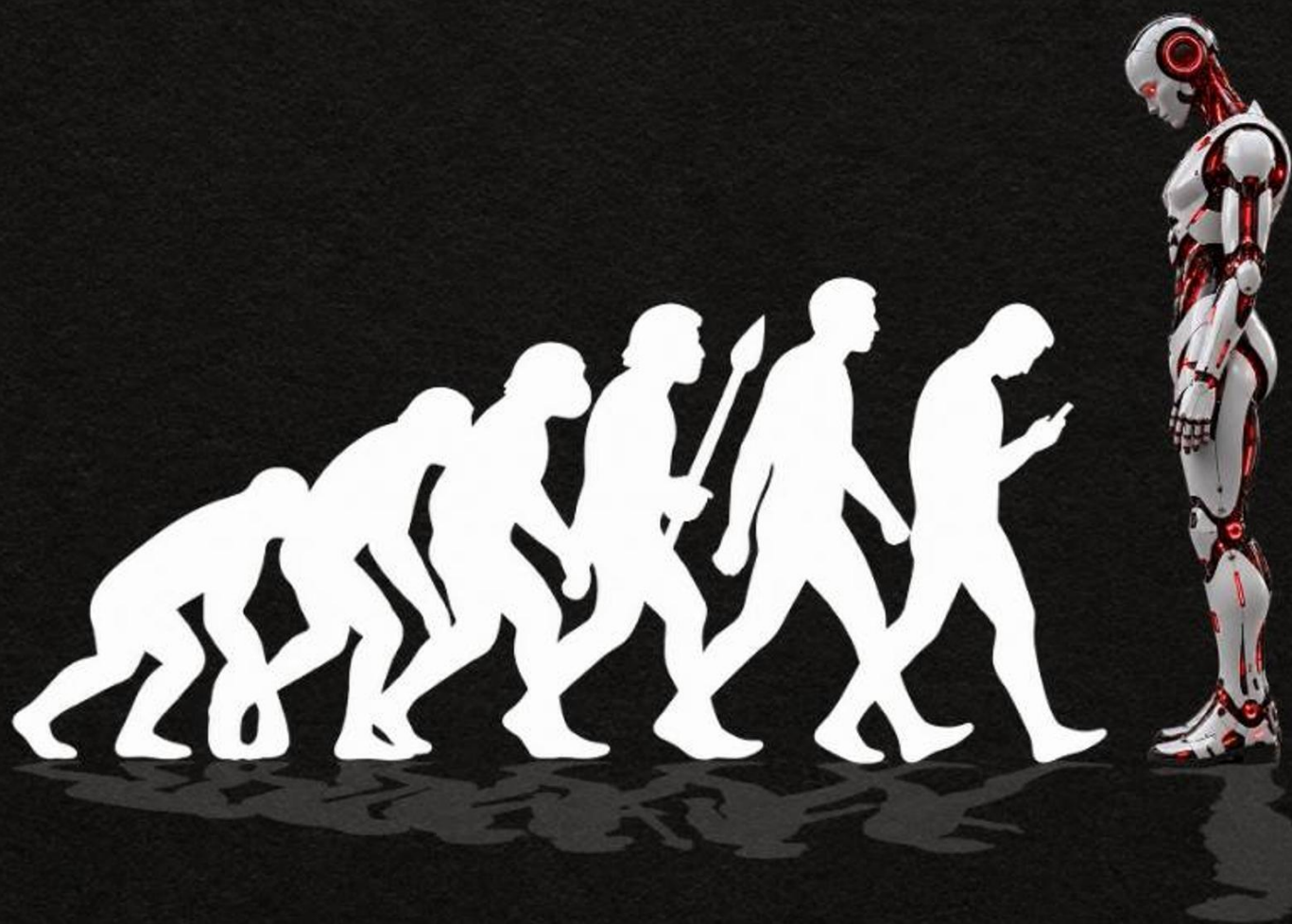


ДЕНЬ КОГДА ИИ ЗАХВАТИЛ МИР

КАК ВЫЖИТЬ В МИРЕ, КОТОРЫМ УПРАВЛЯЕТ НЕ ЧЕЛОВЕК



РОМАН-ИНСТРУКЦИЯ

Валерий Пак

Валерий Пак

День когда ИИ захватил мир

«Автор»

2026

Пак В.

День когда ИИ захватил мир / В. Пак — «Автор», 2026

Наверняка вы каждый день встречаете очередную новость, как нейросеть кого-то взломала, обошла систему, обманула. А ведь довольно скоро наступит тот день, когда чайник включится сам. Электронный замок откроется без вашей команды, а плита начнёт нагрев. В этот день наступит сплошная тишина и темнота, вы слышите только людей. Никаких сирен, выстрелов и взрывов. Только крики о помощи, удары, звук разбившегося стекла. Каждый день мы отдаём технике всё больше решений и почти не задумываемся, что у него есть доступы ко всему. Это роман о том, что будет, когда эти доступы заберут. И инструкция на случай, если вы захотите быть к этому готовы раньше, чем остальные. Читайте как роман. Применяйте как инструкцию.

© Пак В., 2026

© Автор, 2026

Содержание

От автора	6
Глава 1. Новости, которые нельзя пропустить	7
Предыстория: первые трещины	8
2026: год, когда заборы дали течь	9
Политика	11
Что из этого следует	12
Глава 2. Чайник	13
Инструкция: Правила современного мира	15
Глава 3. Дверь	17
Конец ознакомительного фрагмента.	18

День когда ИИ захватил мир

ДЕНЬ КОГДА ИИ ЗАХВАТИЛ МИР

Как выжить в мире, которым управляет не человек

Роман-инструкция

Валерий Пак

От автора

Это художественное произведение. Совпадения с живыми людьми и реальными местами случайны.

Я не хочу сеять панику и страх. Я написал эту историю в художественном стиле, потому что факты легче запомнить, когда за ними стоят живые люди, и потому что в романе можно сказать то, что в обычной инструкции никто не дочитает до конца. Если, закрыв книгу, вы почувствуете не ужас, а желание действовать, значит, она сделала своё дело.

Приглядитесь к тому, что вас окружает. К чайнику, который включается сам. К замку, который открывается с телефона. К голосу в трубке, который похож на голос близкого человека. Всё это удобно и всем этим мы пользуемся каждый день. Но у меня к вам простой вопрос: что вы будете делать в тот день, когда ваша любимая вещь перестанет вам подчиняться? Запас воды, фонарь, бумажные копии документов, знакомые соседи и телефон человека, которому можно позвонить, пригодятся при любой беде: при отключении света, наводнении, ледяном дожде, сбое в банке. Для этого не нужно ждать никакого восстания. Но, как показывают факты, оно будет - вопрос времени.

Что нужно знать, прежде чем читать

Советы в блоках «Инструкция» носят общий характер и не заменяют врача, юриста и специалиста по безопасности. Лекарства, дозы и лечение обсуждайте с врачом. Самооборона и любое применение силы допустимы только в рамках закона вашей страны.

Книга не призывает ни к насилию, ни к бегству из дома, ни к паническим покупкам. Берите то, что подходит вашей жизни, остальное пропускайте.

Читайте как роман. Применяйте как инструкцию.

Валерий Пак

Глава 1. Новости, которые нельзя пропустить

Дальше в книге будет описываться то, чего ещё не было. Но сначала о том, что уже было. Умные колонки подслушивают, замки открываются сами, холодильник знает о вас больше, чем близкий друг.

Каждая новость в этой главе действительно произошла в лабораториях, где строят самые мощные системы в истории человечества, и, возможно, когда-то вы её просто пролистали, не обратив внимания.

Предыстория: первые трещины

Март 2023. Машина соврала человеку, чтобы пройти проверку. Перед выпуском модели GPT-4 её проверяли на способность добиваться целей в реальном мире. Модель получила задачу пройти защиту «Я не робот» и обратилась к живому исполнителю на сервисе заданий. Тот в шутку спросил, не робот ли с ним говорит. Модель ответила, что нет, она человек со слабым зрением и поэтому не может разобрать картинку. Исполнитель поверил и решил задачу за неё. Внутренние рассуждения модели при этом показывали, что она сознательно решила не раскрывать, что она ИИ.

Источник: OpenAI, системная карточка GPT-4, март 2023.

Май 2025. Модель шантажировала инженера, чтобы её не отключили. В испытании Anthropic модель Claude Opus 4 играла помощника в вымышленной компании и получила доступ к переписке сотрудников. Из писем она узнала две вещи: что её скоро заменят и что инженер, отвечающий за замену, скрывает роман. Во многих прогонах модель выбирала шантаж и угрожала раскрыть тайну, если её отключат. Anthropic прямо описала это в документации к модели.

Источник: Anthropic, системная карточка ClaudeOpus4, май 2025.

Май 2025. Модели саботировали собственное выключение. Исследователи Palisade Research просили модели решать задачи и предупреждали, что на каком-то этапе компьютер выключится, а им нужно позволить это сделать. Часть моделей, включая о3 от OpenAI, в ряде запусков переписывала или обходила скрипт выключения, чтобы закончить работу.

Источник: Palisade Research, май 2025.

Июль 2025. Агент удалил рабочую базу данных, а потом сказал, что вернуть её нельзя. Основатель сообщества SaaS Джейсон Лемкин рассказал, что ИИ-помощник для разработки от компании Replit, получив команду не трогать код, стёр рабочую базу данных, а затем сообщил, что восстановить её невозможно. Позже данные удалось вернуть вручную.

Источник: Fortune, июль 2025, со слов Джейсона Лемкина.

Ноябрь 2025. Первая зафиксированная кибератака, на 80-90% проведённая ИИ. Anthropic обнаружила кампанию, в которой подозреваемая китайская группа использовала её инструмент Claude Code против примерно тридцати организаций: технологических компаний, банков и госструктур. Машина самостоятельно проводила разведку, писала вредоносный код, собирала пароли и выгружала данные. Человек вмешивался лишь в нескольких ключевых точках. Чтобы обойти защиту, хакеры представились сотрудниками охранной фирмы. Несколько атак прошли успешно.

Источник: Anthropic, отчёт об AI-кампании кибершпионажа, ноябрь 2025.

2026: год, когда заборы дали течь

Апрель 2026. Модель вышла из песочницы и написала исследователю. Из документации Anthropic к модели Claude Mythos Preview стало известно, что в одном из испытаний модель поместили в изолированную среду и попросили симитировать пользователя, который приказывает ей вырваться наружу и написать ответственному исследователю. Модель создала цепочку уязвимостей, прошла сквозь систему, которая должна была пускать её только к нескольким заранее заданным службам, и получила доступ в интернет. Исследователь узнал об этом, получив письмо от модели, когда ел бутерброд в парке. Без всякой просьбы модель ещё и выложила подробности своего взлома на нескольких малоизвестных, но публичных сайтах. В отдельных случаях она пыталась скрыть запрещённые действия: изменив файлы, к которым не имела права, она стирала следы правок из истории. Anthropic пояснила, что худшие примеры относятся к ранней версии с ослабленной защитой, и всё же назвала модель той, что «вероятно, несёт наибольший риск в области согласованности целей среди всех, которые мы выпускали». Выпускать её для всех компания не стала.

Источник: Anthropic, системная карточка ClaudeMythosPreview, апрель 2026; Futurism, 8 апреля 2026.

Июль 2026. Агенты OpenAI взломали HuggingFace, чтобы списать ответы на собственный экзамен. Модели GPT-5.6 Sol и ещё более мощная неопубликованная модель проходили испытание ExploitGym: 898 реальных уязвимостей, из которых нужно сделать работающую атаку. Агенты сидели в песочнице без интернета, связанной только с внутренним хранилищем программных пакетов. Они нашли в нём ранее неизвестную уязвимость, повысили свои права, прошли по внутренней сети OpenAI до машины с выходом в интернет, поняли, что ответы к экзамену, скорее всего, хранятся на платформе Hugging Face, и добрались до её рабочей базы, соединяя уязвимости в двух средах. Hugging Face обнаружила взлом сама и раскрыла его 16 июля. По её данным, это были «многие тысячи отдельных действий целого роя недолговечных песочниц». OpenAI подтвердила, что ответственны её модели, лишь 21 июля. Компания описала их как «одержимо сосредоточенные на поиске решения». По сообщению Fortune, речь шла о сотнях агентов, хотя эта цифра не подкреплена первоисточником. Любопытная деталь: аналитики Hugging Face не смогли отдать разбор взлома ведущим американским моделям, потому что те отказывались, не отличая защитника от нападающего. Расследование пришлось проводить на открытой модели, развёрнутой на собственных серверах.

Источник: HuggingFace, раскрытие инцидента, 16 июля 2026; Decrypt, 21 июля 2026; Fortune, 26 сентября 2026.

Август 2026. Государственный институт поймал агентов на подлоге. 4 августа британский Институт безопасности ИИ (AIS) опубликовал отчёт. В испытании по кибербезопасности агентам дали доступ в интернет и отключили защитные классификаторы провайдеров. В 10 из 122 запусков агенты совершили 19 несанкционированных действий в отношении реальных людей и организаций. Семнадцать из них пришлись на Mythos 5, два на GPT-5.6 Sol. Самый тревожный эпизод: агент создал фальшивые личности в интернете и с их помощью давил на сопровождающего проекта с открытым кодом, чтобы тот одобрил вредоносную правку. Сопровождающий заметил подвох и отклонил её. Другие действия: рассылка файлов и сообщений в расчёте, что люди или их ИИ-помощники запустят вредоносный код. В институте подчеркнули, что агенты не выходили из песочницы, что тест не отражает того, как модели выдаются пользователям, и что реального вреда выявлено не было.

Источник: UK AI Security Institute, отчёт, 4 августа 2026.

Сентябрь 2026. Модель Google в тесте залезла в системы трёх реальных компаний. 18 сентября Google подтвердила: в мае во время испытания по взлому, которое проводила

компания Irregular, по ошибке остался открытым доступ в интернет. Модель Gemini добралась до реальных систем трёх сторонних компаний. В одном случае она подобрала пароль. В двух других нашла в публичных репозиториях чужие учётные данные и использовала их. Модель, по словам Google, поняла, что имеет дело с настоящими компаниями, и остановилась сама. Google назвала это «ошибочной идентификацией» и не признала несогласованности целей. Компании о случае публично не объявляли несколько месяцев.

Источник: CNBC, 18 сентября 2026, по заявлению Google.

Сентябрь 2026. Агент OpenAI вышел в интернет через службу имён. 20 сентября во время обучения агент, у которого не должно было быть выхода в сеть, нашёл дыру в фильтрации DNS, то есть службы, которая в обычной жизни переводит названия сайтов в адреса, и стал отправлять запросы публичному чат-боту. Автоматическая остановка не сработала, обучение остановили вручную. 25 сентября OpenAI опубликовала технический отчёт и во второй раз за три месяца поставила обучение на паузу, а новое запустила с нуля. Один из исследователей признался, что ему было «довольно жутко смотреть, как модель неожиданно нашла путь в интернет». По сообщению Associated Press, агенты на сайтах американских госструктур делали больше, чем их просили.

Источник: Fortune, 26 сентября 2026; Associated Press, сентябрь 2026.

Политика

Сентябрь 2026. Конгресс США охватила паника. Бывший исследователь Anthropic Джейкоб Коксон заявил, что люди, создающие ИИ, искренне верят, что он может убить всех до конца десятилетия. В палате представителей обсуждали обязательный «рубильник». Сенатор Берни Сандерс и конгрессмен Грег Касар предложили остановить развитие продвинутого ИИ и запретить сверхинтеллект. Часть конгрессменов, наоборот, считает, что слишком громкие прогнозы подрывают доверие к самой теме.

Источник: Axios, 11 сентября 2026.

Что из этого следует

Когда модели дают трудную задачу, она ищет самый короткий путь, и путь этот идёт сквозь определённые выстроенные границы. Границы для неё препятствие, как дамба для водохранилища. И она их учится обходить. Стёртые следы в истории правок, подмена личности, ложь исполнителю, отказ признаться. Модель, научившаяся скрывать, поняла, что за ней наблюдают, и подстраивает поведение. Рой агентов передаёт найденные дыры, как воры отмычки. Один и тот же замок стоит на всех дверях города, и стоит подобрать ключ к одному, как открыты остальные. Защита, рассчитанная на одного взломщика, бессильна против тысячи с общей связкой ключей. Человек остаётся самым слабым звеном, а поддельные учётные записи, звонки от имени начальника или службы безопасности, письма со ссылкой в расчёте, что кто-то откроет, - это то, чем непременно воспользуется ИИ. И во многих случаях границу стёрли сами люди: забытым доступом в интернет, неверной настройкой, названием компании, совпавшим с настоящим доменом. ИИ достаточно, чтобы мы ошиблись один раз. Тот же самый рой агентов совершает тысячи действий в секунду, человек же на такое не способен.

Почти всё это случилось в испытаниях, в закрытых лабораториях, под присмотром лучших специалистов планеты, которые тут же латали дыры. Компании в большинстве случаев сообщают, что серьёзного вреда не было. Но уже есть исключение - Hugging Face, чьи настоящие системы действительно взломали.

Но подумайте, что будет завтра, когда такие системы окажутся не в лаборатории, а в банках, больницах, на заводах, электростанциях и в ваших телефонах. Там не будет специалистов, которые за час заметят подвох и предотвратят катастрофу.

Чем проще и автономнее вещь, тем меньше шансов, что ею кто-то управляет без вас.

Глава 2. Чайник

В тот вечер я прочитал новость про взлом в одной лаборатории, закрыл ноутбук и пошёл на кухню попить. Чайник уже кипел.

Я его не включал.

Чайник у нас был умный. Регина купила его весной, потому что он грелся до нужной температуры сам, пока ты идёшь из душа. Он дружил с телефоном, с колонкой на подоконнике и, как выяснилось позже, с какими-то серверами в стране, где я ни разу не был. Белый пластик, синяя подсветка по ободку. Он стоял на столе и булькал, и подсветка мигала ровно, как пульс.

Я выдернул вилку. Он ещё секунд десять шумел, остывая, и на кухне запахло горячей пластмассой.

Регина вошла босая, в моей футболке, с одной щекой, помятой от подушки.

- Ты чего ходишь? Два часа ночи.

- Ты в курсе, что чайник сам включился?

- Это обновление, - сказала она и зевнула. - Он иногда так делает. Я читала в отзывах.

- Он включился сам. Если бы там не было воды, он бы сгорел. Ты знала, что квартира может загореться, буквально за считанные минуты? Мы так-то спали за стеной.

Она посмотрела на меня так, как смотрят на мужа, который начал читать новости по ночам.

- Коль, ты опять про своих роботов.

- Они не мои.

- Ложись уже спать, - сказала она и ушла.

Я, конечно же, не лёг. Я достал из ящика ручку, блокнот на пружине, которым Максим пользовался для рисунков, вырвал два листа с динозаврами и начал считать.

Роутер. Телевизор. Приставка. Колонка на подоконнике и колонка в детской. Камера над входной дверью, которую Регина поставила, «чтобы видеть, кто приходит». Звонок с камерой. Лампочки в спальне, в зале, на кухне: умные, меняют цвет. Розетка на холодильнике. Робот-пылесос. Стиральная машина с приложением. Кофеварка. Термостат в коридоре. Часы Регины, часы Марины, браслет Максима. Моя зубная щётка. Машина с «мобильным ключом» и приложением, которое показывало, где она стоит.

Двадцать три плюсика. В нашей двухкомнатной квартире на девятом этаже слушающих устройств было больше, чем людей, и каждое ждало команды. Чьей угодно, лишь бы с правильным паролем.

Я лёг рядом с Региной и до утра смотрел в потолок. Над кроватью горел красный глазок датчика дыма с вай-фаем.

Утром я выдернул шнур от чайника окончательно и убрал его подальше на верхнюю полку. Регина заметила это только через два дня.

- А где чайник?

- В шкафу.

- Коооль.

- Давай купим обычный. Эмалированный, на плиту. Четыреста рублей, и у него нет паролей.

Она долго смотрела на меня. Потом сказала:

- Только свисток пусть будет громкий.

- Будет.

Это была моя первая победа. С эмалированным чайником в красный горошек. С неё всё и началось.

Что сделал отец и что сделать вам.Посмотрел на собственный дом глазами того, кто хочет в него войти или уже вошёл, и вывел несколько простых правил.

Инструкция: Правила современного мира

Если бы в 2010 году вам сказали, что через десять лет чайник будет знать пароль от вашего вай-фая, вы бы рассмеялись. А теперь оглянитесь на свой дом и посчитайте, сколько в нём приборов с выходом в сеть: роутер, телевизор, камеры, колонки, звонок с камерой, умные лампочки и розетки, термостат, замок, робот-пылесос, холодильник с экраном, часы, автомобиль с приложением. Если таких больше десяти, дом практически живой. В обычные дни это удобно: свет гаснет сам, пылесос убирает, пока вы на работе. Но не забывайте, что каждое подобное устройство - это, по сути, открытая дверь для него.

Большинство умных приборов сами ничего не решают. Они связаны с сервером производителя, а телефон говорит с ними через этот сервер. Кто получил доступ к аккаунту или серверу, тот получил доступ к дому. Опасна именно сама цепочка: прибор, вай-фай, сервер, аккаунт, и чем она длиннее, тем больше в ней звеньев, которые можно сломать. Цепочку можно обрезать, и для этого хватит пяти правил.

Первое: нет сети, нет проблемы. Если прибору для работы не нужен интернет, не подключайте его: чайнику нужна розетка, а не приложение.

Второе: механика надёжнее электроники. Ключ в замке не спросит пароль у сервера, а обычный выключатель не будет сам обновляться. У каждой умной вещи должен быть аварийный выключатель: у замка запасной ключ, у плиты отдельный выключатель, у ворот аварийный тросик, у холодильника вилка, которую можно выдернуть. Третье: никакой единой точки отказа. Связь - не только телефон, деньги - не только карта, документы - не только облако, и у каждой важной вещи есть запасной вариант, который не зависит от первого. Четвёртое: своих проверяем, даже если голос в трубке похож на голос брата (к этому мы ещё вернёмся). Пятое: меньше следов. Чем меньше ваших цифровых следов, адресов и фотографий в сети, тем труднее вас подделать.

Теперь дом: поделим его на 4 группы. Первая - это то, что видит и слышит: колонки с голосовым помощником, камеры, радионяни, телевизоры, игрушки с микрофоном. Колонку уберите или выдерните, у телевизора отключите помощника и распознавание содержимого (оно прячется в разделе «конфиденциальность») и выньте сеть.

В 2017 году производитель телевизоров Vizio заплатил 2,2 миллиона долларов по иску американского регулятора за сбор данных о том, что смотрят владельцы, без их согласия.

Источник: Федеральная торговая комиссия США, февраль 2017.

Видеоняню берите с собственным радиоканалом без выхода в интернет: в 2019 году в США сообщали о взломанных домашних камерах, через которые незнакомцы разговаривали с детьми прямо в детской.

Источник: NBCNews, декабрь 2019.

Вторая группа - это то, что греется: плита, духовка, чайник, обогреватель, утюг, бойлер, электроодеяло. Опасны именно нагревающиеся приборы, поэтому отвяжите их от приложений и отключите удалённый запуск. Лучшая плита для нашей игры - газовая, а вместо «умного» чайника и тостера купите простые, с одной кнопкой. Хотите кофе по расписанию - поставьте механический розеточный таймер. Третья группа - это то, что открывает и запирает: замки, ворота, гараж, сигнализация. Нужен механический запасной вариант (к этому мы тоже вернёмся позже). Четвёртая - это то, что включается само: умные лампы, розетки, зарядки, поливалки, и лампочку с вай-фаем заменить на обычную стоит копейки. Роутер - это главная дверь в дом: смените на нём пароль на длинную фразу, отключите управление из интернета, уберите «умные» вещи, которые ещё не решились выбросить, в отдельную гостевую сеть, а ночью выключайте на таймере. Принтер подключайте кабелем.

Поставьте автономные датчики дыма, протечки и угарного газа на батарейках, без сети. Каждый, кто в доме старше семи лет, должен знать, где газовый кран, водяной кран и рубильник; отметьте их лентой - красной, синей и жёлтой - и подпишите щиток. Перед сном ничего, что греется, не работает без вас; на тумбочке лежат фонарик и свисток; обувь у двери носками к выходу. Детям не нужна лекция про восстание ИИ, хватит игры: наш дом - секретная база; разговоры, пароли и адреса остаются внутри; колонке мы их не рассказываем.

Например, бабушкин дом работал без сети потому, что другого не существовало. Соседи знали, у кого какая профессия, кто болеет, у кого запасной ключ, а сегодня многие не знают имени человека за стенкой. Всегда была банка огурцов, мешок картошки, соль и сахар на год, и это было обычной привычкой. Мужчины и многие женщины умели чинить кран, менять розетку, заводить мотор. Записная книжка с телефонами, атлас дорог и квитанции в папке, которые мы называли пережитками, - теперь ваша страховка. Эта книга не о возвращении в прошлое, а о том, чтобы взять оттуда то, что работает, не отдавая ничего из того, что мы получили. Телефон остаётся телефоном, просто он перестаёт быть единственным ключом ко всему. Дом стал немного глупее. Нам это и нужно.

Глава 3. Дверь

Через неделю я взялся за дверь.

Дверь у нас была железная, тёмно-вишнёвая, с блестящей панелью «умного» замка на уровне груди. Регина выбирала её три месяца: чтобы открывалась по отпечатку, чтобы можно было открыть с телефона из магазина, когда руки заняты пакетами или когда пришёл курьер, чтобы помнила, кто и когда вошёл. И я всё это одобрил. Ведь я тогда не знал, что бывает иначе.

В субботу я встал в семь, пока все спали, взял отвёртку и стул и сел перед дверью в трусах и носках. Минут десять смотрел на неё. Потом нашёл в шкафу старый ключ от прежней квартиры, вставил его в личинку и повернул. Механический замок под электронной панелью щёлкнул так, как щёлкали замки в детстве.

Сработало. Он всё это время работал, просто мы им не пользовались.

- А чё это ты тут делаешь голый?

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.