора года спустя, в мае 2025-го, CEO Klarna Себастьян Семятковски в интервью Bloomberg сформулировал это прямо: «From a brand perspective, I think it's critical that you are clear to your customer that there will always be a human if you want.» Он же признал, что компания «зашла слишком далеко» в замене людей агентами и возвращается к найму сотрудников клиентского сервиса (Bloomberg).

Между этими двумя точками произошло ещё два публичных события того же типа. В апреле 2025 года CEO Duolingo Луис фон Ан разослал внутри компании меморандум о переходе на «AI-first» режим работы — он быстро ушёл в публичное поле через сотрудников и был подтверждён компанией (The Verge). К августу 2025-го фон Ан был вынужден объяснять смысл этого решения уже в The New York Times и признавать масштаб обратной реакции аудитории (The New York Times). В феврале 2026-го Block — холдинг Джека Дорси, владеющий Square и Cash App, — анонсировал сокращение около 40% штата, примерно 4 000 человек, прямо связав это с интеграцией ИИ в рабочие процессы компании (The Guardian).

Три истории — три разные стадии одного и того же цикла. У Klarna публичный откат уже произошёл. У Duolingo идёт переформулировка позиции в реальном времени. У Block самый громкий старт — и кейс ещё не дозрел до коррекции. Эта статья — карта того, как именно не строится организация поверх агентов и что остаётся за вычетом громких заявлений.

Дешевизна замены людей — в долг

Общая ошибка трёх компаний воспроизводится по одной схеме. Выбирается операционный слой, нагруженный людьми и описываемый понятной метрикой: время ответа в поддержке, объём текста для перевода, количество рутинных операций. Запускается агентный пилот, считается экономия часов и зарплат. Цифра получается крупная, потому что числитель — экономия — считается в

полном объёме задач, а знаменатель — то, что нужно достроить дополнительно, — остаётся пустым.

Дальше следует публичное объявление. У Klarna это пост Семятковски в феврале 2024-го и совместный кейс с OpenAI, в котором цифра «700 человек» вышла на обложку деловой прессы и обвалила котировки Teleperformance — крупного call-center подрядчика (Pragmatic Engineer). У Duolingo — корпоративный меморандум апреля 2025 года, который немедленно ушёл наружу: «AI-first означает, что мы перестанем использовать подрядчиков для работы, которую может сделать ИИ» (The Verge). У Block — пост Дорси сотрудникам в феврале 2026-го, где сокращение прямо связано с «intelligence tools»: автоматизацией внутренних процессов и снижением потребности в повторяемом труде (The Guardian).

Все три объявления формулируют ИИ-перестройку как операцию в плоскости задач — замену слоя людей слоем агентов. Ни одно не описывает изменений в плоскости организации: какие функции переезжают в новые единицы работы, какие границы ответственности перерисовываются, что становится с цепочкой эскалаций, какие новые роли создаются. Это и есть разница между «уменьшить численность на N процентов» и «переписать операционные единицы». Подробнее об этом сдвиге — в заметке про роль вместо места: должность как контейнер для работы устарела раньше, чем большинство компаний это заметили.

Что именно вернулось людям

В 2025–2026 годах у всех троих в той или иной форме случилось одно и то же: тихий возврат функций людям. Интересна не сама траектория отката, а его геометрия.

В Klarna после публичного признания Семятковски компания начала нанимать обратно сотрудников поддержки — но не на старые позиции. Возвращаемые роли сосредоточены вокруг работы с нестандартными случаями: спорные транзакции, эмоционально нагружённые ситуации, эскалации регуляторного характера (Bloomberg). По сути компания перерисовала уровень эскалации, признав, что плоская архитектура «агент-клиент» не покрывает дисперсию реальных запросов. Цифра «700 человек, эквивалент полной команды», на которую опиралась первая декларация, сама по себе скрывала важную деталь: с самого начала бот эскалировал «всё нестандартное» людям (Pragmatic Engineer). Когда команду людей сокращают, граница эскалации становится прозрачной — и оказывается, что она проходит по тонкому, но критичному слою клиентских историй.

У Duolingo откат прошёл без громких аннонсов. CEO в интервью The New York Times и Fortune переформулировал позицию: «AI-first» — это не «без людей», а «по умолчанию через ИИ, с человеком там, где он нужен», и обратная реакция аудитории на меморандум апреля 2025 года была для него неожиданностью (Fortune, The New York Times). На уровне операций часть подрядчиков верну-

лась в виде «контент-редакторов» и «языковых экспертов» — людей, которые верифицируют и финализируют машинный перевод, особенно для языков с малой обучающей выборкой и культурно нагружённых паттернов (Customer Experience Dive). Цена ошибки в продукте, который учит языку, оказалась выше, чем экономия от отказа от проверки.

У Block ситуация другая — это самый свежий и самый громкий случай. Сокращение 40% штата в феврале 2026-го не сопровождалось одновременным hire-back, и публичных свидетельств отката пока нет. Но в этой же истории видны два признака, которые в Klarna и Duolingo проявились до коррекции и в итоге её вызвали. Первый — внутренние сообщения о том, что «intelligence tools» (агенты, в частности проект Goose) применяются не к описанной перестроенной операции, а к существующей структуре труда: автоматизируется то, что и так выполнялось людьми, без переописания самой работы (The Guardian). Второй — слабый крипто-рынок и финансовое давление, на фоне которого AI-нарратив часто становится удобной публичной оболочкой для классической оптимизации затрат, не отвечающей на вопрос «какие операционные единицы перестроены». Аналитики уже называют это «AI-washing» — публичная привязка сокращений к ИИ как способ объяснить инвесторам цикл затрат.

Во всех трёх случаях устойчивая ось одна и та же: люди возвращаются (или возвращение анонсируется) не на старые роли, а на верхний слой суждения, эскалаций и аккаунт-менеджмента. За агентами остаётся то, что поддаётся жёсткой формализации.

Comparison: три кейса в одной таблице

Параметр	Klarna	Duolingo	Block
Год публичного объявления	Февраль 2024	Апрель 2025	Февраль 2026
Замещаемый слой	Клиентская поддержка	Подрядные переводы и контент	Сквозные операционные процессы и поддержка
Цифра, вынесенная в публичное поле	«Эквивалент 700 агентов поддержки»	«AI-first компания», 10% подрядчиков сокращено	40% штата, ~4 000 человек
Форма коммуникации	Совместный пресс-кейс с OpenAI + пост CEO	Внутренний меморандум + утечка	Внутреннее сообщение Дорси + публичная подача
Год коррекции	2025 (явная)	2025 (переформулировка)	Ещё не наступила (по состоянию на май 2026)
Что вернулось людям	Сложные случаи, эскалации, нестандартные транзакции	Языковые редакторы, культурная проверка	Пока ничего публичного
Что осталось за агентами	Первичная обработка типовых запросов	Перевод по языкам с большим корпусом	Внутренние ассистенты, рутинные процессы
Тип ошибки	Недооценка дисперсии запросов	Подмена операционного дизайна лозунгом	Замена в плоскости задач без переописания единиц

Граница не по сложности, а по формализуемости

Если собрать стабильную зону работы агентов по итогам коррекции Klarna и Duolingo и по тому, что про Block уже видно изнутри индустрии, получается узкий, но воспроизводимый набор. Первичная обработка типовых запросов с понятной таксономией. Перевод текста между языками с большой обучающей выборкой и проверкой постредактором. Кодогенерация под чёткие спецификации. Извлечение структурированных данных из документов. Рутинная аналитика по заранее определённым метрикам. Внутренние ассистенты-навигаторы по корпоративным знаниям.

У всех элементов этого списка три общих свойства. **Измеримый выход**: правильность результата проверяется без привлечения организационного контекста. **Повторяемый вход**: классы запросов укладываются в конечное число паттернов. **Низкая цена единичной ошибки**: один плохой перевод или пропущенный тикет не разрушает систему.

Где хотя бы одно из трёх свойств ломается — там агент возвращается под человека. Спорная транзакция в Klarna нарушает первое: измеримого «правильного» ответа нет, есть переговорный исход, в котором учтены история отношений, регуляторные ограничения и риск-аппетит. Локализация культурно нагружённого контента в Duolingo ломает второе: каждый случай уникален, и средний паттерн машинного перевода даёт результат, который в продукте про язык воспринимается как оскорбление. Регуляторная эскалация ломает третье: одна ошибка может стоить лицензии или превратиться в публичный скандал.

Это и есть «хвосты распределения», которые в трёх кейсах публично проявились как причина отката или как зона ближайшего риска. Агенты эффективны в средней части распределения, но плохо ловят края. Хвосты — зона, где требуется суждение, контекст, ответственность за решение. Никакая универсальная агентная оболочка над фронтирной моделью эту зону не закрывает: вопрос не в качестве модели, а в том, что входной сигнал в хвостах слишком разрежен и контекстно нагружен, чтобы строить на нём измеримую таксономию.

Связанный сюжет — почему универсальная агентная обвязка вообще не моат: это разбор в заметке про управляемую обвязку как зависимость. Здесь же ключевой вывод другой: если граница между агентом и человеком определяется не сложностью задачи, а её формализуемостью, то «процент автоматизации» — это плохая метрика трансформации. Она не отвечает на вопрос, какую долю **формализуемого** труда покрыли агенты, и какая доля **неформализуемого** труда осталась нагружённой на людей — и сколько новых ролей при этом не описано.

Что отделяет PR-замену от структурной перестройки?

Все три компании совершили операцию в плоскости задач — заменили людей, выполнявших отдельные функции, на агентов, выполняющих те же отдельные функции. Эта операция выглядит дёшево на бумаге: каждый человек — это зарплата, каждый агент — это API-вызов, разница — в счёт прибыли. Дорого она становится в момент, когда обнаруживается, что между «человеком, выполняющим работу» и «работой, выполненной правильно» лежит слой невидимого контекста: связи с другими функциями, история взаимодействия с клиентом, неявные знания о рисках, культурный паттерн. Этот слой не описан в инструкции и не воспроизводится промптом.

Структурная перестройка устроена иначе. Она начинается не с вопроса «кого можно заменить», а с вопроса «какие операционные единицы у нас вообще есть». В классической организации единица — должность с описанием, KPI и

местом в иерархии. В ИИ-нативной организации операционная единица определяется не носителем, а функцией, которую нужно выполнить: носитель — человек, агент или их комбинация — выбирается под класс случая. Эта инверсия описана в заметке про роль вместо места.

Из этого сдвига следуют три наблюдения о том, чего не было выстроено у Klarna, Duolingo и Block — и что отделяет работающую перестройку от PR-операции.

Первое наблюдение касается того, как запускается работа. Klarna до отката пыталась перевести поддержку в чисто событийную модель — каждый тикет триггерит автоматическое исполнение. Не была достроена обработка событий, которые не закрываются с первого прохода: в исходной архитектуре «всё нестандартное» передаётся людям, но численность этих людей при этом сокращена. В организации, где координация строится через совещания и проектные статусы, агентный слой нагружается асимметрично: автоматизирует ходовую часть и оставляет нестандартные хвосты в неопределённом состоянии. У Duolingo один из признаков проблемы — для редких языков разрыв между поступившей задачей и тем, как именно её решить, оказывается шире, чем система способна закрыть без участия эксперта.

Второе наблюдение — про то, где живёт контекст принятия решения. У Duolingo одна из причин отката — отсутствие машинной памяти о культурных особенностях языков. Постредакторы стали той самой памятью, которой не было в инструментах. У Klarna в спорных транзакциях контекст лежит в истории отношений с клиентом и в неявных правилах риска, не вытащенных в среду, к которой обращается агент. Контекст-в-голове-человека — это и есть незаметная зависимость, которая обнуляет экономику замены: пока он не вытащен в явную форму (память агента, правила маршрутизации, словари исключений), любое сокращение носителей этого контекста уменьшает качество решений на хвостах. У Block эта проблема ещё не проявилась публично, но именно потому, что 40-процентное сокращение свежее: ось «где живёт контекст» проверяется на горизонте 6–12 месяцев, как у Klarna.

Третье наблюдение — про принцип привязки ответственности. В трёх компаниях коммуникация шла в логике «заменим должность А агентом А». В работающей перестройке логика противоположная: сначала описывается зона ответственности, затем определяется, на каком носителе — человек, агент, комбинация — она исполняется в каждом классе случаев. Это не семантическая разница: она определяет, можно ли откатываться без потери репутации. Klarna откатывалась с потерей именно потому, что вернулась к тому же языку «нанимаем агентов поддержки», от которого ушла полтора года назад. Если бы возврат описывался как «нанимаем носителей функции эскалаций» — это уже не откат, а уточнение архитектуры. Duolingo нашла этот язык во второй итерации: «AI-first» переформулирован: «We are not going to replace people with AI», — сказал фон Ан The New York Times. Внутренний смысл сдвига — «AI-

default, с явно описанными зонами человеческого суждения». Block пока в первой итерации, и публичный язык — про сокращение, а не про перестройку.

Что измеряют не там

Связанная ошибка — в выборе метрики, на которую опирается решение об автоматизации. Чаще всего считают экономию часов или экономию фонда оплаты труда. Это знаменатель только частично: он не учитывает рост дисперсии решений, риск «незакрытых хвостов», стоимость новых ролей оператора и эскалатора, репутационную цену публичного отката.

Об этом — отдельная заметка про маржу, которая уходит выше уровня задач: «сэкономленные часы» — это плохая прокси-метрика, потому что она замеряет цену **на уровне задач**, а структурная экономика автоматизации формируется **на уровне операционных единиц**. Цена «забытого» хвоста в Klarna 2025-го (репутация, регуляторное внимание, IPO-нарратив) выше, чем сумма зарплат сокращённой команды поддержки за 12–18 месяцев. Это не означает, что автоматизация поддержки была ошибкой; это означает, что метрика выбрана не на том уровне.

Тесты на трёх ролях

Из кейсов следуют три практические проверки для разных позиций — операционный руководитель, финансовый руководитель, совет директоров и инвесторы.

Операционный руководитель. Возьмите ключевую функцию, которую планируете автоматизировать, и попробуйте описать её без упоминания конкретного человека или должности — что именно должно случиться, при каком условии, и как понять, что результат достигнут. Если функция описывается только через «такой-то человек делает то-то» — компания живёт в плоскости должностей, и любая замена будет PR-операцией. Дополнительный тест: какой процент входящих кейсов укладывается в описанные классы случаев? Если меньше 70% — хвост слишком большой, чтобы строить на нём чисто агентный слой.

Финансовый руководитель. Отделите «среднюю часть распределения» — типовые случаи с измеримым ответом — от «хвостов» — нестандартных кейсов, где правильность определяется в диалоге. Если хвосты занимают существенную долю реальных запросов, экономия от замены людей агентами без слоя эскалации становится отрицательной — узнаётся это обычно через 6–12 месяцев, как у Klarna. Полезное правило: бюджет на новые роли (операторы агентных процессов, редакторы машинного вывода, ревьюеры эскалаций) должен закладываться **сверху** до запуска агентного слоя, а не «когда выяснится, что они нужны».

Совет директоров и инвесторы. В отчётности компании, заявившей агрессивную ИИ-перестройку, ищите два сигнала. Первый — описание операционных единиц в языке функций, а не должностей: это признак реальной перестройки. Второй — структура hire-back или новых ролей вокруг эскалаций,

суждения и аккаунт-менеджмента, а не возврат тех же ролей, что и до автоматизации. Это признак коррекции архитектуры, а не провала эксперимента. Если в годовом отчёте AI-инициатива описана через «сокращено N человек, сэкономлено $M», а не через «перерисованы такие-то операционные единицы и появились такие-то новые роли» — компания, скорее всего, ещё на стадии Klarna 2024 года.

Сигналы 2026 года

Три публичных индикатора покажут, превращается ли паттерн перестройки в массовую практику.

Первый — язык публичной коммуникации зрелых компаний. Если CEO начинает рассказывать не «сколько мы сократили», а «как мы перерисовали границы ответственности» — это сигнал, что урок Klarna дошёл до уровня корпоративной нормы. Пока преобладает первый язык, и кейс Block февраля 2026-го показывает, что даже после публичного отката Klarna инерция нарратива «сокращаем людей через ИИ» сильнее, чем инерция нарратива «переписываем единицы».

Второй — структура hire-back в компаниях, прошедших цикл агрессивной автоматизации. Если возвращаемые позиции сгруппированы вокруг эскалаций, аккаунт-менеджмента и суждения — это нормальная коррекция. Если возвращаются те же роли, что и до отката, — это признание провала эксперимента. Klarna в 2025 году сделала шаг в сторону первого варианта (явно описывая возврат как работу с нестандартными случаями), но в публичной коммуникации язык всё ещё ближе ко второму.

Третий — появление новых операционных ролей, которых раньше не было: операторы агентных рабочих процессов, редакторы машинного вывода, аудиторы агентных решений, тренеры памяти агентов. Чем больше таких ролей оформляется в стабильные должности, тем понятнее, что граница между человеком и агентом стабилизировалась на конкретных функциях, а не дрейфует. Сейчас эти роли существуют у вертикальных AI-компаний и почти не существуют у энтерпрайз-клиентов, которые в массовом порядке закупают агентные платформы.

Главное

- В 2024–2026 годах три публичных кейса — Klarna, Duolingo, Block — прошли (или проходят) цикл «громкая ИИ-замена → тихая коррекция → переформулировка». Это не свидетельство провала технологии, а карта типичной ошибки.
- Граница между тем, что осталось за агентами, и тем, что вернулось людям, прошла не по сложности задачи, а по формализуемости: измеримый выход, повторяемый вход, низкая цена ошибки.
- Все три компании совершили операцию в плоскости задач — заменили людей на агентов на тех же ролях. Структурная перестройка начинается с описания операционных единиц как зон ответственности, а не должностей.

- Откат всегда вернёт людей на верхний слой суждения и эскалаций. Вопрос только в том, проходит ли этот возврат через публичный кризис коммуникации, как у Klarna, или планируется как уточнение архитектуры с самого начала.
- Метрика «процент сокращённых» — плохая прокси трансформации. Лучшая — доля формализуемого труда, покрытого агентами, плюс структура новых ролей, оформленных за тот же период.

FAQ

Что именно сделала Klarna в 2024 году с поддержкой? В феврале 2024-го Klarna в совместной публикации с OpenAI и в посте Семятковски в X сообщила, что её AI-ассистент за первый месяц обработал 2,3 млн диалогов — две трети клиентского сервиса — и выполнил объём работы, эквивалентный нагрузке 700 штатных агентов поддержки (OpenAI). К маю 2025-го компания признала, что зашла дальше, чем стоило, и возобновила найм — но на верхний слой работы с нестандартными случаями (Bloomberg).

Был ли у Duolingo полный отказ от человеческих переводчиков? Нет. В апреле 2025-го компания объявила о переходе на «AI-first» режим работы и сокращении подрядчиков, чья работа поддавалась автоматизации (The Verge). После публичной реакции — включая удаление официальных аккаунтов Duolingo из социальных сетей на короткое время — часть функций вернулась людям в виде языковых редакторов и культурных проверяющих, а сам термин «AI-first» был переформулирован к августу 2025-го (The New York Times).

Что произошло с Block? В феврале 2026-го Block анонсировал сокращение около 40% штата — примерно 4 000 человек, — связав это с интеграцией «intelligence tools» во внутренние процессы и автоматизацией повторяемого труда. Аналитики и пресса указали и на параллельный контекст: слабый крипто-рынок, давление на прибыль, удобство AI-нарратива как публичной оболочки сокращений (The Guardian). Публичного отката пока не было; коррекция, если она нужна, обычно проявляется через 6–12 месяцев.

Какая работа надёжно остаётся за агентами после всех откатов? Та, у которой измеримый выход, повторяемый вход и низкая цена единичной ошибки. Первичная обработка типовых запросов, перевод между языками с большим корпусом, рутинная аналитика по заранее определённым метрикам, кодогенерация по жёсткой спецификации. Всё остальное возвращается под человека или работает как «агент плюс человек».

Что должно быть в компании, чтобы избежать откатов? Описание ключевых функций как зон ответственности, а не должностей. Явное разделение «средней части распределения» (кандидат на агента) и «хвостов» (зона суждения). Заранее спроектированный слой эскалаций. Бюджет на новые роли, оформляемый до запуска агентного слоя, а не после. И язык внутренней и внешней коммуникации, в котором найм и автоматизация не противопоставлены, а описаны как варианты исполнения одной и той же функции.

Почему именно эти три компании, а не другие? Потому что в каждой из них публично, под именем и с цифрой, прошёл громкий шаг автоматизации, на который рынок откликнулся. Это даёт редкую возможность сравнить три отдельных компании по одной оси — что они сделали, что сказали и что вернулось обратно. Большинство аналогичных случаев в энтерпрайзе проходит без громкого публичного аккорда, и проследить откат снаружи невозможно.