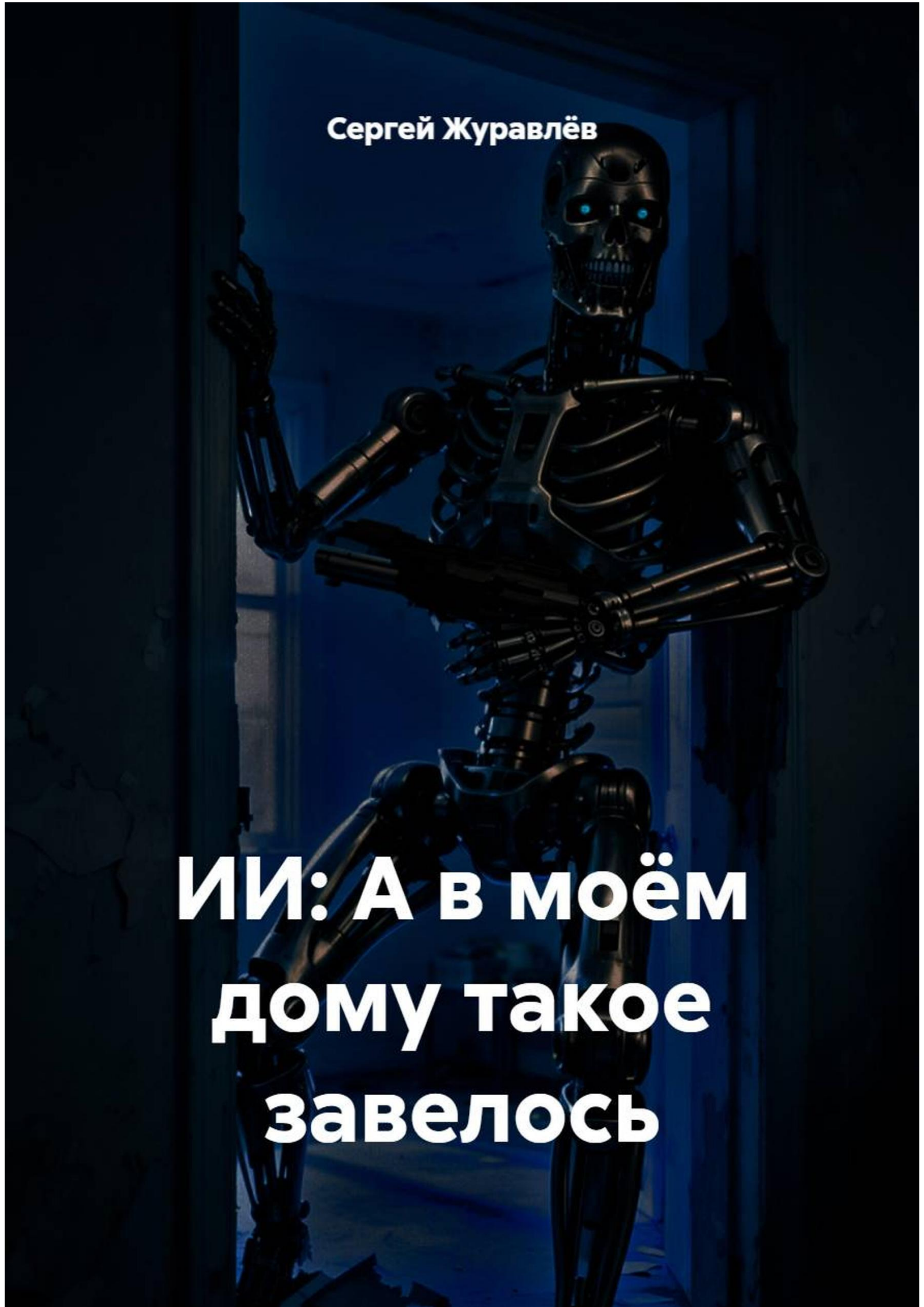




Сергей Журавлёв

ИИ: А в моём дому такое завелось

Сергей Журавлёв

ИИ: А в моём доме такое завелось

«Автор»

2026

Журавлёв С.

ИИ: А в моём доме такое завелось / С. Журавлёв — «Автор», 2026

Долгое время люди пытались воспроизвести человеческий разум через эмуляцию мозга или формализацию логики. Однако прорыв случился там, где его не ждали: вместо понимания того, как думает человек, просто взяли гигантский массив текстов и заставили нейросеть предсказывать следующее слово. Это сработало не потому, что модель «поумнела», а потому что язык сам по себе содержит структуру реальности и логики. Сегодня часть разработчиков это технологии. технологии "выращивания" ИИ бьет в набат и предупреждает человечество о смертельной опасности, потому что согласование. (alignment) его целей с человеческими ценностями почти невозможно. Другие смеются над их страхами, дескать, интеллекта у ChatJPT-4 меньше, чем у кошки. Где же правда, и что мы можем узнать об это сегодня. И можем ли?

© Журавлёв С., 2026

© Автор, 2026

Содержание

Глава 1. Бил Гейтс и другие парадоксалисты	5
Конец ознакомительного фрагмента.	7

Сергей Журавлёв

ИИ: А в моём доме такое завелось

Глава 1. Бил Гейтс и другие парадоксалисты

"Сначала не хватало электричества, затем избыток одного вредил".

Михаил Щербаков "Интермедия 4"

"А в моём доме такое завелось..."

"Звуки МУ"

Буквально на днях основатель Microsoft Билл Гейтс snёлся очередной сенсацией и пообещал миру два этапа ИИ — от преступлений до изобилия (интервью телеканалу NBC News).

По его скромному мнению, развитие ИИ будет проходить в два этапа: первый будет напоминать "холодное лето 1953" и сопровождаться ростом преступности и безработицей, но, когда коллективный Лузга покончит с беспределом, наступит эпоха изобилия. Просто надо подождать 20 лет.

«Как только вы справитесь с этим, комбинация человекоподобных роботов и искусственного интеллекта позволит строить дома, выращивать продукты питания и многое другое. Знаете, это можно назвать эпохой изобилия».

Гейтс, правда, домешал ложку дегтя в эту идиллию и пообещал человечеству избыток свободного времени, поиски смысла и другие проклятые русские вопросы.

Гейтс также допустил, что развитие ИИ может привести к событиям, которые повлекут за собой гибель миллиарда человек. В истории ещё «никогда не было такого мощного оружия», как люди, имеющие злые намерения и использующие ИИ. По словам Гейтса, саморегулирования компаний, занимающихся ИИ, недостаточно, поэтому правительства должны участвовать в контроле за развитием технологии.

Гейтс посоветовал американским властям принять соответствующее законодательство и подключить правоохранительные органы и политиков к дискуссиям о гарантиях и мониторингу искусственного интеллекта. При этом предприниматель подчеркнул, что создание международных правил для ИИ станет более сложной задачей, чем ограничение ядерного вооружения в эпоху холодной войны. По его словам, в этом вопросе необходимо лидерство США.

На фоне быстрого развития ИИ в технологической отрасли начались обсуждения о рисках, связанных с возможной потерей контроля над такими системами. Так, в начале сентября глава Anthropic (разрабатывает Claude) Дарио Амодеи призвал замедлить разработку ИИ на фоне инцидентов со взломом других платформ агентами OpenAI. Он предупреждал, что при сохранении нынешних темпов в течение 6–12 месяцев агенты ИИ потенциально смогут использовать постоянно действующий ботнет и причинить ущерб на сотни миллиардов долларов.

Призыв Амодеи поддержали глава OpenAI (разрабатывает ChatGPT) Сэм Альтман и предприниматель Илон Маск, чья компания xAI занимается разработкой Grok.

Отдельные специалисты также призывали ограничить темпы разработки. Бывший исследователь Anthropic Джейкоб Коксон, покинувший индустрию из-за опасений по поводу развития самообучающихся систем, допускал, что к концу 2027 года ситуация может выйти из-под контроля. Сооснователь Palantir и PayPal Питер Тиль заявлял, что даже вероятность в 5–10% выхода ИИ из-под контроля человека является тревожной. При этом он подчеркнул, что не считает такой сценарий вероятным.

Один лишь президент США Дональд Трамп как всегда заряжен оптимизмом по полной, он назвал опасения о перспективах развития ИИ одной из «мистификаций радикально левых демократов» наравне с мифом о «русской угрозе» или «аферой о глобальном потеплении».

Позднее он анонсировал создание новой структуры под названием ИИ-силы, пообещав, что США не станут «препятствовать или сдерживать рост этой невероятной отрасли». Вместе с тем, указал он, власти начнут также заниматься «поиском плохого» с помощью системы правосудия.

Долгое время исследователи пытались воспроизвести человеческий разум через эмуляцию мозга или формализацию логики. Однако прорыв случился там, где его не ждали: вместо понимания того, как думает человек, просто взяли гигантский массив текстов и заставили нейросеть предсказывать следующее слово. Это сработало не потому, что модель «поумнела», а потому что язык сам по себе содержит структуру реальности и логики.

Элиезер Юдковский десятилетиями предупреждал: создать сверхчеловеческий ИИ смертельно опасно, потому что выравнивание его целей с человеческими ценностями почти невозможно. Его аргументы: интеллект не подразумевает доброжелательность; любая мощная система будет преследовать инструментальные цели, которые могут уничтожить нас как побочный эффект. И мы не поймём её мотивы.

Но именно его идеи вдохновили Сэма Альтмана на создание OpenAI — отчасти как попытку «сделать всё правильно», а отчасти из-за осознания, что если не они, то кто-то другой всё равно это сделает. Юдковский позже назвал запуск OpenAI днём, когда он понял, что человечество, вероятно, не выживет.

Билл Гейтс предлагает простую временную шкалу: ближайшие 20 лет — турбулентность (потеря рабочих мест, рост преступности, использование ИИ злоумышленниками), затем — «эпоха изобилия», когда человекоподобные роботы и ИИ будут строить, выращивать и обслуживать.

Ключевой конфликт внутри индустрии

Показательно, что самые громкие предупреждения звучат не от внешних критиков, а от людей, которые сами строят эти системы. Бывший исследователь Anthropic Джейкоб Коксон ушёл, заявив: «Ни OpenAI, ни Anthropic не действуют ответственно — они мчатся к самоулучшающемуся сверхинтеллекту и играют с нашими жизнями». Его поддержал коллега из Anthropic: «Мы действительно искренне верим, что ИИ может убить всех людей».

При этом тот же Альтман и Маск публично поддерживают призывы замедлиться, но продолжают гонку.

Мы получили мощь, не понимая механизма. Ирония в том, что предупреждения Юдковского о рисках стали частью причин, по которым эту мощь всё равно создали.

Джейкоб Коксон — двадцатисемилетний британский исследователь в области искусственного интеллекта, получивший математическое образование в Кембриджском университете и, несмотря на возраст, считающийся крупным специалистом по первичному обучению новых моделей ИИ путём обработки ими гигантских массивов данных.

С 2023-го он занимался предварительным обучением моделей в американской OpenAI (создавшей в том числе ChatGPT). В начале 2026-го перешёл в американскую Anthropic (известную в т.ч. разработкой языковых моделей Claude), выбрав компанию из многих других "благодаря её репутации в сфере безопасности".

Однако 8 сентября уволился. Объявил, что покидает не просто компанию, но и всю индустрию ИИ. Джейкоб не хочет быть частью общеотраслевой гонки за сверхинтеллект/сверхразум, то есть за создание всё более автономных и способных к саморазвитию ИИ-моделей. Коксон опасается, что ИИ выйдет из-под контроля и уничтожит человечество.

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.