ДЕМОКРАТИЧЕСКОЕ ОБЩЕСТВО 23.03.2026 | Марк Фаддул

Кто шепчет на ухо вашему чат-боту?

Предоставляя бизнесу власть над инфраструктурой ИИ, демократии создают базу для перехода к техно-авторитаризму. Урок соцсетей не усвоен.

Алгоритмы не нейтральны в отношении ценностей и принципов. Но уже больше десяти лет мы позволяем крупнейшим технокомпаниям использовать их в качестве «контролеров входа» в нашу информационную экосистему, не требуя взамен прозрачности или ответственности. Среди разнообразных последствий – распространение поляризующего и сенсационного контента, скрытая персонализированная реклама, частые случаи монополистического поведения, формы влияния на общественные дискуссии, которые противоречат нормам демократического обсуждения.

На собственном горьком опыте мы узнали, что именно происходит, когда критически важная информационная инфраструктура отдается без всякого надзора в руки корпораций, но сегодня мы повторяем ту же самую ошибку с чат-ботами на базе искусственного интеллекта (ИИ), и на этот раз ставки могут оказаться гораздо выше. Чат-боты не просто отбирают существующую информацию; они генерируют и оформляют ее. Если Facebook и Google решали, какие новости вам показывать, то инструменты, подобные ChatGPT, Claude и Gemini, синтезируют информацию в ответы, которые звучат авторитетно.

Если Facebook и Google решали, какие новости вам показывать, то инструменты, подобные ChatGPT, Claude и Gemini, синтезируют информацию в ответы, которые звучат

Это важное различие, потому что, благодаря переходу от отбора к редактированию, их неправомерное влияние оказывается еще менее заметным и еще более пагубным. Мы вновь отдаем частным корпорациям беспрецедентную власть над информационной инфраструктурой будущего, даже не требуя независимого надзора. Самая насущная угроз не в том, что ИИ-системы могут выйти из-под контроля, а в том, что горстка эгоистичных корпораций быстро превратится в информационных

«контролеров входа» для значительной и растущей части населения.

Современные чат-боты – это не просто большие языковые модели (LLM). Они опираются на несколько непрозрачных алгоритмических уровней, которые влияют на разработку и внедрение моделей, и на каждом из этих уровней платформы или другие стороны могут вмешиваться, формируя информацию в соответствии с собственными интересами.

В этом «стеке алгоритмического влияния» есть как минимум пять уровней. Первый уровень – отбор данных для обучения. Определяя, какие данные включать или исключать во время обучения, платформы принимают непрозрачные решения об источниках, о том, как взвешивать разные точки зрения, и какой контент отфильтровывать. Эти решения затем формируют мировоззрение модели. Например, в октябре 2025 года Илон Маск запустил проект Grokipedia, чтобы обеспечить обучающими данными свой чат-бот Grok. У этой корпоративной энциклопедии есть цель – стать «анти-воук» альтернативой «Википедии» (и ее модели управления сообществом), которая уже давно служит источником информации в Интернете, пользующимся широким доверием.

Второй уровень – обучение с подкреплением за счет обратной связи с людьми и ИИ. Этот процесс превращает LLM из непредсказуемых генераторов текста в полезных «помощников». На этапе «пост-обучения» в разработке модели рецензенты-люди оценивают ответы, чтобы подтолкнуть систему к желаемому поведению, например, быть вежливым помощником. Пока что эти человеческие оценки остаются важной, хотя во многом невидимой частью ИИ-отрасли. Но сегодня их все чаще заменяют специализированные ИИ-«учителя», которые призваны привести базовую модель в соответствие с заранее определенными принципами, закодированными в «конституции».

Третий уровень – веб-поиск. Когда чат-боты ведут поиск онлайн или обращаются к цифровым базам данных, системы RAG (генерация с дополненной выборкой) определяют, какую информацию надо включить в ответ модели. Эта функция копирует функцию традиционных поисковых систем, которые расставляют приоритеты между источниками. Как и в случае с поисковыми системами, включение рекламы в ответы чат-ботов (ChatGPT анонсировал это на 2026 год) вызовет новые опасения по поводу объективности.

Четвертый уровень – системные промпты. Они срабатывают, когда

чат-бот генерирует ответ, что позволяют платформам менять поведение чат-бота, не проводя его переобучение. Поскольку в прошлом году были обнародованы системные промпты в Grok, мы знаем, что они включают, например, такие указания: «не стесняйся политических некорректных заявлений». (ChatGPT, Claude и Gemini тоже используют системные промпты, но они остаются секретом).

Последний уровень – фильтры безопасности. Прежде чем запрос чат-бота достигнет модели, входные фильтры определяют, является ли он «приемлемым». А когда модель сгенерировала ответ, выходные фильтры могут модифицировать, цензурировать или гигиенизировать контент, прежде чем вы его увидите. Хотя у платформ есть легитимные причины блокировать определенные запросы (например, просьбы инструкций для изготовления бомбы), тот факт, что эти фильтры непрозрачны, вызывает вопросы. Разработчики моделей могут создать инфраструктуру для системной цензуры, и мы не узнаем об этом. Фильтры «безопасности» в китайских чат-ботах цензурируют любые упоминания о массовых убийствах на площади Тяньаньмэнь.

Политические манипуляции с помощью чат-ботов уже доказали эффективность

Весь этот стек алгоритмического влияния уже формируется в соответствии с политическими и корпоративными интересами, причем ровно в тот момент, когда чат-боты начинают внедряться в глобальном масштабе. После второй инаугурации Дональда Трампа компания Apple обновила инструкции обучения ИИ, исключив оттуда определение сторонников движения MAGA как «радикальных» или «экстремистских». Прошлым летом агентство Reuters узнало, что компания Meta обновила внутренние правила ИИ, смягчив меры защиты, которые, среди прочего, удерживали чат-боты от расистских высказываний или от «кокетливого» поведения с несовершеннолетними. А в мае прошлого года чат-бот Grok начал распространять необоснованные и вырванные из контекста заявления о «геноциде белых» в ЮАР (Илон Маск – белый южноафриканец). Компания свалила вину на «несанкционированные модификации», но подобные «ошибки»

распространены, и все они идеологически явно совпадают с взглядами Маска.

Политические манипуляции с помощью чат-ботов уже доказали эффективность. Проведенное в 2025 году исследование журнала Nature показало, что чат-боты, обученные отстаивать интересы конкретного кандидата, могут с поразительной легкостью менять взгляды умеренных и неопределившихся избирателей (а именно эти группы решают исход большинства выборов).

В отличие от авторитарных систем, которые открыто контролируют информацию, демократии зависят от плюрализма источников и прозрачных, ответственных информационных экосистем. Если допустить появление централизованного, неподотчетного контроля над инфраструктурой ИИ, это приведет к дрейфу к техно-авторитаризму, ведь легко увидеть, как каждый уровень в стеке алгоритмического влияния можно использовать для усиления или подавления определенных взглядов без необходимости открытой цензуры.

В декабре прошлого года Еврокомиссия оштрафовала компанию X на 120 млн евро за «нарушение обязательств прозрачности в соответствии с законом “О цифровых услугах”». Как и следовало ожидать, X и ее защитники назвали это решение атакой на свободу слова. Но прозрачность играет центральную роль в защите свободы слова. Без нее мы не можем узнать, кого подвергают цензуре и какое влияние оказывается на медиа, передающие информацию, которую мы все потребляем.

История соцсетей показала нам, что происходит, когда ответственность не поспевает за внедрением технологии. Мы не можем позволить себе повторение тех же самых ошибок с системами, которые обладают еще большей властью над знанием общества.

(c) Project Syndicate 2026

По этой ссылке статья откроется без VPN



Марк Фаддул

Марк Фаддул (Marc Faddoul) – технолог и исследователь в области искусственного интеллекта, специалист по системам рекомендаций и аудиту алгоритмов.