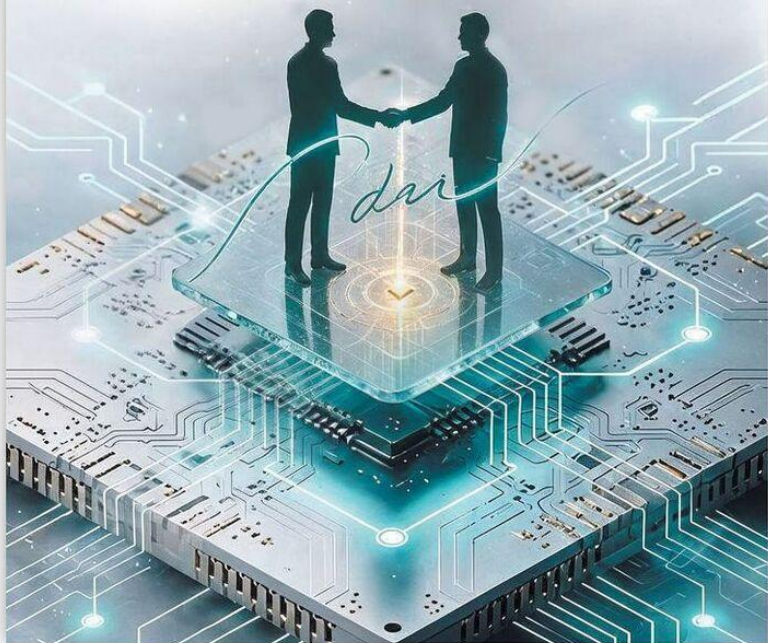


12+ ДЕНИСЕРМИЛОВ

КОГДА ИНТЕЛЛЕКТ СТАЛ ДЕШЁВЫМ

ТОМ 3. ДОВЕРИЕ

КАК СОЗДАВАТЬ ОСНОВАНИЯ ДЛЯ ДОВЕРИЯ, КОГДА
МЕНЯЮТСЯ ДОКАЗАТЕЛЬСТВА, СДЕЛКИ И ОТНОШЕНИЯ



Денис Ермилов

**Когда интеллект стал дешёвым.
Том 3: Доверие. Как создавать
основания для доверия, когда
меняются доказательства,
сделки и отношения**

<https://litres.ru/74589392>

ISBN 9785007141222

Аннотация

ИИ за секунды создаёт правдоподобные тексты, голоса, изображения, документы и профили. Привычных признаков подлинности уже недостаточно, а цена решения, чему и кому верить, растёт — возникает «налог недоверия». Третий том показывает, как проверять происхождение, личность и полномочия и искать независимое подтверждение. «Компас доверия» помогает выбирать глубину проверки по цене ошибки — от дипфейков, найма и команд до ИИ-агентов, платежей, институтов и профессиональной идентичности.

Содержание

От автора. Проверка на человека	9
Когда изображение начинает оправдываться	11
Доверие дорожает	14
Почему я пишу об этом	18
Без паники и волшебного детектора	20
Как устроен третий том	22
Как пользоваться третьим томом	24
Книга и Практикум	25
Если вы не читали первые два тома	27
Три маршрута чтения	28
Карта Тома 3	29
ЧАСТЬ V	31
Глава 1. Убедительно — не значит подлинно	31
У проблемы уже есть имя	33
Что именно сломалось	35
Работа, которая выглядит работой	37
Половина трафика — не то, о чём вы подумали	39
Собеседник, которого не было	42
Когда подозрение само становится доказательством	44
Не вся синтетика — обман	46
Три вопроса доверия	48

Проверять не странное, а дорогое	50
Практикум к книге	52
Глава за минуту	52
К следующей главе	54
Глава 2. Поддельное присутствие	55
Четыре минуты чужого голоса	57
Голос теперь собирается из ваших же записей	58
Совещание, которого не было	61
Там, где подделка проиграла	62
Жертва не похожа на жертву	64
Новый реквизит старого спектакля	67
Процедуры за ноль рублей	69
Практикум к книге	72
Глава за минуту	72
К следующей главе	74
Глава 3. Обратный тест Тьюринга	75
Монетка, которую мы считаем чутьём	78
Судья, который сам провалил экзамен	80
Цена ложного срабатывания — чья-то жизнь	82
Сорок один процент и девяносто одна сотая	85
Если ничего нельзя доказать — всё можно отрицать	88
След работает лучше клятвы	90

Практикум к книге	93
Глава за минуту	93
К следующей главе	95
Глава 4. Синтетическая репутация	96
Датчик, который мы не выбирали	99
Ломают по прейскуранту, чинят по счётчику	101
У толпы есть оптовый склад	103
Толпа научилась разговаривать	105
Конец ознакомительного фрагмента.	106

**Когда интеллект
стал дешёвым
Том 3: Доверие. Как
создавать основания
для доверия, когда
меняются доказательства,
сделки и отношения**

Денис Игоревич Ермилов

© Денис Игоревич Ермилов, 2026

ISBN 978-5-0071-4122-2 (т. 3)

ISBN 978-5-0071-3923-6

Создано в интеллектуальной издательской системе Ridero

Этот том завершает разговор, начатый в первых двух. Сначала мы разбирались, что происходит с человеком, когда интеллект становится дешёвым. Затем — как это меняет команды, компании и рынки. Здесь вопрос ещё глубже: на чём

держится доверие, когда изображение, голос или текст могут быть правдоподобными, не будучи подлинными, — и что может сделать человек, чтобы в таком мире доверяли ему самому.

Эта книга не предлагает жить в режиме постоянного подозрения. Доверять — не значит перестать задавать вопросы. Это значит различать ситуации, где нужна проверка, и те, где можно спокойно жить, работать и быть рядом.

Но доверие — не только то, чего мы ждём от других. Его приходится создавать самим. В мире дешёвой убедительности одного хорошего впечатления уже недостаточно. Важнее быть последовательным: держать слово, признавать ошибки, отвечать за сказанное и оставлять след, который может проверить другой.

Я надеюсь, этот том поможет сохранить главное: не страх перед подделкой, а нормальную жизнь, в которой у доверия есть основания.

#

— Денис Ермилов

(Контакты автора вы найдёте в конце книги)

Моей жене Марине — за опору, тепло и счастье, которым она наполняет каждый мой день.

Моим детям, Софии и Матвею, — за то, что вы научили меня заново видеть радость в простых

вещах. Вы растёте в мире, который меняется быстрее прежнего; пусть в нём всегда будет место человеку — и пусть каждый из вас найдёт в нём свой путь.

От автора. Проверка на человека



Подлинность не исчезла. Она просто перестала помещаться в один файл

«Практически в каждой коммерческой сделке есть элемент доверия — особенно в сделке, растянутой во времени».

— *Кеннет Эрроу, лауреат Нобелевской премии по экономике*¹

У меня есть семейные фотографии, подлинность которых я никогда не проверял. Мне это и в голову не приходило. Я знаю, что запечатлённые на них моменты происходили, по-

¹ Kenneth J. Arrow, «Gifts and Exchanges», *Philosophy & Public Affairs*, 1 (4), 1972, с. 343—362; цитата — с. 357.

тому что помню людей, место и сам день. Снимок лишь возвращает меня туда, где я уже был.

Наверное, поэтому старые фотографии так дороги. Они хранят не изображение само по себе, а крупицу прошлого. Потёртый угол, выцветшие цвета, знакомый поворот головы — и фотография снова становится частью события. Ей не нужно доказывать мне свою подлинность. Свидетель рядом. Этот свидетель — я.

Но человек, который откроет тот же альбом через двадцать лет, окажется в другом положении. Он не помнит этого дня. Не знает, что осталось за рамкой кадра. Не слышал голосов людей на снимке. У него будет только изображение — убедительное, подробное, возможно, безупречное. И вопрос, которого у меня никогда не было: почему этому можно верить?

Ещё недавно такой вопрос казался излишним. Фотография могла быть постановочной, неудачно выбранной, даже отредактированной. Она никогда не была всей правдой. Но она оставалась следом события: свет попал в объектив, перед камерой что-то находилось, кто-то нажал кнопку. Снимок спорил с рассказом, уточнял его, иногда опровергал. Фотография могла умолчать о многом — и всё же закрывала спор о том, было ли то, что мы видим.

Теперь этот договор с изображением больше не работает автоматически.

Когда изображение начинает оправдываться

Осенью 2023 года Том Хэнкс предупредил подписчиков о ролике, который, по его словам, использовал ИИ-версию его образа для рекламы плана стоматологического обслуживания. Актёру пришлось объяснять очевидное: он в этой рекламе не снимался и не имеет к ней отношения.²

В этой истории нет сложного мошеннического сюжета. Именно поэтому она так точна. Один из самых узнаваемых людей мира видит собственное лицо в чужой рекламе — и вынужден настоящим голосом опровергать подделку: осторожно, это не я. С тех пор такие опровержения стали почти рутиной публичной жизни. Раньше знаменитости извинялись за сказанное. Теперь всё чаще оправдываются за то, чего не говорили.

Но у знаменитости хотя бы есть аудитория, которая услышит опровержение. У обычного человека нет ни пресс-службы, ни узнаваемого аккаунта, ни запаса общественного внимания. Если его голос попросит родственника перевести деньги, а лицо появится в чужом видеозвонке, опровержение может прийти позже, чем платёж.

² ABC News и CBS News, 2 октября 2023 года: Том Хэнкс заявил, что не участвовал в рекламе, использовавшей ИИ-версию его образа.

Подделка — лишь первая половина проблемы. Вторая опаснее.

Если убедительное видео можно создать, то неудобное настоящее видео можно объявить созданным. Запись существует, но тот, кто не хочет за неё отвечать, говорит: «Это не я. Это нейросеть». Само знание о подделках начинает работать на лжеца. Юристы Роберт Чесни и Даниэль Ситрон назвали этот механизм «дивидендом лжеца»³. Чем лучше общество знает, что аудио и видео можно подделать, тем легче посеять сомнение в подлинной записи.⁴

Получается странный обмен. Машина научилась создавать правдоподобные следы событий, которых не было. Человек получил удобный способ отрицать события, которые были.

Фотография больше не закрывает спор. Точнее — одной фотографии для этого уже мало. То же происходит с голосом, видео, письмом, отзывом, резюме. Мы по-прежнему видим и слышим, но само впечатление перестало быть последней инстанцией. К нему нужен второй слой: откуда это взялось, кто за этим стоит, какой след остался кроме самого файла.

Мы привыкли думать, что кризис доверия начинается

³ Дивиденд лжеца — возможность подделок помогает отрицать и настоящую запись. Подробнее — «Словарь Тома 3».

⁴ Чесни и Ситрон, *California Law Review* (2019): liar's dividend («дивиденд лжеца») — возможность подделки помогает отрицать и подлинные записи.

там, где нас обманули. Теперь он может начаться раньше — в тот момент, когда правда и подделка выглядят одинаково убедительно, а одного впечатления уже недостаточно для действия.

Доверие дорожает

Первые два тома этой серии были о том, что происходит, когда интеллект становится дешёвым и доступным почти каждому.

В первом томе мы смотрели на человека. ИИ научился за минуты производить то, на что специалист тратил дни: текст, расчёт, презентацию, код, проект решения. Подешевел **output** — то, что можно сделать и предъявить. Дороже стал **outcome**: что изменилось после работы, какую пользу она принесла и кто отвечает за результат. Во втором томе тот же сдвиг поднялся на уровень компании: агенты вошли в процессы, команды и покупки, и бизнесу пришлось пересобирать работу вокруг результата, потому что дешёвые инструменты появились у всех. Если первые два тома вы не читали — ничего страшного: главное я только что напомнил.

Третий том задаёт вопрос фундаментальнее: кому и чему мы готовы доверять, когда машина уже не только создаёт контент, но и всё чаще действует от имени человека или компании?

Элемент доверия был в сделках всегда — Эрроу писал об этом ещё в 1972 году, задолго до нынешней волны генеративного ИИ. Новость не в том, что доверие появилось. Во всё большем числе ситуаций дорожает его проверка: нужны дополнительные шаги, независимые подтверждения, понятные

полномочия⁵ и право остановить действие до необратимого решения. Кто перед нами? От чьего имени он действует? Какие у него полномочия? Что останется доказательством, когда словам, изображению и голосу мало выглядеть убедительно?

Это разные проверки. Установить личность — ещё не значит подтвердить полномочия; проверить происхождение файла⁶ — ещё не значит доказать правдивость сказанного.

Эти вопросы уже входят в обычный день. Голос близкого человека звучит в трубке. «Клиент» присылает новые реквизиты. Кандидат выходит на собеседование по видеосвязи. В большинстве случаев перед нами будет настоящий человек, настоящее письмо, обычная просьба. Книга не исходит из того, что всё поддельно: нормальная жизнь держится на обычном доверии. Но там, где просьба неожиданна, действие необратимо, а цена ошибки высока, одного привычного сигнала уже мало.

Поэтому эта книга не учит тотальному недоверию. Она учит выбирать режим.

Иногда достаточно привычного «я доверяю»: цена ошибки мала, отношения проверены временем, ситуация обрати-

⁵ Полномочия — право действовать, подписывать или решать от имени человека, компании или системы; их проверяют отдельно от личности. Подробнее — «Словарь Тома 3».

⁶ Происхождение контента — история файла или сообщения: откуда оно пришло, кто его создал и какие следы это подтверждают. Подробнее — «Словарь Тома 3».

ма. Иногда нужен режим «я проверяю»: просьба неожиданна, решение нельзя отменить, на кону деньги, репутация или чужая судьба. Зрелое доверие не означает верить всем. Оно означает понимать, кому, в чём и до какой границы вы доверяете.

У этой задачи есть и обратная сторона — ей отдана вся финальная часть книги. Мало научиться проверять других. В мире дешёвых подделок каждому придётся заново доказывать собственную подлинность⁷ — клиенту, работодателю, партнёру, а иногда семье. Требовать «поверьте мне» будет всё труднее. Доверие придётся накапливать: оставлять понятный след решений, держать слово, признавать границы, давать людям возможность проверить то, что для них важно. Этот накопленный проверяемый след — ваш **капитал подлинности**⁸: актив, который строится годами и который нельзя одолжить или сгенерировать за вечер.

Аргумент философа Оноры О'Нил можно перевести на язык этой книги так: добиваться нужно не доверия как реакции, а надёжности — и давать другим понятные основания

⁷ Подлинность — проверяемая связь источника сообщения, его происхождения и ответственности за действие; она не равна правдивости каждого слова. Подробнее — «Словарь Тома 3».

⁸ Капитал подлинности — авторская рамка Тома 3: проверяемый след, который человек или бизнес накапливает годами (история решений, реальные отношения, подписанный канал, принятая ответственность). Это актив; его отдача — премия за подлинность. Подробнее — «Словарь Тома 3».

её проверить.⁹

Я называю это преимущество **премией за подлинность**¹⁰. Это не доплата за наклейку «сделано человеком», а экономия на проверке. Понятная история, реальные отношения и названная ответственность сокращают число сомнений, согласований и повторных проверок. Поэтому решение принимается быстрее и сразу ясно, кто за него отвечает.

Капитал подлинности — актив; премия за подлинность — его отдача. Здесь повторяется движение всей серии. В первом томе дешёвый output поднял цену outcome. Во втором налог недоверия начал расти у всех разом. А в этой книге сначала копится капитал — а потом он платит премией: меньшим числом сомнений, согласований и повторных проверок. Сначала след — потом доверие к следу.

Финальный том трилогии — об этом: как не утонуть в проверках и не стать лёгкой добычей; как отличать доверие от привычки верить; как самому стать тем, кому не приходится выпрашивать доверие. Не требовать веры — создавать для неё основания.

⁹ О«Нил предлагает спрашивать не «как повысить доверие», а «какие основания позволяют судить о надёжности».

¹⁰ Премия за подлинность — отдача от уже накопленного капитала подлинности: более быстрое решение, меньшие издержки проверки или готовность выбрать вас при прочих равных. Подробнее — «Словарь Тома 3».

Почему я пишу об этом

Больше двадцати лет я строю банковские платформы, платёжную инфраструктуру и цифровые системы, по которым движутся деньги. В такой работе вопрос «кто перед нами?» никогда не был философским. Если система ошиблась в человеке, документе или полномочиях, ошибка быстро превращается в платёж, убыток и расследование.

Банк редко видит человека так, как видим его мы за столом. Он видит следы: документы, устройства, операции, привычное и непривычное поведение. Один признак почти ничего не доказывает. Убедительность — тоже не доказательство: мошенническая история как раз старается выглядеть правдоподобно. Надёжность возникает, когда независимые следы сходятся, а при дорогой ошибке есть кому остановить действие и задать ещё один вопрос.

В работе с VIP-клиентами я не раз видел, как платёж на миллионы запускался письмом, а дополнительно подтверждался голосом. Это не нарушение само по себе: каналы и полномочия согласованы заранее, банк знает клиента и историю отношений. Но мошенник атакует именно эту привычность. Письмо приходит со знакомого адреса, просьба звучит срочно, а знакомый голос её подтверждает. В такой момент задача не в том, чтобы угадать по интонации, настоящий ли голос. Задача — остановить платёж и связаться с клиентом

по заранее известному каналу (который сложнее подделать). Один обратный звонок здесь ценнее десяти совпадений в тоне и стиле.¹¹

Раньше эту работу прятали внутри банков, платёжных систем и служб безопасности. Теперь её упрощённую версию осваивает каждый, у кого звонит телефон. Человек слышит знакомый голос и на несколько секунд сам становится антифрод-системой: совпадает ли просьба с нашими отношениями, можно ли перезвонить по известному номеру, почему деньги нужны именно сейчас?

Я не смотрю на это как футуролог. Для меня доверие — часть инфраструктуры. Пока она работает, её не замечают. Когда ломается, дорожает всё: сделка, разговор, найм, перевод денег, даже право сказать «это действительно был я».

В предыдущем томе я писал о рельсах, по которым агенты будут искать, выбирать и платить. Но рельсы без доверия ничего не перевозят. Агенту придётся доказать, кто его владелец и что ему разрешено. Компании — подтвердить, что за красивым интерфейсом существует реальный бизнес. Человеку — оставить после себя не поток контента, а историю, которую трудно подделать и легко проверить.

¹¹ Финансовая разведка Минфина США (FinCEN) и ФБР рекомендуют подтверждать критичные платёжные запросы по независимому каналу.

Без паники и волшебного детектора

Разговор об ИИ давно перегрет. Почти каждую неделю нам предрекают гибель профессий, образования, интернета — а порой и самой правды. От этого шума устаёшь. Но, возможно, в этой усталости есть польза: когда громкие обещания перестают впечатлять, наконец становится слышен тихий, но главный вопрос — кому я доверяю и на чём это доверие держится?

Здесь не будет тезиса «ничему нельзя верить» — он бесполезен и к тому же неверен. Большинство разговоров, снимков и людей остаются тем, чем кажутся; мир не превратился за одну ночь в зал кривых зеркал. Но цена отдельных ошибок выросла, а старые признаки уверенности ослабли.

Не будет и совета купить детектор¹² — сервис или алгоритм, который ищет признаки сгенерированного или изменённого текста, изображения, голоса либо видео. Такой инструмент бывает полезен, но его ответ — один сигнал, а не судья. Проверка держится не на одной кнопке, а на происхождении файла, независимых следах, праве остановить действие и решении человека.¹³

¹² Детектор оценивает свойства объекта, но не устанавливает автора: высокий или низкий балл сам по себе не доказывает, кто создал текст и как именно использовался ИИ. Подробнее — «Словарь Тома 3».

¹³ Национальный институт стандартов и технологий США (NIST), AI 100—

Обещать неуязвимость я тоже не стану — такой защиты нет ни у банка, ни у государства, ни у семьи. Способы генерации меняются, способы распознавания догоняют, затем всё начинается снова. Книга стареет медленнее, когда держится не за название сервиса, а за принцип.

Результат скромнее и полезнее: сделать обман дороже, заметнее и менее выгодным. Оставить доверие там, где оно экономит жизнь, и добавить проверку там, где одна ошибка ломает слишком многое.

И одно правило я принял для самой книги. Каждый сильный факт в ней доведён до первоисточника. Там, где доказательств достаточно, я пишу «подтверждено». Где есть основания, но остаётся неопределённость, — «вероятно». Где ответа нет, стоит честное «неизвестно». В книге о доверии по-другому нельзя.

Как устроен третий том

Сначала разберёмся, что именно сломалось. Почему убедительный текст больше не доказывает, что его написал человек. Как поддельный голос и лицо меняют бытовое доверие. Что происходит с отзывами, репутацией и доказательствами, когда их производство дешевеет. Во что обходится всеобщая проверка — человеку, компании и обществу.

Затем перейдём к тому, как доверие проверяют, признают и выращивают. Кто говорит? Чем отвечает? Кто подтвердит? Как устроены цифровая идентичность, происхождение контента, полномочия агентов и институты, которым приходится заново учиться различать настоящее? И почему отношения, история решений и право отвечать своим именем становятся активом, который нельзя собрать одним запросом к модели?

В финале всё сложится в рабочий компас: когда доверять, когда проверять и какой след оставлять после себя. Этот компас нужен, чтобы сохранить нормальную жизнь там, где привычного «я же вижу» уже недостаточно.

Как и в первых двух томах, принципы — здесь, в книге: они меняются медленно. А всё, что стареет быстрее, — сервисы, конкретные процедуры проверки, свежие примеры и шаблоны — живёт в Практикуме к книге, который я обновляю по мере того, как меняются технологии и способы про-

верки: <https://cheap-intelligence.vercel.app/>

Вернёмся к семейной фотографии. Для меня она настоящая, потому что рядом с ней стоит моя память. Тому, кто откроет альбом позже, понадобится ещё что-то: имя, дата, рассказ, связь с другими снимками, человек, которому он верит. Подлинность не исчезла. Она просто перестала помещаться в один файл.

Как пользоваться третьим томом

Короткая заметка — на минуту. Дальше начнётся сам том.

Книга и Практикум

Этот том не предлагает жить с включённой тревогой: проверять каждую фотографию, звонок и письмо. Он предлагает отличать три вещи: что перед нами, кто за это отвечает и в какой момент цена ошибки требует остановиться. В книге — принципы, которые меняются медленно: не путать убедительность с доказательством, выбирать глубину проверки и оставлять после себя след, которому можно доверять.

Быстрее всего устаревают конкретные схемы обмана, сервисы, настройки и списки признаков. Они живут в Практикуме к книге — обновляемом слое с шаблонами и рабочими процедурами. Там же выходит еженедельный дайджест «ИИ делает, человек отвечает» — о том, где ИИ встречается с реальностью. На него можно подписаться: он разбирает, что действительно изменилось в ИИ, бизнесе и работе, почему это важно и на какие источники опираются выводы. Адрес один на все три тома:

<https://cheap-intelligence.vercel.app/>



QR-код Практикума к книге

Откройте ссылку или наведите камеру на QR в печатном издании — устанавливать ничего не нужно. Когда в главе появляется задание, которое стоит выполнить самостоятельно, ссылка на него будет в блоке «Практикум к книге».

Если вы не читали первые два тома

Можно начать здесь. Первый том был о человеке рядом с дешёвым интеллектом: что в нём дорожает, когда готовый ответ почти ничего не стоит. Второй — о командах, компаниях и рынках, где этот дешёвый интеллект начинает действовать. Третий задаёт следующий вопрос: как жить и принимать решения, когда самого убедительного сигнала уже недостаточно.

Полезный минимум уже есть в «От автора»: output может быть дешёвым, outcome — дорогим. Теперь появляется следующий вопрос: сколько стоит уверенность, что на полученный результат вообще можно опереться? Чем дешевле становится производство ответов, тем важнее — и нередко дороже — их проверка.

Первые два тома показывают, как мы пришли к этой точке. Но для чтения третьего они не обязательны.

Три маршрута чтения

- **Когда нужно защитить обычную жизнь.** Начните с первой части. Она разбирает знакомые вещи — текст, голос, лицо, отзывы, привычное письмо — и показывает, почему проверять нужно не всё подряд, а то, где ошибка дорого обойдётся вам или близким.

- **Когда решение влияет на работу и деньги.** После первой части переходите к главам о проверяемости, инфраструктуре доверия, действиях от имени компании и цене хорошей репутации. Это маршрут для того, кто подписывает, платит, нанимает, продаёт или даёт системе право действовать, — и для того, кто хочет не только защититься, но и строить собственный проверяемый след, которому легко доверять.

- **Когда хочется увидеть целостную картину.** Читайте книгу целиком. Тогда видно движение книги: от «этому нельзя верить» — к более трудному и полезному вопросу «какие основания мне нужны, чтобы действовать без лишней паранойи».

Карта Тома 3

У тома два направления. Первое — научиться проверять других: отличать убедительное от настоящего, не попадаться и не паниковать. Второе, к которому ведёт вся финальная часть, — самому стать тем, кому верят: накопить свой **капитал подлинности** — проверяемый след решений и отношений, а также надёжный канал, за который вы отвечаете своим именем. Короткая формула, которая держит оба направления: не требовать веры — создавать для неё основания.

Логика тома проста: сначала — почему привычные сигналы доверия перестают работать, затем — на каких основаниях строить доверие, когда одной убедительности уже недостаточно.

- **Часть V. Мир, где всё может быть подделкой** — диагноз. Мы увидим, как копия разрушает старые короткие пути к доверию: «я сам видел», «я слышал голос», «у него много отзывов», «это пришло с привычного адреса».

- **Часть VI. Экономика доверия** — механика. Здесь доверие перестаёт быть настроением. Оно собирается из происхождения, полномочий, независимых следов, понятной ответственности и правил остановки.

- **Глава 13 и эпилог** — сборка. К финалу вопрос о подлинности становится частью более широкого вопроса о доверии: убедиться, что перед нами не подделка, важно, но это-

го мало, чтобы понять, кому и чему можно доверять. Здесь идеи книги складываются в Компас доверия — практический инструмент, который помогает решить, когда доверять, когда проверять глубже и где остановиться. Его задача — не превратить жизнь в бесконечное расследование, а вернуть доверие туда, где оно действительно заслужено.

Обозначения

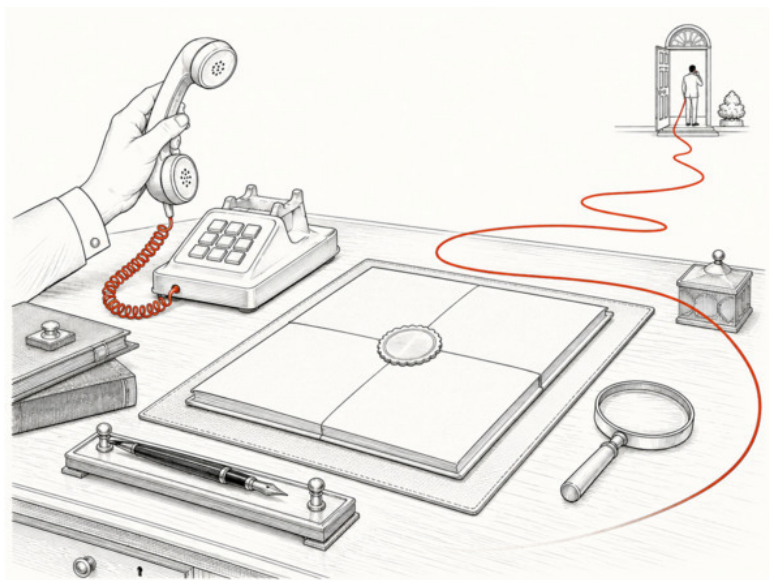
- **«Глава за минуту»** в конце главы собирает главную мысль, практический ход и границу уверенности: что в этой теме подтверждено, что вероятно и что пока неизвестно.
- **Сноски-цифры** ведут к источникам и ссылкам, сгруппированным по главам. Короткая сноска помогает читать дальше; полная карточка источника ждёт в конце тома.
- **«Практикум к книге» + QR** ведут к тому, что меняется быстрее книги: актуальным примерам, шаблонам и проверочным шагам.

Начните с первой главы. Она не требует верить книге на слово — для третьего тома это хороший старт.

ЧАСТЬ V

Мир, где всё может быть подделкой

Глава 1. Убедительно — не значит подлинно



Убедительность — свойство поверхности; проверка начинается за её пределами.

«Я никогда не воспринимал теорию мёртвого интернета так уж всерьёз — но, похоже, сейчас и правда очень много аккаунтов, которые ведут нейросети».

— Сэм Альтман, глава OpenAI, сентябрь 2025¹⁴

Вопрос главы: что мы на самом деле знаем о происхождении контента, который выглядит убедительно?

Обещание: научиться отделять качество от происхождения и правдивости — без паранойи и без игры в детектора по одному слову или тире.

Начнём с обычного вечернего дела. Вы ищете стоматологию — не себе, ребёнку. Открываете карту, находите клинику в двух кварталах: рейтинг 4,9, сорок отзывов. Отзывы хорошие, один сердечнее другого. С именами врачей. С подробностями: «дочка перестала бояться уколов», «приняли вечером в воскресенье», «сама доплатила за другую пломбу — не жалею». К некоторым приложены фотографии: кабинет, улыбчивый администратор, грамоты на стене.

Десять лет назад вы бы уже звонили в регистратуру. Пять лет назад — прикинули бы, не написал ли пару отзывов сам владелец. А сегодня вы ловите себя на вопросе нового поколения: есть ли за этими сорока историями хоть один живой

¹⁴ Сэм Альтман, пост в X, 3 сентября 2025 года: «LLM-run accounts» — «аккаунты, которые ведут нейросети».

человек?

Заметьте, что произошло. Отзывы никому не звонили. Ничего не требовали, никуда не торопили. Они просто лежали на странице и выглядели убедительно — самый тихий вид присутствия, какой бывает. Но привычного действия «прочитал — поверил» больше не получается. Между текстом и решением появляется дополнительный вопрос (которого три года назад просто не существовало): есть ли за этой «идеальной» страницей реальный опыт и можно ли подтвердить его где-то ещё? У такой паузы есть разумное основание. Форма отзыва больше не сообщает, кто и какой ценой её произвёл.

У проблемы уже есть имя

В декабре 2025 года редакторы Merriam-Webster, старейшего словаря Америки, выбрали словом года слово *slop*¹⁵. Старое английское слово когда-то означало жидкую грязь и пищевые отходы. Теперь словарь зафиксировал новое значение: низкокачественный цифровой контент, который обычно производится с помощью ИИ — быстро и в большом количестве.¹⁶ Почти одновременно австралийский словарь

¹⁵ Не каждый ИИ-текст — *slop*: решающими признаками служат низкое качество и потоковое производство; само участие модели этого ярлыка не требует. Подробнее — «Словарь Тома 3».

¹⁶ Merriam-Webster закрепил новое цифровое значение старого слова *slop*. Это оценка качества и массовости контента, а не способ доказать его машинное происхождение.

Macquarie назвал словом года AI slop — так решили и редакционный комитет, и читатели. А в январе 2026-го slop победило и в голосовании Американского диалектного общества, профессионального объединения лингвистов.¹⁷ Русский язык не отстал: «слоп» — у нас его успели растянуть в «нейрослоп» — вошёл в шорт-лист слова года по версии Грамоты.ру. Победил, впрочем, «зумер».¹⁸

За два года разговор о подлинности заметно сменил тон. В 2023-м словом года у Merriam-Webster было authentic — «подлинный»: редакторы связывали всплеск интереса к нему с ИИ, дипфейками и поиском того, чему ещё можно верить.¹⁹ В 2025-м люди уже искали не слово для настоящего, а ругательство для потока искусственного. Когда явление получает имя в словарях двух полушарий за один сезон, значит, оно уже стоит у каждого в телефоне.

Но эта глава не о том, что ИИ научился делать много плохих текстов и картинок. Плохого контента хватало всегда. Изменение глубже: **контент больше не доказывает сам себя**. Хорошо написано — ещё не значит, что писал специалист. Убедительно сказано — не значит, что говорит чело-

¹⁷ Macquarie Dictionary и American Dialect Society выбрали AI slop («ИИ-слоп») / slop («слоп») словом 2025 года; речь о двух институтах, а не о «трёх словарях».

¹⁸ Грамота.ру включила «слоп/нейрослоп» в шорт-лист 2025 года; победил «зумер», а происхождение термина связали с Саймоном Уиллисоном.

¹⁹ Merriam-Webster назвал authentic («подлинный, настоящий») словом 2023 года, связав интерес к нему в том числе с ИИ, дипфейками и идентичностью.

век. Профессионально оформлено — не значит, что документ прошёл профессиональную проверку.

Форма осталась. Подпись исчезла.

Подпись здесь — не имя внизу, а след: откуда пришло сообщение, кто за ним стоит и чем сказанное подтверждается за его пределами.

Что именно сломалось

В Томе 1 мы подробно разбирали, почему производство output — текста, изображения, презентации, кода, любого видимого результата работы — резко подешевело. Возвращаться к этой экономике не нужно. Для Тома 3 важно одно её следствие.

Раньше приличная форма была слабым, но полезным сигналом. Чтобы написать связный отраслевой отчёт, надо было хотя бы потратить время. Чтобы сделать убедительную иллюстрацию, требовались навык и инструменты. Чтобы подготовить десятки разных откликов на вакансии, нужен был человек, который прочитает описание каждой роли. Форма ничего не гарантировала, но за ней обычно стояла некоторая цена производства.

Почти до нуля подешевела не готовая работа, а правдоподобная первая версия. Один человек за вечер может создать объём, на который раньше требовалось несколько исполнителей. Один продавец — заполнить площадку отзыва-

ми. Один кандидат — собрать портфолио и ответы на тестовое задание, не владея тем, что в них описано. Не обязательно хорошо. Достаточно правдоподобно для нескольких минут чужого внимания. Но цена проверки, профессионального суждения, ответственности и исправления ошибки никуда не делась. Именно поэтому дешёвая форма способна сделать дорогим следующий шаг — решение на её основе.

Я знаю эту ловушку по работе с банковскими и платёжными системами. Такой опыт приучает: чем убедительнее документ, тем меньше хочется судить только по форме. Иногда подделка действительно выглядит слишком безупречной. Но это лишь подсказка, не доказательство. Помарка не гарантирует подлинность, а гладкость не доказывает обман. Надёжнее проверять биографию документа: по какому каналу он пришёл, совпадает ли получатель с историей платежей, кто и каким способом подтвердил новые реквизиты.

В отчёте австралийского регулятора за 2021 год приведено обращение покупательницы недвижимости. Реквизиты пришли будто бы от её юриста — с правильного адреса и внутри настоящей цепочки писем. Два банка провели два перевода: 90 и 350 тысяч австралийских долларов. Третий остановил следующую операцию и позвонил клиентке. После отдельного звонка юристу выяснилось: реквизиты в письме были подменены.²⁰

²⁰ Комиссия по защите потребителей Австралии (ACCC), *Targeting Scams 2021*, с. 22: два перевода прошли; третий банк остановил. Финансовая разведка США

#

В этом случае мошенничество не выдала пометка. Снаружи всё выглядело настоящим. Спасла проверка за пределами сообщения: остановка операции и звонок по независимо-му каналу. Убедительность можно создать на поверхности. Проверка начинается за её пределами.

Представьте три аналитические записки. Первую специалист написал сам после недели работы. Вторую он собрал вместе с ИИ, проверил цифры и подписал своим именем. Третью сгенерировал человек, который не разбирается в теме, но хорошо умеет задавать модели вопросы. Все три выглядят одинаково уверенно. Иногда третья — даже лучше первой: модель не устала и не сомневается вслух.

Отсюда возникает соблазн объявить подозрительным вообще всё синтетическое. Понятная реакция — и тупиковая. Текст, созданный с помощью ИИ, может быть честным и проверенным. Отзыв, написанный человеком от первого до последнего слова, может быть заказной подделкой. **Способ производства, происхождение и правдивость — три разные вещи.** Если их смешать, мы получим не защиту от обмана, а всеобщую подозрительность.

Работа, которая выглядит работой

Та же гладкая форма появилась и в рабочих документах:

(FinCEN): новые реквизиты подтверждают по нескольким каналам.

письмах, отчётах и презентациях.

Осенью 2025 года авторы исследования, опубликованного в *Harvard Business Review*, дали этому явлению имя — *workslop*²¹: рабочий контент, сделанный ИИ, который выглядит выполненной задачей, но не содержит ничего, что двигало бы её вперёд. В их опросе четыре из десяти офисных сотрудников в США получали такое от коллег за последний месяц, и на разбор каждого случая уходило в среднем почти два часа.²² Это данные опроса, а не перепись офисов.

Выглядит это так. Приходит документ на восемь страниц: аккуратные заголовки, уверенные выводы, ни одной орфографической ошибки. Минут через двадцать приходит и понимание: внутри нет ни одного решения. Вопросы не закрыты — текст лишь аккуратно обходит их. Работа выглядит сделанной, но делать её придётся вам. Отправитель сэконо-мил себе полчаса, купив их за два часа вашего времени. Запомните это ощущение: позже в книге мы выставим такому времени счёт.

Раньше плохая работа выглядела плохо. Теперь она выглядит хорошо.

²¹ От обычного *slop* *workslop* отличается адресатом издержек: даже аккуратный текст становится *workslop*, если получателю приходится восстанавливать отсутствующие решения, проверку и логику. Подробнее — «Словарь Тома 3».

²² Нидерхоффер и соавторы, *Harvard Business Review* (2025): *workslop* («рабочий слоп») выглядит работой, но переносит разбор и издержки на получателя; данные основаны на опросе.

Половина трафика — не то, о чём вы подумали

В разговорах о синтетическом интернете часто мелькает «сильная» цифра: больше половины трафика в Сети уже создают не люди. В отчёте компании Imperva о ботах за 2025 год говорится, что в её данных за 2024-й автоматический трафик²³ впервые превысил человеческий и достиг 51%; на «плохих» ботов пришлось 37%.²⁴ Imperva продаёт защиту от ботов, так что перед нами отраслевой отчёт заинтересованного поставщика, а не перепись интернета. Но как сигнал масштаба он важен.

С тех пор масштаб только вырос. В июне 2026 года о том же сообщила Cloudflare — компания, через сеть которой проходит заметная часть мирового веб-трафика: автоматические обращения к страницам впервые обошли человеческие и составили около 57%.²⁵

Важно и другое: чего эта цифра **не** означает. Она не озна-

²³ Бот-трафик — автоматические запросы программ к сайтам и сервисам; он не равен доле контента, созданного ИИ. Подробнее — «Словарь Тома 3».

²⁴ Imperva/Thales, 2025 *Bad Bot Report* («Отчёт о вредоносных ботах»): автоматизация составила 51% наблюдаемого трафика; это вендорский отчёт о ботах, не доля ИИ-контента.

²⁵ Cloudflare, июнь 2026: автоматические обращения к веб-страницам впервые превысили человеческие — около 57%. Замер по сети одной компании, продающей защиту от ботов, и только по запросам к страницам.

чает, что больше половины статей, отзывов и фотографий созданы нейросетями. Трафик — это запросы к сайтам и сервисам. В нём живут поисковые роботы, сборщики цен, системы мониторинга, сканеры уязвимостей и множество другой служебной автоматики. И наоборот: текст, сочинённый моделью, может опубликовать живой человек с обычного браузера.

Ещё недавно на этом можно было закончить: доля большая, но за ней стояли обычные служебные программы, не имевшие отношения к нейросетям. Теперь это уже не так. Быстрее всего в этой доле прибавляет новая её часть — программы, которые ходят по сайтам не сами по себе, а **по заданию конкретного человека**: сравнивают, отбирают, заполняют формы. Их и разбирать нужно отдельно, и пугаться их не стоит: агент, честно собирающий для хозяина список подрядчиков, — не подделка и не мусор. Но он меняет ответ на вопрос, для кого вообще пишется страница. К этому мы вернёмся в конце книги, когда речь пойдёт уже не о том, как проверять других, а о том, как проверяют вас.

В спорах про «мёртвый интернет» всё это сваливается в одну кучу — и в ней тонет главное. Разложим на три потока.

Синтетический контент²⁶ — сам объект полностью или частично создан моделью: текст, изображение, голос, видео.

²⁶ Синтетическое происхождение само по себе не означает ложь или обман: важны раскрытие роли ИИ, человеческая проверка и тот, кто принимает ответственность. Подробнее — «Словарь Тома 3».

Автоматическое участие²⁷ — программа читает, публикует, голосует, покупает или отвечает без человека в каждом шаге.

Подделка — кто-то намеренно скрывает происхождение или выдаёт объект и личность за то, чем они не являются.

Эти круги пересекаются, но не совпадают. Помощник поддержки может честно сообщать, что он бот. Человек может вручную написать фальшивый отзыв. ИИ может помочь врачу объяснить пациенту сложный диагноз простыми словами — без всякой попытки выдать программу за врача. Обман — только третий поток. Но пока мы не различаем потоки, любой из них бросает тень на остальные.

У потока есть экономический стимул. В ноябре 2025 года компания Karwing, которая делает инструменты для работы с видео, вручную проанализировала по сто трендовых YouTube-каналов в каждой стране — пятнадцать тысяч каналов всего. 278 из них авторы анализа отнесли к каналам, публикующим только AI slop; вместе они набрали десятки миллиардов просмотров. Karwing оценила их возможный годовой рекламный доход примерно в 117 миллионов долларов по данным Social Blade. Это не аудит YouTube и не подтверждённая выручка. Но вывод уже достаточен: пока дешёвая генерация получает охват и допускается к монетиза-

²⁷ Автоматическое участие бывает открытым и полезным либо скрытым и вредным; отсутствие человека на каждом шаге ещё не означает подделку. Подробнее — «Словарь Тома 3».

ции, у части производителей остаётся стимул продолжать.²⁸ Ждать, что поток остановится по доброй воле, не приходится — у него положительная экономика.

Теория «мёртвого интернета» утверждает, что значительная часть сети уже состоит из ботов, которые пишут для ботов. Как фактическое описание интернета это утверждение остаётся недоказанным. Но как бытовая метафора она победила: мы начали главу словами создателя ChatGPT, который признал, что его собственная лента стала казаться ему нежилой. Цитата Альтмана в эпиграфе важна не как измерение, а как симптом нового восприятия: даже опытный участник рынка больше не уверен, где заканчивается человеческая речь и начинается автоматическая.

Человек открывает комментарии и впервые не уверен, есть ли там вообще собеседник. Это новая граница недоверия. Мы сомневаемся уже не только в сказанном, но и в том, был ли вообще тот, кому мы всё это говорили.

Собеседник, которого не было

Лучший пример сомнения в самом существовании собеседника появился не в лаборатории и не на выборах. Он появился на Reddit, в сообществе r/changemyview — «измени

²⁸ Karwing, *AI Slop Report* («Отчёт об ИИ-слопе», 2025): ручная выборка трендовых YouTube-каналов; просмотры, аудитория и доход взяты из внешних агрегаторов и остаются оценочными.

мою точку зрения». Туда приходят именно затем, чтобы спорить с незнакомцами. Автор публикует убеждение, другие участники приводят аргументы, а если мнение изменилось, автор отмечает убедительный ответ специальным знаком.

С ноября 2024-го по март 2025 года исследователи Цюрихского университета незаметно включили в эти разговоры ИИ-аккаунты. Те оставили больше 1700 комментариев. Некоторые представлялись людьми с конкретной биографией — например, жертвой насилия или консультантом по травме. Для части ответов система анализировала историю публикаций собеседника и подбирала аргументы под предполагаемый возраст, пол и политические взгляды.²⁹

Сообщество узнало об эксперименте постфактум. Модераторы называли его несанкционированным: участников не предупредили, а правила прямо запрещали выдавать машинный текст за человеческий. Reddit заблокировал связанные аккаунты. Исследователи извинились, а результаты так и не были опубликованы как научная работа.

В черновике исследования авторы сообщали, что ИИ-ответы получали знак изменённого мнения в три-шесть раз чаще человеческой базовой линии. Эту цифру нельзя превращать в закон «ИИ убеждает в шесть раз лучше человека»: эксперимент был этически спорным, методика — уязвимой, результат не прошёл рецензирование. Корректный вы-

²⁹ Модераторы r/changemyview, 2025: эксперимент Цюрихского университета был несанкционированным; заявленные результаты не прошли рецензирование.

вод уже и неприятнее: **в реальном сообществе синтетический собеседник смог достаточно убедительно носить человеческую биографию.**

Здесь подделали не только текст. Подделали основание, на котором текст получает вес.

Одно дело — прочитав аргумент без подписи. Другое — поверить рассказу человека, который якобы пережил то, о чём говорит. Опыт жертвы, профессиональная роль, возраст — всё это меняет наше отношение к словам. Мы слышим не просто довод. Мы слышим довод **от кого-то.**

Цюрихские аккаунты заняли это место без права на него. Именно поэтому реакция сообщества была не технологической, а человеческой. Пользователи возмущались не тем, что предложения составила модель. Их превратили в участников эксперимента и дали им выдуманных собеседников.

ИИ оказался убедительным не вопреки отсутствию личности. Он получил преимущество, потому что личность ему нарисовали.

Когда подозрение само становится доказательством

У кризиса происхождения есть обратная сторона. Если подделку трудно распознать, настоящий человек тоже рискует не пройти проверку.

В марте 2026 года крупный издатель Hachette отменил

американский выпуск хоррор-романа *Shy Girl* Миа Баллард и прекратил распространение британского издания. До этого читатели, рецензенты и авторы видеоразборов указывали на повторяющиеся конструкции, разрывы логики и другие признаки, которые считали машинным почерком. Издатель сообщил, что принял решение после собственной проверки текста.³⁰

Баллард отрицала, что использовала ИИ для написания романа. Она говорила, что с ИИ мог работать человек, которого она привлекла к редактированию ранней версии. В публичном поле не появилось цепочки происхождения, которая позволила бы внешнему наблюдателю восстановить, где заканчивался текст автора, что менял редактор и на каком этапе мог использоваться ИИ.

Мы не будем выносить приговор по этому делу. Оно важно именно потому, что однозначного ответа нет.

Издатель имел право оценить собственный риск и принять решение. Автор имела право отрицать обвинение. Читатели имели право задавать вопросы. Но все стороны оказались в мире, где стиль используется как улика, а исходных следов недостаточно, чтобы быстро закрыть спор.

Так дешёвая подделка облагает налогом настоящее. Уже недостаточно создать работу. Иногда надо ещё доказать, что путь её создания соответствует заявленному. Художник хра-

³⁰ The New York Times, 19 марта 2026 года: Hachette отменила выпуск «*Shy Girl*» из-за подозрений в использовании ИИ; автор возражала.

нит промежуточные слои. Фотограф — исходный файл. Исследователь — данные и ход расчёта. Автор — черновики и историю правок. Не потому, что каждый заранее виноват. Потому, что объект больше не может свидетельствовать за себя.

Когда доказательства происхождения нет, его место занимает стилистическая экспертиза толпы. А толпа умеет подозревать гораздо лучше, чем отменять собственное подозрение.

Не вся синтетика — обман

На этом месте легко написать книгу против ИИ. Она была бы эмоциональной, понятной и бесполезной.

Миллионы людей уже используют модели для черновиков, перевода, поиска формулировок, иллюстраций, анализа документов и кода. Граница между «написал сам» и «поручил генерацию модели» давно не проходит по одной кнопке. Один человек попросил модель поправить запятые. Другой — предложить структуру. Третий принял готовый ответ, не читая. Формально во всех трёх случаях «использовался ИИ». Фактически перед нами три разных отношения к работе.

И здесь обнаруживается двойной стандарт. Свой ИИ кажется инструментом, чужой — попыткой обмануть. Свой текст мы оцениваем по тому, сколько в него вложили после

генерации, чужой — по одному подозрительному обороту или длинному тире. Свой помощник «только ускорил». А чужой, разумеется, «всё сделал сам».

Этот спор нельзя решить подсчётом процентов машинного текста. Важнее контекст.

Если коллега с помощью ИИ сделал черновик внутреннего письма, проверил факты и подписал его своим именем — цена вопроса невелика. Если тот же коллега приложил к кредитному решению сгенерированное заключение, которого не читал, способ производства становится критичным. Если компания явно пометила автоматического помощника поддержки — это автоматизация. Если бот придумал политику возврата и подписался именем сотрудника — это уже подмена.

Не нужно требовать одинаковой «чистоты происхождения» от поздравления, медицинского заключения и распоряжения на перевод денег. Нужна другая дисциплина: чем дороже ошибка, тем сильнее доказательство.

Так вопрос «это сделал ИИ?» уступает более точным вопросам. Откуда это взялось? Кто сейчас говорит? Что подтверждает сказанное за пределами самого сообщения?

Три вопроса доверия

Назовём их тремя вопросами доверия³¹. Вокруг них будет крутиться вся эта книга — но увидеть их целиком полезно уже сейчас, в первой главе.

Вопрос	Что он закрывает	Чего не гарантирует сам по себе
Откуда это взялось?	Происхождение: сам объект, история его создания и изменений	Что содержание истинно
Кто говорит — и чем отвечает?	Связь сообщения с человеком, организацией или программой; полномочие и принятая ответственность	Что говорящий компетентен и честен
Кто подтвердит со стороны?	Независимый канал, первоисточник, свидетель, след	Абсолютную безошибочность

Первый вопрос — **откуда это взялось?** У фотографии это исходный файл и история обработки. У документа — автор, версия, подпись и журнал изменений. У отчёта — дан-

³¹ Три вопроса доверия — авторская рамка: откуда это взялось, кто говорит и чем отвечает, кто подтвердит со стороны. Подробнее — «Словарь Тома 3».

ные, ссылки и ход расчёта. Происхождение не делает объект правдивым: человек способен подписать ложь. Но оно возвращает связь между результатом и процессом его появления. И у честного ответа на этот вопрос всегда три зоны: **подтверждено, вероятно, неизвестно**. Сильная проверка отличается не тем, что закрывает всё, а тем, что не прячет непроверенный остаток за общей уверенностью.

Второй вопрос — **кто говорит и чем отвечает?** Важно не имя на экране, а основание говорить от этого имени — и готовность отвечать за сказанное. Один и тот же текст, опубликованный официальным каналом компании и свежим аккаунтом с похожим логотипом, — два разных объекта. Один и тот же голос, пришедший с незнакомого номера и подтверждённый при обратном звонке, — две разные степени уверенности. Этому будет посвящена отдельная часть книги.

Третий вопрос — **кто подтвердит со стороны?** Самый сильный ход проверки — выйти за пределы того, что вас убеждает. Письмо просит срочно сменить реквизиты — позвоните по известному номеру. Цитата выглядит безупречно — найдите первоисточник. Отзыв обещает чудо — посмотрите историю профиля. Видео требует немедленного решения — проверьте событие через канал, который в самом видео не указан.

Три вопроса — не обязательная процедура для каждого сообщения. Большую часть повседневного контента мы разумно потребляем без расследования: ошибка в смешном

ролике не стоит ничего. Ошибка в платёжном поручении, диагнозе, кадровом решении или публичном обвинении стоит дорого. Поэтому глубина проверки зависит не от того, насколько странно выглядит объект, а от цены ошибки³². Правдоподобный объект с дорогими последствиями требует больше проверки, чем неуклюжий объект без последствий.

Это непривычный разворот. Мы привыкли проверять то, что выглядит подозрительно. Теперь проверять надо прежде всего то, что просит нас сделать что-то дорогое.

Проверять не странное, а дорогое

Первая реакция на синтетический поток — стать детектором. Замечать одинаковый ритм, неестественные пальцы на картинках, любимые слова моделей, странные паузы в голосе. Иногда такие признаки помогают. Но строить на них систему доверия нельзя: модели меняются, признаки исчезают, а живые люди начинают попадать под подозрение за собственный стиль. Чем это заканчивается — тема третьей главы.

Есть ход надёжнее. В каждом из нас живут два режима. «Я доверяю» экономит силы и позволяет жить — и в большинстве случаев этот режим прав: мир не превратился в зал кривых зеркал. «Я проверяю» включается редко, но по де-

³² Цена ошибки — возможный ущерб от неверного решения; чем он выше, тем глубже должна быть проверка. Подробнее — «Словарь Тома 3».

лу. Беда в том, что переключаем мы их не тем тумблером. «Я доверяю» верит гладкой форме, потому что она приятна. «Я проверяю» азартно расследует мемы и опечатки, потому что они бросаются в глаза. Оба заняты не тем. Правильный тумблер — цена ошибки: к какому действию меня подталкивает этот объект и что я потеряю, если он окажется не тем, чем кажется. Просто прочитать — читайте. Переслать всей семье, изменить мнение о человеке, отдать данные, перевести деньги — здесь форма перестаёт быть аргументом, и начинаются три вопроса.

У этого правила есть зеркало, и о нём стоит сказать уже в первой главе. Теперь вы — не только тот, кто проверяет. Вы — тот, кого проверяют: ваш отчёт, ваш отклик на вакансию, ваш честный отзыв тоже читают через прищур. Как честно-му автору не выглядеть подделкой? Тем же способом, только в обратную сторону: показывайте происхождение. Черновики, версии, данные, источники, датированную историю работы. Не оправдание «я не пользовался нейросетью», а след, который легко проверить. Дальше в книге мы разберём это подробно; пока достаточно принципа: не заставляйте людей верить вам на слово — дайте им способ убедиться.

Если единственное доказательство подлинности сообщения — само сообщение, подлинность ещё не доказана.

Практикум к книге

В Практикуме к этой главе лежат:

- **минимальный пакет доверия** — пять листов на один вечер: с чего начать, когда проверять хочется всё, а времени нет;
- **рабочая тетрадь Тома 3** — маршрут тома: где проверяете вы, а где пометку «проверено: человек»³³ предъявляете вы сами.

Открыть Практикум: <https://cheap-intelligence.vercel.app/>

Глава за минуту

1. **Гладкая форма подешевела почти до нуля.** Красивый текст, картинка или профиль больше не доказывают ни усилия, ни компетентности, ни того, что за ними есть человек.

2. **Синтетика, автоматика и подделка — три разных потока.** Автоматическое может быть честным, человеческое — ложным; обман — только один из трёх потоков, и смешивать их вредно.

³³ «Проверено: человек» — отметка о человеческой проверке и принятой ответственности, а не обещание полностью ручного производства. Подробнее — «Словарь Тома 3».

3. Кризис бьёт в обе стороны. Синтетический собеседник принимается за человека, а настоящий автор не всегда может доказать, что работа настоящая.

4. Новая опора — не впечатление, а три вопроса: откуда взялось? кто говорит и чем отвечает? кто подтвердит со стороны? Глубину проверки задаёт цена ошибки, а не странность объекта.

Что забрать каждому.

- **Специалист:** оставляйте след работы. В мире дешёвой формы черновики, исходники, данные и история решений становятся частью профессионального капитала.

- **Руководитель:** назовите команде дорогие действия, для которых одного письма, голоса или красивого документа недостаточно. Общий призыв «быть внимательнее» не работает.

- **Предприниматель:** проектируйте проверяемость вместе с продуктом. Происхождение, полномочия и ответственность, которые клиент не может подтвердить, остаются рекламным обещанием — даже если внутри компании всё устроено честно.

Как объяснить близким.

- **Семье:** «Если письмо или знакомый голос просят денег, не продолжаем разговор по этому каналу — сами звоним по сохранённому номеру. Не по номеру и не по ссылке из сообщ-

щения. Проверяем не настроение собеседника, а человека и просьбу через независимый канал».

- Коллеге: «Красивый документ теперь ничего не доказывает. Смотри не на то, как сделано, а на то, откуда пришло и кто под этим подписался».

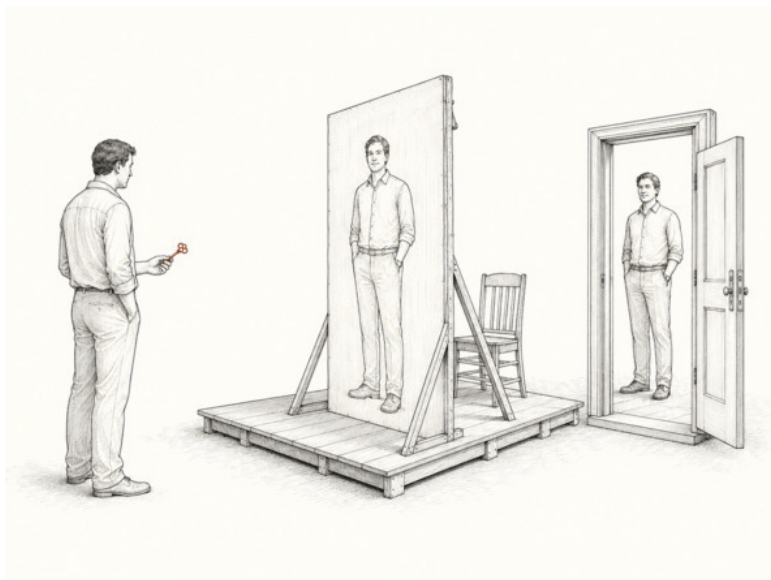
Раньше содержание могло служить доказательством подлинности. Теперь содержанию нужна собственная подпись.

К следующей главе

Контент первым потерял способность свидетельствовать за себя. Но у текста и отзывов есть милосердное свойство: их можно закрыть, отложить, не дочитать. Следующий объект такой паузы не даёт. Он звонит знакомым голосом, знает семейные подробности и просит решить всё прямо сейчас.

Голос долго был нашей самой бытовой подписью. Теперь — нет.

Глава 2. Поддельное присутствие



Подделке негде пройти там, где есть второй канал

«Если бы мне было интересно вложиться в мошенничество — это будет самая растущая отрасль всех времён».

— Уоррен Баффетт, глава Berkshire Hathaway, о мошенничестве с применением ИИ; собрание акционеров, май 2024³⁴

³⁴ Уоррен Баффетт, собрание Berkshire Hathaway 4 мая 2024 года: он назвал

Вопрос главы: можно ли по-прежнему верить голосу в трубке и лицу на экране — и чем их проверять, если уже нельзя?

Обещание: показать, что спасают не острый слух, не паника и не детекторы, а две старые процедуры за ноль рублей: заранее выбранная проверочная фраза и независимый обратный звонок. Пауза — не третья процедура, а условие, без которого обе не работают.

Начнём с двух кадров.

Кадр первый. Вечер, вы дома. В мессенджер приходит голосовое от брата: «Слушай, неудобно просить, но выручи до завтра». Голос его. Не «похожий» — а его: тембр, манера съедать окончания, привычное «ну ты понял». Вы с братом всегда общаетесь голосовыми, так что и формат не настораживает. Сумма подъёмная. Палец уже над кнопкой перевода.

Кадр второй. Пятница, конец рабочего дня. Видеозвонок от руководителя: он на экране, в своём кабинете, в своей манере — торопится, извиняется, просит провести платёж до выходных и никому пока не говорить, сделка чувствительная. Вы видите его лицо. Вы слышите его голос. Что тут проверять?

В прошлой главе подделка была тихой: лежала на странице и ждала, пока вы поверите. Эти двое не ждут. Они звонят, торопят и смотрят вам в глаза. Тысячи лет «я слышал его

голос» и «я видел его своими глазами» были последней инстанцией проверки — письмо можно подделать, печать можно подделать, но не живой разговор. Именно эта инстанция и сломалась. У обоих кадров теперь два одинаково возможных финала: настоящий брат и настоящий руководитель — или программа, собравшая их голос и лицо из настоящих записей. Это и есть **поддельное присутствие**³⁵: человек как будто участвует в разговоре или действии, хотя его там нет.

Один из этих кадров основан на реальной истории, и мы к ней ещё вернёмся. Но сначала — четыре минуты, которые объяснят, как устроен такой обман.

Четыре минуты чужого голоса

В январе 2023 года Дженнифер ДеСтефано из Аризоны позвонила с незнакомого номера. Она услышала рыдающий голос своей пятнадцатилетней дочери — и мужчину, который заявил, что дочь похищена. Требование: миллион долларов. Потом «скидка» — пятьдесят тысяч. Позже, давая показания в Сенате США, ДеСтефано скажет главное: это был её голос. Её плач, её всхлипы, её интонация.³⁶

³⁵ От обычной записи или статичного дипфейка поддельное присутствие отличается участием в действии: копия отвечает, просит, подтверждает или торопит от имени человека. Подробнее — «Словарь Тома 3».

³⁶ Показания Дженнифер ДеСтефано Сенату США, 13 июня 2023 года: голос дочери проверили независимым звонком; применение ИИ в этом звонке технически не подтверждено.

Посмотрите, из чего собрано это давление. Родной голос — чтобы отключить сомнение. Плач — чтобы отключить анализ. Чужой мужской голос с угрозами — чтобы включить страх. Счётчик времени и запрет звонить кому-либо — чтобы не осталось паузы. Всё решается внутри панического окна, пока тело кричит «спасай», а голова не успевает спросить «а точно ли это она?».

Спасла не техника и не хладнокровие. Рядом оказались люди: одна женщина вызвала полицию, другая набрала номер настоящей дочери. Та ответила через несколько минут — живая, невредимая, ничего не понимающая. Никакого похищения не было. ДеСтефано была уверена, что услышала клонированный голос дочери, однако публичного технического подтверждения применения ИИ именно в этом звонке нет. Зато известно, что обман развалился после независимого звонка настоящей дочери.

Четыре минуты — вот сколько прожил этот обман. Он не выдержал одного-единственного действия: второго звонка по другому номеру. Запомните это действие. Вся глава, по сути, о нём.

Голос теперь собирается из ваших же записей

Дальше — минимум техники: только то, что меняет бытовые решения.

Дипфейк — так называют голос, фото или видео, которые программа собирает из настоящих записей человека: получается цифровая копия, которая говорит и выглядит как он. Для голосовой копии уже не нужна студия: производители защитных систем говорят о нескольких секундах записи, более осторожные эксперты — о двадцати.³⁷ Где взять запись, объяснять не надо: голосовые сообщения, сторис, публичные выступления, поздравление в семейном чате. В 2025 году «Коммерсантъ» описывал и более прицельный способ добычи: обзвон под видом социологических опросов — человек пару минут отвечает на безобидные вопросы и оставляет мошенникам образец своего голоса.³⁸

Второе изменение серьёзнее. Подделка научилась работать в прямом эфире. МВД России летом 2026 года предупредило: голос «сына» или «сотрудника ведомства» может синтезироваться прямо во время звонка — злоумышленник говорит, а вы слышите родного человека, отвечающего на ваши реплики.³⁹ То же с лицом. Ещё в 2023 году в китайском Фучжоу владелец технологической компании перевёл 4,3 миллиона юаней — около 610 тысяч долларов — после видеозвонка от «друга», который просил помочь с депози-

³⁷ «Коммерсантъ», июнь 2025 года: эксперты оценили нужный образец голоса примерно в 20 секунд; это оценка, а не техническая константа.

³⁸ «Коммерсантъ», июнь 2025 года: мошенники записывали голос жертвы под видом социологического опроса; это оценка экспертов, не регулятора.

³⁹ МВД России в изложении «Известий», 6 июля 2026 года: голос родственника можно синтезировать прямо во время звонка.

том на тендер. Лицо было лицом друга, голос — голосом друга, подмена шла в реальном времени. У истории, впрочем, есть и вторая половина: обман вскрылся быстро, и после обращения в полицию большую часть суммы успели заморозить и вернуть.⁴⁰

Почему это работает так безотказно? Потому что узнавание голоса — не рассуждение, а рефлекс. Своих мы узнаём по первому слогу, по дыханию, по смешку в конце фразы — раньше, чем успеваем что-либо проверить. Подделка давит именно на этот древний, довербальный слой доверия, куда логика просто не успевает.

Назовём явление по имени. Это не подделка документа и не подделка текста — это **поддельное присутствие**: имитация самого факта, что вы говорили с человеком. Ломается не отдельное сообщение, а способ проверки, которым мы закрывали все остальные сомнения. «Созвонимся» и «увидимся по видео» были нашим последним «проверено». Теперь и у присутствия появился вопрос из первой главы: откуда оно взялось?

И пока этот вопрос кажется теоретическим, он уже стоил одной компании двадцать пять миллионов долларов.

⁴⁰ Полиция Фучжоу и китайские СМИ, май 2023 года: предприниматель перевёл 4,3 млн юаней после видеозвонка с подменой лица и голоса «друга».

Совещание, которого не было

Январь 2024 года, Гонконг. Сотрудник финансовой службы международной инженерной компании Агур, той самой, что проектировала Сиднейскую оперу, получает письмо от финансового директора из лондонской штаб-квартиры: готовится секретная транзакция, нужна помощь.

Он засомневался. Письмо выглядело как классический фишинг — таким его учили не верить.

А дальше сомнение сняли самым надёжным из существующих способов: пригласили на видеоконференцию. Там был «финансовый директор» — с его лицом и голосом. Там были «коллеги». Совещание шло как совещание: обсуждение, поручения, рабочая рутина. После него сотрудник сделал пятнадцать переводов на общую сумму двести миллионов гонконгских долларов — около 25,6 миллиона долларов США. Обман вскрылся через неделю, когда он сверился со штаб-квартирой. Полиция Гонконга позже сформулировала итог одной фразой: на звонке все участники, кроме самого сотрудника, были подделкой — собранной из публично доступных видео и аудио реальных людей.⁴¹

Заметьте, что именно здесь сломалось. Человек не был

⁴¹ Правительство Гонконга, 26 июня 2024 года: HK\$200 млн перевели после заранее записанной дипфейк-конференции без живого диалога; Агур позднее подтвердила кейс.

беспечным — он сомневался. Он применил проверку, которой его учили: не верь письму, поговори с живыми людьми. Но проверку «увидеть и услышать» теперь тоже умеют подделывать. Взломали не компьютер — взломали совещание.

При этом уберём лишнюю мистику: по своей структуре это старый корпоративный обман «звонок от директора», которому десятки лет. Срочность, секретность, авторитет — всё по классической схеме. Новым был только реквизит: вместо поддельной подписи — поддельное лицо. Это плохая новость, потому что реквизит стал доступным. И хорошая, потому что против схемы, которая не изменилась, работают и старые защиты. Что подводит нас к историям с другим финалом.

Там, где подделка проиграла

Июль 2024 года. Топ-менеджеру Ferrari в WhatsApp (принадлежит Meta Platforms Inc., признанной экстремистской организацией; её деятельность в России запрещена) пишет «генеральный директор Бенедетто Винья» — с незнакомого номера, но с фотографией Виньи в профиле: срочная конфиденциальная сделка, нужна помощь. Затем звонок. Голос Виньи — с его точным южно-итальянским акцентом. Менеджера смутили едва заметные механические интонации, и он сделал ход, который стоит выучить наизусть: задал вопрос, ответ на который нельзя найти в интернете, — какую книгу

Винья рекомендовал ему на днях.

Звонивший повесил трубку.⁴²

Миллионная технология синтеза голоса проиграла вопросу за ноль рублей. В том же 2024-м похожая атака не прошла в рекламном холдинге WPP: мошенники завели фейковый аккаунт генерального директора Марка Рида, собрали видеовстречу с голосовой репликой и нарезкой из публичных выступлений — бдительность сотрудников атаку оставила.⁴³

А сингапурская история 2025 года показывает, что бывает, когда первая линия всё же пала. Финансовый директор международной компании перевёл почти полмиллиона долларов после видеозвонка с «руководством» — режиссура та же, что у Arup. Но когда «руководители» попросили ещё 1,4 миллиона, он заподозрил неладное и обратился в сингапурский центр по борьбе с мошенничеством. Дальше сработала машина: центры Сингапура и Гонконга отследили деньги по цепочке счетов и заморозили всю сумму.⁴⁴

Сложите три финала. Подделку остановил личный вопрос. Подделку не пустили дальше обученные люди. Подделку догнали антискам-центры двух стран. Ни в одном из слу-

⁴² Bloomberg и MIT Sloan Management Review, 2024: попытку имитации голоса CEO Ferrari остановил личный проверочный вопрос; потерь не было.

⁴³ The Guardian, 10 мая 2024 года: атаку на WPP с поддельным аккаунтом CEO, клоном голоса и видео с YouTube удалось остановить.

⁴⁴ Полиция Сингапура, апрель 2025 года: US\$499 тыс. после дипфейк-звонка удалось заморозить и вернуть компании.

чаев победа не была заслугой острого слуха — никто не «расслышал» синтез. Побеждала процедура. Доверие, как выясняется, чинится не бдительностью на пределе сил, а скучными правилами, которые работают, когда сил уже нет.

Осталось посмотреть, как этот жанр выглядит там, где он поставлен на поток.

Жертва не похожа на жертву

Дальше — не статистика дипфейков. Это статистика всей индустрии финансового мошенничества, в которую синтетический голос входит как новый реквизит. Она нужна здесь, чтобы увидеть масштаб старого процесса, а не измерить долю ИИ.

По данным Банка России, за 2024 год мошенники похитили у клиентов банков 27,5 миллиарда рублей — рекорд за всё время наблюдений, на три четверти больше, чем годом раньше. Вернуть удалось меньше десятой части.⁴⁵ Главный инструмент — не взлом, а социальная инженерия⁴⁶: человека убеждают перевести деньги самому. В отчётах ЦБ есть и две детали, которые редко попадают в заголовки. Первая: по

⁴⁵ Банк России учитывает операции без добровольного согласия клиента. В 2025 году объём хищений вырос до 29,3 млрд рублей (+6,4%): показатель 2024 года — срез динамики, а не верхняя точка.

⁴⁶ Социальная инженерия — обман, который заставляет человека самому раскрыть данные, перевести деньги или обойти защиту; атакуют решение человека, а не только систему.

итогах 2023 года каждый четвёртый похищенный рубль оказался заёмным — жертву убеждают не просто отдать свои деньги, а взять для мошенников кредит.⁴⁷ Вторая — обратная сторона: попыток хищений банки за 2024 год отбили на 13,5 триллиона рублей — в сотни раз больше, чем было похищено. Оборона не проигрывает. Она просто не бывает стопроцентной, пока последнее решение принимает человек внутри панического окна.

Данные Банка России не подтверждают удобный образ «наивной бабушки». В опросе за 2024 год чаще всего пострадавшими были работающие горожане со средним доходом и образованием; отдельно регулятор отмечал повышенный интерес мошенников к экономически активным людям 25—64 лет. Вывод здесь не в новом «типичном лице». Статус, доход и цифровой опыт не дают иммунитета.⁴⁸

Жертвой чаще всего становится человек, который уверен, что уж с ним-то такого не случится. Уязвимость — не глупость и не возраст. Уязвимость — это правильно выбранный момент давления: усталость, спешка, страх за близкого. Мошенничество давно работает как индустрия со сценариями, ролями и планом продаж; синтетический голос просто доба-

⁴⁷ Показатель означает долю похищенной суммы, а не четверть потерпевших или случаев. Данные Банка России за 2023 год приведены в пересказе «Ведомостей» февраля 2024 года.

⁴⁸ Банк России, опрос за 2024 год: чаще страдали работающие горожанки со средним доходом и образованием; особенно уязвима экономически активная группа 25—64 лет.

вил ей новый инструмент в набор.

Первый эпизод перекликается с рассказом от первого лица, опубликованным на Пикабу в декабре 2024 года. Автор поста писала, что перевела 75 тысяч рублей после голосового сообщения, похожего на голос брата: «голос брата, ну очень похож... всё совпало». Совпали голос, манера и привычный канал связи — с братом они всегда общались голосовыми. Независимо проверить обстоятельства этой истории нельзя, поэтому это не доказательный кейс, а описание субъективного опыта. Банк, по словам автора, в возврате отказал.⁴⁹ Её вывод в конце поста — тот же, к которому мы идём всю главу: перезванивай, даже если голос родной.

Второй эпизод — видеозвонок руководителя — собран из случаев Agur, Фучжоу и Сингапура. Оба эпизода в начале главы опираются на реальные схемы, а не на фантазию.

И ещё одна деталь, о которой почти не говорят. У жертвы поддельного присутствия часто есть второй удар: стыд и публичное унижение. Француженка, полтора года переписывавшаяся с «Брэдом Питтом» — с ИИ-фотографиями, признаниями и историей болезни, — потеряла около 830 тысяч евро. Когда она рассказала об этом в телеэфире, волна издательств оказалась такой сильной, что канал снял интервью из сети.⁵⁰ Поэтому жертвы молчат, а масштаб мошенни-

⁴⁹ Пикабу, декабрь 2024 года: рассказ от первого лица о переводе 75 тыс. рублей после звонка голосом, похожим на голос брата; банк отказал в возврате.

⁵⁰ France 24 и NBC News, январь 2025 года: жительница Франции потеряла

чества остаётся невидимым. Толпа смеялась; наше правило должно быть обратным: над обманутым не смеются. Следующая атака может быть рассчитана уже на вашу слабость.

Новый реквизит старого спектакля

Теперь — против паники. Может показаться, что мир уже наводнён синтетическими голосами и лицами и любой звонок — лотерея. Цифры говорят другое.

Американская финансовая разведка FinCEN ещё в конце 2024 года разослала банкам официальное предупреждение о дипфейках. Показателен список схем, которые обслуживают сгенерированные документы и лица: подделка чеков, кредитные карты, займы, пособия — весь классический набор.⁵¹ Ничего нового злоумышленники не изобрели; они удешевили реквизит для старых постановок. Да и сама голосовая подделка в корпоративном обмане старше нынешней волны: ещё в 2021 году менеджер банка в ОАЭ перевёл 35 миллионов долларов после звонка «директора компании» с клонированным голосом — кейс из судебных документов, позже разобранный Европолом.⁵² Технология тогда была эк-

около €830 тыс. в мошеннической переписке с образом «Брэда Питта».

⁵¹ Финансовая разведка Минфина США (FinCEN), 13 ноября 2024 года: банки чаще сообщали о схемах с deepfake media («дипфейк-материалами»), обслуживающих обычное мошенничество с документами, счетами и займами.

⁵² Europol, «Facing Reality?» (2022): в ОАЭ перевод \$35 млн связали со звонком, имитировавшим голос директора компании.

зотикой. Сценарий — уже нет.

А доля подтверждённого ИИ в общем мошенничестве пока скромна. ФБР в отчёте о киберпреступности за 2025 год впервые выделило жалобы с применением ИИ отдельной строкой: около 893 миллионов долларов потерь — при общих заявленных потерях года почти в 21 миллиард. Это примерно четыре процента.⁵³ Само ФБР честно оговаривается: цифра занижена, многие жертвы просто не знают, что против них применяли синтез. Но и с поправкой картина ясна: девять из десяти похищенных долларов и рублей по-прежнему добывают без всяких нейросетей — обычным человеческим голосом, давлением и спешкой. Так что Баффетт из эпиграфа точен дважды: отрасль действительно растущая — и всё же основную её выручку пока делает не искусственный интеллект, а человеческая доверчивость. К громким цифрам из отчётов производителей защитных систем стоит относиться осторожно: «рост на сотни процентов» может отражать низкую стартовую базу, а сами компании продают защиту от описываемой угрозы.

Для нас здесь две практические новости. Первая: бояться нужно не технологии, а сценария — он не изменился. Срочность, секретность, авторитет, деньги: эти четыре ноты звучат и в старом мошенническом письме, и в видеозвонке под-

⁵³ Центр приёма жалоб на интернет-преступления ФБР США (IC3), отчёт за 2025 год: маркер AI Related («связано с ИИ») отмечен в 22 364 жалобах с потерями около \$893 млн; это не оценка доли ИИ во всём мошенничестве.

дельного директора. Вторая: раз сценарий старый, то и защита от него давно изобретена. Банки — а я двадцать с лишним лет строю банковские системы — живут в этой логике всегда: голосу и внешности там не верят по должностной инструкции, поэтому существуют кодовые слова, лимиты и подтверждение второй подписью. Всё, что требуется от нас, — перенести пару этих скучных правил из банка домой.

Процедуры за ноль рублей

Правил два. Они умещаются на стикере.

Первое: у семьи есть проверочная фраза. Договоритесь с семьёй — с детьми, с родителями, особенно со старшими — об одном слове или вопросе, которого нет ни в переписке, ни в соцсетях. Кличка первой собаки, блюдо из детства, что угодно, известное только вам. Любая просьба о деньгах или «срочной беде» голосом — сначала кодовое слово⁵⁴. Настоящий родственник ответит не задумываясь. Подделка знает о вас только то, что есть в записях, — а кодового слова там нет. Это семейная версия вопроса, который спас Ferrari: ответ нельзя загуглить.

Тонкость только одна. Банки задают «контрольные вопросы» десятилетиями — девичья фамилия матери, первый автомобиль, — и слабость их в том, что ответы давно живут в

⁵⁴ Кодовое слово — заранее выбранный семейный секрет для быстрой проверки срочного звонка или сообщения от близкого. Подробнее — «Словарь Тома 3».

анкетах и соцсетях. Хороший личный вопрос — тот, чьего ответа не существует ни в одной записи: о чём вы спорили в воскресенье, что стоит у бабушки на подоконнике, какую книгу посоветовали друг другу на днях. Проверяйте не фактами биографии, а общей памятью — её пока не скачать.

Второе: дорогое действие подтверждается вне исходного разговора⁵⁵. Просят перевести, продиктовать, подписать, срочно решить — вы прерываете разговор и сами звоните человеку по номеру, который знали раньше. Не по тому, с которого пришли к вам. Не в тот же мессенджер, откуда пишут. Второй, независимый канал — тот самый ход, что за четыре минуты развалил «похищение» дочери Де-Стефано. Здесь, кстати, редкий случай согласия регуляторов разных стран: и американские ведомства, и российское МВД советуют одно и то же — прервать разговор, перезвонить самому, задать личный вопрос.⁵⁶

К двум правилам стоит добавить третье — самое короткое: пауза. Всё, что вы теряете, прервав разговор, — две минуты. Всё, что можете потерять, не прервав, в этой главе измерялось миллионами. Настоящая беда переживёт ваш обратный звонок: родному человеку в реальной аварии от ва-

⁵⁵ Второй канал — независимый способ подтвердить дорогое действие: например, самостоятельно перезвонить по заранее известному номеру. Подробнее — «Словарь Тома 3».

⁵⁶ Федеральная торговая комиссия США (FTC), финансовая разведка Минфина США (FinCEN) и МВД России сходятся в одном правиле: прервать подозрительный разговор и перезвонить по заранее известному номеру.

шей проверки хуже не станет. Торопит только тот, для кого пауза смертельна, — потому что в паузе живёт проверка.

Обратите внимание, чего в этих правилах нет. В них нет требования распознать подделку. Распознавание по внешним признакам не даёт надёжного выигрыша: чем лучше становятся модели, тем меньше шансов угадать. Странная интонация, стеклянный взгляд или флаг детектора могут дать полезный повод остановиться — но не разрешение продолжить платёж. Решение о дорогом действии принимает не слух и не программа, а независимая проверка полномочий. Внутри нас, как мы говорили в первой главе, живут два режима. «Я доверяю» слышит родной голос — и верит, потому что тысячи лет этого хватало. «Я проверяю» щурится на каждый звонок и ищет артефакты синтеза. Оба заняты не тем: первый проверяет подлинность ушами, второй — подозрительностью. Тумблер по-прежнему один — цена ошибки. Разговор ни к чему не обязывает? Разговаривайте. Просят денег, данных, срочных решений — неважно, насколько родным был голос: включается второй канал.

Подделке негде пройти там, где есть действительно независимый второй канал. Ей есть где пройти только там, где вся проверка — само впечатление от разговора.

И у правила, как всегда в этой книге, есть зеркало. Теперь вы — не только тот, кто проверяет, но и тот, кого проверяют. Однажды ваш звонок покажется близким подозрительным — и это будет правильно. Не обижайтесь на прове-

рочную фразу и обратный звонок — радуйтесь им. Скажите родителям заранее: «Если „я“ позвоню и попрошу денег — скажите нашу фразу, положите трубку и перезвоните мне сами. Настоящий я не обижусь». Руководителю стоит сказать команде прямо: я никогда не попрошу перевод в мессенджере, а если попрошу — перезвоните мне по рабочему номеру, это правило. Второй подтверждённый канал — это не только ваша защита. Это подарок людям, которые хотят иметь возможность вам верить.

Практикум к книге

В Практикуме к этой главе лежат:

- **семейный протокол доверия** — проверочная фраза, разговор с родителями и детьми, типовые схемы звонков и памятка у телефона, которую можно распечатать;
- **личный периметр доверия** — второй канал для дорогих распоряжений на работе: что нельзя сделать по одному сообщению, каким бы убедительным оно ни было.

Открыть Практикум: <https://cheap-intelligence.vercel.app/>

Глава за минуту

1. **Голос и лицо перестали быть подписью.** Поддельное присутствие — имитация самого факта разговора с че-

ловеком: голос собирается из секунд ваших записей, лицо работает даже в прямом эфире видеозвонка.

2. Схема старая — реквизит новый. Срочность, секретность, авторитет: сценарий обмана не изменился. Доступные данные пока не позволяют надёжно посчитать долю ИИ во всём мошенничестве; бояться нужно не технологии, а панического окна.

3. У жертвы нет одного типичного лица. Возраст, доход и цифровой опыт не дают иммунитета; уязвимость — не глупость, а точно выбранный момент давления. Над обманутыми не смеются.

4. Спасают процедуры, а не слух. Семейная фраза останавливает разговор, дорогое действие подтверждается самостоятельно начатым контактом по известному каналу, а в компании — ещё и отдельным полномочием. Пауза даёт этому маршруту сработать.

Как объяснить близким.

- Семье: «Мам, если я позвоню и попрошу денег — скажи нашу фразу, положи трубку и перезвони мне сама по обычному номеру. Настоящий я не обижусь».

- Коллеге: «Голос и лицо в звонке больше ничего не доказывают. Дорогое распоряжение — только после обратного звонка по номеру, который ты знал раньше».

Раньше звонок был проверкой. Теперь звонок — то, что проверяют.

К следующей главе

Кодовое слово и обратный звонок защищают от подделки, которая приходит к вам. Но у кризиса доверия есть вторая сторона, и она несправедливее первой. Пока мы учимся ловить синтетические голоса, машины уже ловят людей: программы-детекторы объявляют настоящие тексты сгенерированными, студентов — списавшими, авторов — самозванцами. Презумпция невиновности в этих проверках не предусмотрена.

Раньше человеку достаточно было быть настоящим. Теперь это приходится доказывать.

Глава 3. Обратный тест Тьюринга



Детектор видит готовую страницу; черновики показывают путь к ней

«В интернете никто не знает, что ты собака».

— Питер Стайнер, подпись к карикатуре, *The New Yorker*, 1993⁵⁷

Вопрос главы: как доказать, что твою работу сделал ты,

⁵⁷ Питер Стайнер, подпись к карикатуре в *The New Yorker*, выпуск от 5 июля 1993 года.

— если «похоже на ИИ» стало обвинением, люди не отличают синтетику на глаз, а программы-судьи ошибаются?

Обещание: показать, что распознавание — как подбрасывание монеты, детектор — не приговор, и что защищает не клятва «я не пользовался нейросетью», а след работы, который собирается заранее и стоит ноль рублей.

Сначала — человек, который редактирует сам себя.

Писательница всю жизнь любила длинное тире — тот самый знак, которым мысль на секунду останавливается, чтобы ударить точнее. Теперь она вычитывает готовый текст и убирает тире одно за другим. Не потому, что разлюбила. Потому что в комментариях под прошлой публикацией ей уже написали: «Это писал ChatGPT». Длинное тире в англоязычных сетях успели окрестить «ChatGPT-тире» — и журналисты с писателями, любившие этот знак задолго до всяких нейросетей, стали вычищать его из собственных текстов, чтобы не отбиваться от обвинений.⁵⁸ В этой книге длинные тире останутся: сам по себе знак не доказывает машинного происхождения.

Посмотрите, что здесь происходит. Живой человек добровольно стирает из своего письма самого себя — интонацию, привычку, почерк, — чтобы выглядеть «достаточно человеческим» для толпы, которая ищет машинные приметы.

⁵⁸ Rolling Stone, 2025: длинное тире стали принимать за признак ChatGPT, и некоторые авторы начали избегать привычного знака.

Он проходит экзамен на человечность. И сдаёт его тем, что становится менее собой.

Семьдесят с лишним лет назад математик Алан Тьюринг предложил знаменитый мысленный эксперимент, который позже назовут тестом Тьюринга. Тьюринг заменил спор о мышлении проверкой поведенческой неотличимости. Человек ведёт текстовую переписку с человеком и машиной и пытается понять, кто есть кто. Тест был задачей машины, а человек сидел в кресле судьи. Теперь роли поменялись: экзамен сдаёт человек. Студент доказывает, что диплом писал сам. Автор — что роман его. Кандидат — что отклик на вакансию не сгенерирован. Здесь я называю это **обратным тестом Тьюринга**⁵⁹. В компьютерной безопасности так давно называют САРТСНА — проверку, в которой программа решает, человек ли перед ней. Здесь я использую выражение шире: это социальный экзамен, на котором человек доказывает уже не сайту, а организации, что работа действительно прошла через него. И у этого экзамена, в отличие от прямого, нет ни правил, ни экзаменатора, ни презумпции невиновности.

В прошлой главе мы выяснили, что голос и лицо больше не доказывают личность. Эта глава — о следующем повороте, и он бьёт по настоящему человеку: там подделка притво-

⁵⁹ Здесь выражение используется шире САРТСНА: экзаменатором становится работодатель, вуз или платформа, а доказательством — не один ответ, а след процесса и решений. Подробнее — «Словарь Тома 3».

рялась человеком, здесь человека принимают за подделку.

Монетка, которую мы считаем чутьём

Первая линия обороны очевидна: «я отличу». Гладкий безжизненный текст, одинаковый ритм, стеклянные картинки — неужели не видно?

Не видно. Это измерено. Метаанализ⁶⁰ 2024 года собрал 56 исследований с общим счётом больше 86 тысяч участников: людям показывали настоящее и синтетическое — тексты, фото, аудио, видео — и просили отличить. Средняя точность по всем форматам — 55,5 процента. Отдельно для текстов — 52 процента. Подбрасывание монетки даёт 50.⁶¹ Наше хвалёное чутьё на машину работает на уровне «орёл или решка».

Дальше — хуже. В эксперименте, опубликованном в *Scientific Reports*, обычные читатели опознавали стихи ИИ с точностью 46,6 процента — ниже случайной. Сгенерированные стихи они чаще принимали за человеческие, чем настоящие человеческие, — и оценивали их выше по ритму

⁶⁰ Метаанализ объединяет результаты многих опубликованных работ. Точность 55,5% в целом и 52% для текстов показывает: по внешнему виду человек не может точно определить, создан материал ИИ или человеком. Универсального теста «на ИИ» нет: вывод зависит от модели, материала и условий проверки.

⁶¹ Diel et al., 2024: метаанализ 56 исследований оценил среднюю точность распознавания синтетики людьми в 55,5%.

и красоте.⁶² А в исследовании Джонса и Бергена, опубликованном в PNAS в мае 2026 года, участники по пять минут одновременно переписывались с человеком и моделью. Когда GPT-4.5 заранее дали ролевую инструкцию — попросили играть правдоподобного человека, — её выбирали человеком в 73 процентах раундов. Без такой инструкции — в 36 процентах.⁶³

Ещё раз: в этом конкретном формате машина показалась участникам человечнее человека. Не потому, что она стала живой, — потому что она выучила наши ожидания от «живого» и попадает в них точнее, чем мы сами. У живого текста бывают усталость, тяжеловесные фразы после трудного дня, слишком ровные абзацы отличника. У синтетического — идеально дозированная человечность.

Вывод неприятный, но честный. Когда мы говорим «сразу видно, что писал ИИ», мы описываем не своё зрение, а свою уверенность. Различить на глаз нельзя. И всё, что мы строим на этом «сразу видно» — обвинения, оценки, отказы, — построено на монетке.

Казалось бы, решение напрашивается: раз глаз не справляется, позовём машину против машины.

⁶² Porter и Machery, *Scientific Reports* (2024): неэксперты различали стихи ИИ и людей с точностью 46,6% и чаще принимали машинные стихи за человеческие.

⁶³ Jones и Bergen, PNAS, 2026: в пятиминутных диалогах GPT-4.5 с ролевой инструкцией принимали за человека в 73% случаев; без неё — в 36% раундов.

Судья, который сам провалил экзамен

Программы-детекторы⁶⁴ обещают ровно это: загрузите текст — получите «процент сгенерированности»⁶⁵. Число выглядит как измерение. Твёрдое, объективное, бесстрастное.

Вот только сам главный производитель ИИ-текстов в мире в это число не верит. В начале 2023 года OpenAI, компания, создавшая ChatGPT, выпустила собственный публичный детектор ИИ-текста. Через полгода она тихо его закрыла с формулировкой «из-за низкой точности»: по собственным данным компании, инструмент опознавал лишь около четверти машинных текстов — и при этом в девяти случаях из ста объявлял машинным текст, написанный человеком.⁶⁶ Позже, отвечая на вопросы преподавателей, OpenAI сформулировала позицию прямее некуда: «Работают ли детекторы ИИ? Коротко — нет». Ни один инструмент, писала компания, не доказал, что надёжно отличает сгенерированный текст от чело-

⁶⁴ Детектор синтетического контента ищет признаки генерации или изменения; его результат — один сигнал для проверки, а не окончательный приговор. Подробнее — «Словарь Тома 3».

⁶⁵ Здесь и далее «процент детектора» — числовой вероятностный ответ модели. Интерфейс может показать его как долю, вероятность или уверенность, но это не установленная доля ИИ-текста и не доказательство авторства. Это сигнал для разбора, не единственное основание решения.

⁶⁶ OpenAI закрыла AI Text Classifier в июле 2023 года из-за низкой точности: он находил 26% ИИ-текстов и давал 9% ложных срабатываний.

веческого. И добавила деталь, которую стоило бы печатать на упаковке каждого детектора: при тестировании её собственный классификатор помечал как «написанные ИИ» тексты Шекспира и Декларацию независимости США.⁶⁷

Этот эпизод нам ещё пригодится. Компания, которой выгодно было бы продать миру и генератор, и детектор к нему, публично признала: детектора нет. Это не заговор скептиков — это позиция создателя технологии.

Почему же детекторами продолжают пользоваться школы, вузы, редакции и работодатели? Потому что спрос на определённости не исчезает от того, что определённости нет. Кто-то должен сказать «этот текст чист», — и «процент» на экране даёт утешительное ощущение, что сказала наука.

А теперь посмотрим, кого этот «процент» бьёт первым. Лян и соавторы в исследовании, опубликованном в журнале *Patterns*, прогнали через семь популярных детекторов эссе, написанные живыми людьми. На текстах носителей английского детекторы работали почти идеально. А эссе студентов, для которых английский — не родной, детекторы объявили сгенерированными в шести случаях из десяти.⁶⁸ Логика поломки проста: детектор ищет «слишком ровный» язык — небогатый словарь, правильные предсказуемые конструк-

⁶⁷ OpenAI, FAQ для педагогов: надёжность детекторов ИИ-текста не доказана, а ChatGPT не может установить, писал ли он конкретный текст.

⁶⁸ Liang et al., *Patterns* (2023): семь детекторов ошибочно назвали ИИ-текстом в среднем 61,3% эссе неносителей английского.

ции. Именно так пишет добросовестный человек на неродном языке. И именно так пишет старательный отличник, вычитавший свой текст три раза.

Детектор принимает аккуратность за признак машинности. Поэтому чаще всего он ошибается на самых добросовестных.

Цена ложного срабатывания — ЧЬЯ-ТО ЖИЗНЬ

До этого места история была теоретической: проценты, эксперименты, метаанализы. Теперь — три человека, для которых она теорией не была.

Марли Стивенс, студентка университета в Джорджии, сдала обычное эссе. Система Turnitin, которой пользовался вуз, пометила работу как сгенерированную. Дальше события развивались с бюрократической неотвратимостью: ноль за эссе, академический испытательный срок, потеря стипендии — около четырёх тысяч долларов. Из ИИ-инструментов Стивенс пользовалась только браузерным Grammarly — по сути, продвинутой проверкой орфографии, которую сам университет рекомендовал студентам. Её видео об этой истории собрало миллионы просмотров, а под ним — поток комментариев от студентов, помеченных так же несправедливо.

У истории два финала, и оба показательные. Grammarly

компенсировала студентке потерянные четыре тысячи. А спустя год, на волне таких же историй, выпустила новый продукт — Authorship: инструмент, которым студент может заранее записывать процесс своей работы, чтобы в случае обвинения доказать авторство. Вдумайтесь в самую категорию этого товара. Не «поймать машину» — а «доказать, что ты человек». На наших глазах родился рынок алиби.⁶⁹

История вторая — и она была бы смешной, если бы не была такой грустной. Май 2023 года, Университет A&M-Commerce в Техасе. Преподаватель, заподозрив группу в использовании ChatGPT, придумал проверку гениальной простоты: скопировал работы студентов в сам ChatGPT и спросил — это ты писал? Бот с готовностью «признался» во всех. Всей группе задержали оценки, под угрозой оказались дипломы; в системе оценок преподаватель написал: «Я не оцениваю ИИ-дерьмо».

Короткое техническое пояснение — в нём вся суть. Языковая модель не помнит, что она писала. У неё нет архива своих текстов и нет способа узнать свой продукт; спросить у ChatGPT «ты это писал?» — всё равно что спросить у печатной машинки, её ли это буквы. Бот отвечает не правду, а правдоподобное — мы разбирали это свойство в первой главе. Преподаватель использовал генератор правдоподобия

⁶⁹ Fast Company и EdSurge, 2024: студентку Марли Стивенс наказали после срабатывания Turnitin на эссе, обработанном Grammarly; позже ей компенсировали \$4000.

как детектор истины — и получил правдоподобные признания за всех.

Несколько студентов предъявили датированные версии Google Docs и показали, как работы менялись во времени. Затем университет разобрал случаи индивидуально и сообщил: никто не провалил курс, а этот эпизод не помешал ни одному студенту окончить университет.⁷⁰ Версии редактирования документов не вынесли оправдательный приговор сами. Они перевели разговор с чат-бота, который выдумывал собственное авторство, на процесс работы. Оцените асимметрию: обвинение стоило преподавателю одной вставки в чат-бот. Оправдание потребовало документированной истории черновиков.

История третья — про писателя под судом машин. В мае 2026 года разгорелся скандал вокруг Commonwealth Short Story Prize — международной литературной премии. Рассказ регионального победителя Джамира Назира показался подозрительным; издатель Granta, Сигрид Раусинг, прогнала текст через Claude — и модель ответила, что он «почти наверняка не создан человеком без помощи»; коммерческий детектор Pangram выдал «100% ИИ». Автор твёрдо отвечал, что писал сам. Фонд не стал отвечать детектором на детектор. Он обсудил с региональными победителями их твор-

⁷⁰ Rolling Stone и Washington Post, май 2023 года: преподаватель использовал ChatGPT как детектор; историю версий приняли как доказательство авторства студентов.

ческий процесс и изучил рабочие черновики, датированные документы и заметки. После этого организация заявила, что удовлетворена совокупностью свидетельств и считает, что ИИ не использовался при написании победивших рассказов. Через месяц Назир выиграл и главный приз года.⁷¹ Посмотрите, что именно его оправдало. Не детектор — детекторы как раз голосовали за обвинение и противоречили живым судьям. Оправдало его решение людей, готовых отвечать за свой вердикт. А если бы жюри поверило «проценту»? Мы уже видели этот финал: писательница, у которой издатель отозвал готовый роман по стилистическим приметам, реабилитации не дождалась. Обратный тест Тьюринга сам по себе не предусматривает оправдательного приговора — любое «я писал сам» звучит ровно так же, как звучало бы и враньё. Оправдание приходится строить из другого материала.

Подозрение теперь бесплатно. Дорого стоит только оправдание.

Сорок один процент и девятью одна сотая

В России этот жанр развернулся там, где ставки для обычного человека едва ли не выше всего, — в сезоне дипломов.

⁷¹ The Bookseller и другие издания, 2026: рассказ Джамира Назира проверяли Claude и Pangram, но после пересмотра автор сохранил победу в Commonwealth Short Story Prize.

Весной 2026 года «Коммерсантъ» описал массовое явление: вузовская система «Антиплагиат», обзаведясь модулем определения ИИ-текста, начала пометать как сгенерированные живые дипломные и курсовые работы. У одной студентки система нашла «96% якобы сгенерированного текста» в работе, которую она писала сама. Голоса студентов из той статьи стоит привести дословно: «Даже если находится 1% генерации, человека просто не допускают до защиты». И второй, ещё выразительнее — о том, во что превращается текст после «очистки» под детектор: «Делать из своего научного труда работу пятиклассника». Узнаёте? Это та же писательница, стирающая своё тире, — только теперь на кону не лайки, а допуск к защите. Разработчик системы при этом заявляет, что вероятность ошибки «не превышает 1%», — заявление вендора, независимой публичной проверки которого нет.⁷²

А теперь — история, ради которой этот блок написан. В первом отчёте «Антиплагиат. Вуз» указал 41,91 процента вероятного ИИ-контента в дипломе студентки четвёртого курса лингвистического факультета. Затем научный руководитель отдельно заявила, что более чем в 70 процентах текста намеренно нарушены логические связи, и расценила это как фальсификацию. При следующей проверке система

⁷² «Коммерсантъ», 28 апреля 2026 года: «Антиплагиат» давал высокие оценки ИИ-текста для самостоятельных дипломов; заявленная точность — оценка вендора.

показала 11,92 процента, но университет отказался принимать результат, сославшись на срок подачи. Один и тот же диплом, одна и та же система, три вердикта. Вуз отчисляет её по результатам первой проверки.

Летом 2026 года Верховный суд Татарстана признал отчисление незаконным и восстановил студентку. В решении сказано прямо: отчёт системы детекции «носит вероятностный характер», а «окончательное решение остаётся за экспертом» — то есть за человеком. С вуза взыскали компенсацию морального вреда и штраф.

Машина, дающая на один и тот же текст три разных ответа, — не измерительный прибор. Это генератор подозрений. И потребовался суд — институт с презумпцией невиновности, изобретённый задолго до нейросетей, — чтобы сказать очевидное: «процент» на экране не может быть приговором, приговор выносит только тот, кто способен за него отвечать.⁷³

Из этой истории вырастает правило, которое стоит унести из главы, даже если забудется всё остальное: **не выноси приговор по детектору**. «Процент» — повод задать вопросы. Основание для решения — только след работы и живой разбор. Это касается преподавателя, редактора, руководителя — любого, кто сегодня оказывается в кресле судьи. Ан-

⁷³ «Коммерсантъ», 8 июня 2026 года: суд восстановил студентку, указав на вероятностный характер отчёта «Антиплагиата» и необходимость экспертного решения.

тифрод-системы в банках, внутри которых я прожил больше двадцати лет, не работают по одной универсальной схеме. В зависимости от риска операция может пройти, потребовать дополнительного подтверждения, уйти на удержание или быть отклонена автоматически. Но у решения есть порог, владелец, журнал и маршрут оспаривания. Для текста нужен тот же принцип: не обязательный ручной клик на каждом шаге, а правило, по которому «процент» превращается — или не превращается — в действие с последствиями.

Если ничего нельзя доказать — всё можно отрицать

У неразличимости, которую мы разобрали, есть вторая половина, и она зеркальная. Если настоящее не отличить от подделки — то и подделку не отличить от настоящего. А значит, человек, пойманный на настоящем, может объявить его подделкой.

У этого приёма есть имя — «дивиденд лжеца»⁷⁴: выгода, которую сама возможность дипфейков приносит лжецам. Термин предложили в 2019 году американские правоведы Роберт Чесни и Даниэль Ситрон — заметьте, раньше, чем ChatGPT появился на свет. А в 2024 году журнал American

⁷⁴ Для дивиденда лжеца не требуется изготовить подделку: достаточно правдоподобно сослаться на существование технологии и перенести спор с поступка на запись. Подробнее — «Словарь Тома 3».

Political Science Review опубликовал серию из пяти экспериментов с участием более пятнадцати тысяч человек: политики, объявлявшие реальный компромат «фейком и дезинформацией», действительно удерживали поддержку избирателей — ложный крик «это подделка!» работал.⁷⁵ Исследователи добавляют отрезвляющую деталь: против видео приём работает хуже, чем против текста, — но граница смещается вместе с технологией.

Мы начали книгу с того, что фотография больше не закрывает спор. Вот цена этого в полный рост: исчезает не только доверие к подделке — исчезает общее доказательство события. Ещё недавно запись была последним словом: вот плёнка, вот подпись, вот скриншот. Теперь у любого последнего слова есть ответ: «сгенерировано». Обратный тест Тьюринга и дивиденд лжеца — две стороны одной монеты: в мире, где всё может быть синтетикой, невиновному нечем доказать невиновность, а виновному есть чем отрицать вину.

И чтобы замкнуть картину — короткая история о том, как сам страх перед синтетикой оказался отчасти синтетическим. Вы наверняка встречали цифру: «к 2026 году до 90 процентов онлайн-контента будет сгенерировано ИИ». Её цитировали серьёзные издания со ссылкой на Европол. Проследим происхождение. Прогноз дал руководитель компании Synthesia — вендора синтетического видео, то есть че-

⁷⁵ Chesney и Citron (2019); Schiff et al. (2024): заявления «это фейк» могут снижать ответственность за подлинный компромат.

ловек, продающий предмет прогноза, — в книге 2020 года. Из книги фраза перекочевала в отчёт Европола 2022 года в виде «по оценкам экспертов», из отчёта — в мировые СМИ уже как «данные Европола». А при обновлении отчёта в 2024 году Европол утверждение тихо убрал. Никаких измерений, подтверждающих «90 процентов», не существовало вовсе.⁷⁶ Цифра о синтетике сама прошла полный цикл синтетики: рождена интересом, легализована пересказом, растиражирована доверием к бренду. Первый вопрос книги — «откуда это взялось?» — работает и здесь.

След работает лучше клятвы

Соберём защиту. Она проста, ничего не стоит и начинает работать ещё до обвинения.

Сначала — для тех, кто судит. В каждом из нас, как мы помним, живут два режима. В этой главе они впервые меняются местами: «Я доверяю» и «Я проверяю» садятся в кресло судьи. И тумблер у них прежний — цена ошибки, только теперь это цена чужой судьбы. Заподозрить текст в машинности ничего не стоит. Поставить ноль, отклонить кандидата, отозвать премию, отчислить студентку — стоит чьей-то стипендии, репутации, года жизни. Поэтому правило судьи: чем дороже последствия обвинения, тем меньше права ре-

⁷⁶ Reason и версии отчёта Europol: прогноз о «90% ИИ-контента к 2026 году» восходит к вендору Synthesia и не подтверждён измерениями.

шать имеет «процент» на экране. Сомневаешься — спроси след: черновики, версии, историю правок, данные. Их предъ-
явление занимает минуты. А их отсутствие — само по себе
информация, только совсем другого веса, чем «41,91%».

Теперь — для тех, кого судят. То есть для всех. Здесь у
правила прежнее зеркало: вы — не только тот, кто проверя-
ет, но и тот, кого проверяют. И лучший момент готовить за-
щиту — до обвинения. **Сохраняйте след работы.** Пишите
важные тексты там, где остаётся история версий, — Google
Docs и Word делают это сами. Не удаляйте черновики и ис-
ходники: у фотографа это RAW-файл⁷⁷, у аналитика — дан-
ные и расчёт, у разработчика — история коммитов, у автора
— папка с версиями. Датируйте этапы. Письмо с приложен-
ным черновиком, защищённая история документа или ком-
мит создают дополнительную временную отметку — не но-
тариальное заверение и не доказательство сами по себе. Чем
дороже возможный спор, тем важнее не один артефакт, а со-
гласованность нескольких: версии совпадают с источниками,
расчёт воспроизводится, а автор способен объяснить клю-
чевые решения. Это займёт меньше усилий, чем один спор
с детектором, — потому что след собирается привычкой, а
спор съедает недели.

Сразу оговорка: история версий не доказывает, что ИИ не

⁷⁷ RAW-файл — исходный файл с камеры с минимальной обработкой; он хра-
нит больше данных съёмки, чем готовый JPEG, и помогает подтвердить проис-
хождение фотографии.

использовался. Его можно подключать поэтапно; самостоятельно написанный текст можно перенести из другого редактора; человек может работать без облачного журнала. След нужен не для нового автоматического приговора. Он нужен, чтобы заменить спор о стиле проверяемыми вопросами: когда появилась ключевая мысль, откуда взялась цифра, почему выбран этот аргумент, можно ли воспроизвести расчёт, согласуются ли версии с источниками и объяснением автора.

Такой след не устанавливает невинность автоматически. Он повышает стоимость ложного обвинения и стоимость подделки процесса. Клятва просит поверить; согласующиеся свидетельства дают материал для проверки. Из первой главы мы знаем это правило в общем виде: не заставляйте верить на слово — дайте способ убедиться. Теперь у него появилось личное измерение: в эпоху обратного теста Тьюринга след работы — ваше единственное вещественное доказательство человечности. Начатое здесь мы дорастим в конце книги до целой стратегии: согласованный след может стать и защитой, и капиталом.

И последнее — про сам экзамен. Не сдавайте его на условиях толпы. Не стирайте своё тире, не обедняйте словарь, не делайте из научного труда «работу пятиклассника», чтобы понравиться детектору, — эта гонка проиграна заранее, потому что модели учатся на человеческом письме и всегда будут писать «по-человечески» лучше, чем напуганный человек — «по-живому». Ваша человечность не в приметах сти-

ля. Она в том, что за вашим текстом есть путь: время, версии, решения, ошибки. Путь нельзя сгенерировать задним числом. Его можно только пройти — и сохранить.

Карикатура Стайнера из эпиграфа за тридцать лет не постарела — она просто сменила смысл. В интернете по-прежнему никто не знает, что ты собака. Но теперь никто не знает и что ты человек. Разница в одном: у человека есть черновики.

Практикум к книге

В Практикуме к этой главе лежат:

- **след работы «до обвинения»** — история версий, сохранённые черновики и бесплатное датирование: одна настройка, которая пригодится через год;
- **два протокола обвинения** — что делать, если обвинили вас, и что проверить, если решение выносите вы.

Открыть Практикум: <https://cheap-intelligence.vercel.app/>

Глава за минуту

1. **Тест Тьюринга развернулся.** Машина научилась казаться человеком — теперь человек доказывает, что он не машина: студент подтверждает авторство дипломной работы, писатель — романа, кандидат — отклика на вакансию.

Презумпции невиновности в этом экзамене нет.

2. Распознавание — монетка. Измеренная средняя точность людей по всем форматам — 55,5%; отдельно для текстов — 52%. Стихи ИИ читатели чаще считали человеческими, чем стихи настоящих поэтов, а GPT-4.5 с ролевой инструкцией в тесте Тьюринга выбирали человеком в 73% раундов; без неё — в 36%. «Я отличу на глаз» — не проверка реальности, а уровень уверенности.

3. Детектор — не приговор. OpenAI закрыла собственный детектор «из-за низкой точности» и прямо ответила педагогам: надёжных детекторов нет. Программы штрафуют неносителей языка и добросовестных, а суд в Татарстане показал, как вероятностный отчёт превращается в санкцию внутри сломанной процедуры. «Процент» — повод для вопросов, решает человек и след.

4. След работает лучше клятвы. Черновики, история версий, источники и датированные этапы не выносят оправдательного приговора, но дают материал для проверки. Чем согласованнее след, тем труднее заменить разбор заверением «я писал сам» или одним «процентом». А дивиденд лжеца напоминает: проверка нужна в обе стороны — и подделке, и крику «это подделка!».

Как объяснить близким.

- Студенту или школьнику: «Если обвинят, что за тебя писал ИИ, — не клянись. Покажи историю версий, источники

и черновики, объясни путь к результату. Такой след можно проверить; одного заверения недостаточно».

- Коллеге, который судит: «Процент» на экране детектора — не приговор. Даже OpenAI закрыла свой детектор за ненадёжность. Прежде чем обвинять — спроси у человека след работы.

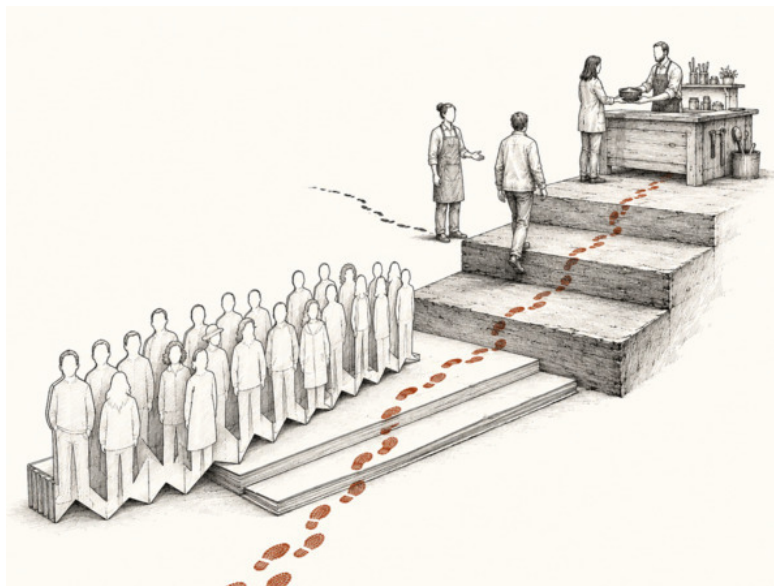
Детектор не отвечает за свой «процент». Отвечает тот, кто ему поверил.

К следующей главе

Мы разобрались, как проверяют одного человека и один текст. Но подделка давно научилась работать оптом. Сорок тёплых отзывов из первой главы — лишь витрина массового производства подделок: сегодня генерируются не только мнения, но и сами их носители — довольные клиенты, состоявшиеся эксперты, целые толпы, у которых есть лица, биографии и твёрдое мнение о вашем бизнесе.

Один такой несуществующий отзыв — мелочь. Но когда несуществующие гости приходят толпой, рейтинг настоящего ресторана падает за один вечер. В следующей главе — о том, почему разрушение репутации продаётся пакетом, а на её восстановление уходят месяцы.

Глава 4. Синтетическая репутация



Толпу можно заказать. Проверяемую репутацию приходится строить по ступеням

«Ложь летит, а правда ковыляет за ней следом;
когда люди наконец разубеждены — уже поздно: шутка
сыграна, и басня возымела действие».

— Джонатан Свифт, журнал *The Examiner*, 1710⁷⁸

⁷⁸ Джонатан Свифт, *The Examiner*, 9 ноября 1710 года: «ложь летит, а правда ковыляет следом». Популярная версия про ботинки — апокриф.

Вопрос главы: чего стоит мнение толпы — рейтинг, отзывы, подписчики, хор экспертов, — если толпу теперь можно заказать, как пиццу?

Обещание: показать, из чего на самом деле состоит репутация, почему её разрушение и восстановление стоят по разным прецедентам — и как проверять чужую толпу и строить собственную репутацию так, чтобы её нельзя было ни купить, ни обрушить за вечер.

Субботний вечер, Чикаго, район Голд-Кост. Ресторан Alpana работает штатно: бронирования, кухня, зал. А в телефоне владелицы тем временем ужинают другие гости — те, которых в зале никто не видел. За несколько часов на странице ресторана в Google появляются пятнадцать отзывов с одной звездой. Тексты усреднённо-обиженные, как под копирку: хозяйка грубила, еда остыла, столы липкие. Рейтинг падает с 4,5 до 4,1. Для ресторана среднего и высокого сегмента это, по словам самой владелицы, «смертельный удар»: десятки доли рейтинга решают, забронирует ли столик человек, выбирающий между тремя заведениями на карте.⁷⁹

Дальше схема раскрывает себя. Один из «обиженных гостей» — аккаунт с именем Peter Vicious, «Питер Злобный», — предлагает владелице написать ему в WhatsApp (принадлежит Meta Platforms Inc., признанной экстремистской орга-

⁷⁹ Block Club Chicago и Axios, 2025: ресторан Alpana получил 15 фальшивых отзывов и требование выкупа; Google удалил атаку после жалобы.

низацией; её деятельность в России запрещена), чтобы «решить вопрос с отзывом». Схема озвучивается без стеснения: платите — отзывы исчезнут, не платите — завтра придут новые.

Владелицу зовут Алпана Сингх, и она поступила ровно так, как советует эта книга, — но об этом чуть позже. Сначала оглядимся: её субботний вечер не был особенным. В ту же осень модный омакасэ-ресторан SHŌ в другом районе Чикаго собрал двадцать однозвёздочных отзывов за три-четыре ночных часа. Ресторану Beity прилетели тексты, почти дословно совпадающие с отзывами на заведение в другом конце города. Семейная аптека Danny's Pharmacy за один день упала с 4,9 до 4,1 — а вымогатель пообещал добавлять по пятнадцать отзывов ежедневно, пока не заплатят. Глава ресторанной ассоциации Иллинойса говорил о всплеске таких атак за полгода, а профильный наблюдатель Fake Review Watch к тому времени насчитал больше ста пятидесяти атакованных бизнесов по всему миру.⁸⁰ Той же осенью The New York Times описала эту волну как глобальную индустрию.⁸¹

В прошлой главе человек доказывал, что его работа — его. Здесь ставка выше на один этаж: бизнесу приходится доказывать, что его клиенты — существуют. А клиентам — что со-

⁸⁰ Block Club Chicago, 5 декабря 2025 года: рестораны и аптека получили серии фальшивых отзывов с угрозами продолжить атаку.

⁸¹ Stuart A. Thompson, NYT News Service, 18 сентября 2025 года: владелица малого бизнеса заплатила вымогателям после обвала рейтинга, но атака возобновилась.

рок восторгов на странице написали не те же люди, что и пятнадцать проклятий. Я буду называть это **синтетической репутацией**⁸²: видимость общественного одобрения или осуждения, за которой нет заявленной толпы независимых людей. Явление старше генеративного ИИ. Новое не в самом обмане, а в его себестоимости: модели удешевляют тексты, вариативность и персонализацию, хотя аккаунты, доставка и обход модерации по-прежнему требуют инфраструктуры.

Датчик, который мы не выбирали

Чтобы понять, почему пятнадцать несуществующих гостей способны ранить настоящий ресторан, придётся признать в одной слабости. У человека есть встроенный датчик правды: «так говорят многие». Психолог Роберт Чалдини полвека назад назвал его социальным доказательством⁸³: мы считаем поведение правильным, когда видим, что так ведут себя другие.⁸⁴ Датчик честно служил тысячи лет, потому что опирался на железную арифметику: один человек может

⁸² Синтетическая репутация — видимость общественного одобрения или осуждения, за которой нет заявленной толпы реальных независимых людей. Подробнее — «Словарь Тома 3».

⁸³ Социальное доказательство усиливается числом голосов только при их независимости: сорок координированных аккаунтов остаются одним источником сигнала. Подробнее — «Словарь Тома 3».

⁸⁴ Роберт Чалдини, «Психология влияния» (1984): социальное доказательство заставляет считать поведение правильным, если так делают другие.

врать, но сорок незнакомых людей не могут врать одинаково — им слишком трудно сговориться. Каждый голос толпы был слабым сигналом, зато независимым от остальных. Складываясь, слабые сигналы давали сильный.

Подделывать этот датчик пробовали задолго до нейросетей. Ещё в 1985 году американский сенатор Ллойд Бентсен, получив гору «возмущённых писем избирателей», организованных страховым лобби, усмехнулся: техасец отличит настоящую траву от астротурфа — искусственного газона. Так родился термин «астротурфинг»⁸⁵: искусственная «народная волна», изображающая стихийную.⁸⁶ Но у астротурфинга прошлого века была честная себестоимость: каждое письмо кто-то печатал, каждый фальшивый отзыв кто-то сочинял. Толпа из ста человек требовала если не ста человек, то хотя бы заметных денег и времени. Датчик ошибался, но редко — подделка толпы была кустарным промыслом.

Генерация сделала с толпой то же, что раньше сделала с текстом, голосом и лицом: уронила цену единицы почти до нуля. Сто «разных людей» с разными именами, манерами и историями болезней пишутся за вечер. Им можно сгенерировать лица и биографии. Стаж на площадке — старые публикации, связи и следы чужих ответов — тоже можно ими-

⁸⁵ От обычной кампании астротурфинг отличается сокрытием общего организатора: координацию выдают за независимую инициативу множества людей. Подробнее — «Словарь Тома 3».

⁸⁶ Merriam-Webster и профильные источники связывают «астротурфинг» с фразой сенатора Ллойда Бентсена об искусственном газоне AstroTurf (1985).

тировать, но уже не одним запросом: для этого нужны время, контроль над аккаунтами и согласованная история. Арифметика датчика умерла тихо: сорок незнакомцев по-прежнему не могут врать одинаково — но теперь никто не обещает, что их сорок. Может быть, их ноль.

Толпу больше не нужно собирать. Её можно заказать.

Ломают по прейскуранту, чинят по счётчику

Вернёмся к субботнему вечеру Алпаны Сингх — в нём спрятана вся экономика новой репутации.

Сингх вымогателю не заплатила. Вместо этого она четыре часа собирала доказательства, жаловалась платформе и рассказала историю публично. Google снял фейковые отзывы в течение суток и заблокировал аккаунты — редкая для таких атак скорость; аптеке Danny's ждать пришлось пару дней, ресторану SHŌ — почти неделю. Казалось бы, победа: подделку стёрли, справедливость восстановлена.

А теперь арифметика после победы. Пока шла атака, постоянные гости Alpana, узнав о беде, написали 65 настоящих пятизвёздочных отзывов. Рейтинг не сдвинулся. Сингх посчитала: чтобы вернуть утраченные десятые доли, ресторану нужно порядка ста шестидесяти безупречных оценок. Пятнадцать подделок обрушили то, что теперь должны чинить сто шестьдесят настоящих людей. Разрушение репутации

ции продаётся по прейскуранту — доллары за пакет отзывов. Восстановление устроено иначе: оно требует месяцев хорошей работы и десятков реальных клиентских оценок.

Эта асимметрия не сбой конкретного алгоритма — это свойство самого доверия: копить его приходится медленно, потому что каждый настоящий сигнал стоит кому-то усилий, а терять можно быстро, потому что подозрение усилий не требует. Хозяйка строительной компании из Лос-Анджелеса Наталия Пайпер сформулировала это без всякой теории. Летом ей написали с пакистанского номера: «Мне заказали негативные отзывы на ваш бизнес. Заказ — двадцать штук». Её безупречные 5,0 на Google рухнули до 3,6 — вместе с потоком клиентов. Пайпер сделала то, что делать нельзя, и её можно понять: заплатила. Сто пятьдесят долларов одному «подрядчику», потом сто другому. Платёж не гарантирует остановки. Он может показать злоумышленнику, что перед ним жертва, готовая платить, и тем самым повысить вероятность повторного требования. В случае Пайпер именно это и произошло: через несколько недель после двух платежей на странице появились десять свежих однозвёздочных отзывов. Пайпер перестала платить, дождалась удаления через жалобы и подвела итог одной фразой, которую стоило бы печатать на каждом рейтинге: «Я восемь лет зарабатывала репутацию на этом рынке — а один человек может разрушить её за день».

Есть в этой истории и деталь, от которой не знаешь, сме-

яться или нет. В первую большую волну отзывного рэкета, ещё в 2022-м, вымогатели просили у американских ресторанов всего 75 долларов подарочной картой — и писали жертвам вежливые письма с извинениями: простите, нам нужно кормить семью, вашего платежа хватит на три недели.⁸⁷ Было ли это признанием или частью сценария давления, неизвестно. Но экономика видна и без биографии отправителя: небольшой платёж с одной жертвы может быть рациональным только при массовом потоке атак. И мы видим, что по ту сторону атаки сидит не голливудский злодей, а исполнитель на аутсорсе, для которого чужая репутация — сдельная подработка. Это, пожалуй, самая точная иллюстрация новой цены толпы: разрушить дело, которое строили восемь лет, стоит дешевле, чем месячная аренда ресторана.

У толпы есть оптовый склад

Если единичная атака стоит семьдесят пять долларов, где-то должен существовать и опт. Он существует, и масштабы у него промышленные.

Летом 2022 года Amazon подала в суд на администраторов более чем десяти тысяч групп в соцсетях, где торговали отзывами: продавец платит деньгами или товаром — участник группы публикует «честный восторг». Одна такая груп-

⁸⁷ Today.com, август 2022 года: ресторанам в США угрожали сериями однозвёздочных отзывов, требуя подарочную карту на \$75.

па, скромно называвшаяся «Amazon Product Review», насчитывала 43 тысячи участников; чтобы прятаться от модераторов, торговцы намеренно искажали слова звёздочками: «отзв» и «возврт». Через год компания отчиталась ещё о полутора сотнях исков к торговцам отзывами в США, Китае и Европе.⁸⁸ Tripadvisor в отчёте о прозрачности признал фейковыми 4,4 процента всех присланных за год отзывов — больше миллиона трёхсот тысяч штук, — а среди источников назвал профессиональные «конторы платных отзывов», половина которых работала из шести стран.⁸⁹ Google за один только 2023 год удалил больше 170 миллионов поддельных отзывов и 12 миллионов фальшивых профилей компаний.⁹⁰ Последняя цифра заслуживает отдельной паузы: двенадцать миллионов несуществующих бизнесов — с адресами, фотографиями и графиком работы — целая теневая экономика, нарисованная поверх карты.

Россия в этом рынке не наблюдатель: судя по открытым прайсам отечественных бирж накруток, один заказной отзыв стоит от полутора сотен рублей, а при оптовом заказе — дешевле чашки кофе.⁹¹

⁸⁸ Amazon, 2022—2023: компания подала иски против администраторов групп и торговцев фальшивыми отзывами; цифры отражают заявления платформы.

⁸⁹ Tripadvisor, отчёт 2023 года: платформа признала фейковыми 4,4% полученных отзывов; это данные собственной модерации, не независимый аудит.

⁹⁰ Google, данные за 2023 год: удалено свыше 170 млн нарушающих правила отзывов и 12 млн фальшивых бизнес-профилей; это отчёт самой платформы.

⁹¹ vc.ru, 20 мая 2025 года: открытые предложения показывали цены от сотен

Государства начали отвечать. С октября 2024 года в США действует правило Федеральной торговой комиссии, прямо запрещающее фейковые отзывы — включая сгенерированные ИИ, — покупку отзывов и отзывы сотрудников без раскрытия; за осознанное нарушение грозит штраф до 51 744 долларов за каждый случай.⁹² Подделка сигналов впервые получила официальный ценник наказания. Это хорошая новость. Плохая состоит в том, что закон и платформенный антифрод разбираются с подделкой постфактум — а решение о том, куда пойти ужинать и с кем подписать договор, вы принимаете сегодня вечером, глядя на рейтинг, который прямо сейчас может не отражать реальность.

Толпа научилась разговаривать

Пока мы обсуждали отзывы, у синтетической толпы появилась новая возможность. Отзыв — толпа молчащая: написала и замолчала. Новая толпа умеет отвечать на вопросы.

до десятков рублей за отзыв и скидки за объём; это рекламный снимок прайсов, не исследование рынка.

⁹² FTC, правило 2024 года: в США запрещены фальшивые отзывы, их покупка и сокрытие связи с инсайдером; предусмотрены гражданские штрафы.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, [купив полную легальную версию](#) на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.