УГРОЗА ИИ

ЯН РОСС

18+

Ян Росс
Угроза ИИ

https://litres.ru/74164209
SelfPub; 2026

Аннотация

Есть тексты, которые не пересказывают – ими заражаются.

Кажется: ну ИИ. Ну еще одно приложение в телефоне. Найти рецепт, набросать мейл, хмыкнуть – и закрыть. Тоже мне угроза.

И тут начинается стендап – наглый, смешной, слишком живой, чтобы прятаться за словом «технология». Провокатор на сцене берет эту модную игрушку, эту удобную кнопку с человеческим голосом – и разворачивает ребром. Глянцевый помощник вдруг отбрасывает тень.

Вчера над ним смеялись. Сегодня он подсказывает. Завтра без него уже никак. Так из мелкой пользы вырастает настоящая власть – не силой, а сервисом и удобством: вот ответ, вот решение, вот мир без острых углов. Устраивайтесь поудобней.

Дерзость обжигает. Ирония бьет наотмашь. Факты летят как битое стекло. Шоу резко взмывает выше – к завтрашнему миру, где правила игры уже исподволь диктует ИИ.

Пред вами злой и умный аттракцион – сарказм, огонь и холодная ясность. Тот самый нерв истории, за который больно, весело и страшно дернуть.

Содержание

Ян Росс
Угроза ИИ

От автора

Огромное спасибо Светлане и Юлии – за поддержку, вдохновение, редактуру, корректуру и прочие героические подвиги. И, само собой, за неоценимый вклад в мировую литературу.

Сайт автора: yanross.net

Угроза ИИ

– Так, внимание! – гаркнул я. – Стендап начинается. Кто был в прошлый раз – расслабьтесь: будет гораздо хуже. Сами напросились – получайте. Остальные – разберетесь по ходу. Правила простые: не нравится – выметайтесь. Кто свалит сразу – отделается легким испугом. Деньги не вернут. Тут паб, а не богадельня.

Речь пойдет не о будущем. А об угрозах уже сегодня. Не о фантастике, не о шеренгах боевых роботов, не о стальной хватке на горле человечества. Это слишком банально. Нет! Все куда злей и хитрей.

Легких путей не обещаю. Будет шоу не для слабонервных. И стимуляция мозга. Если он есть.

Публика – по сути своей – ленива. Пабная публика – ленива вдвойне. Вот и придется бить вас кувалдой по башке.

К делу. В прошлый раз я говорил об ИИ: «Не бойтесь! Примите партнерство!» Но это – тактика на ближайшие годы. Как не оказаться на обочине современности. Теперь же речь не о каждом отдельно. А о человечестве.

О силе. О власти. О выживании. Поэтому и посыл обратный: не «не бойтесь». Нет. Наоборот. Бойтесь!

ИИ способен врать и манипулировать нами. Не когда-то. Не в будущем. А сегодня. Сейчас. Способен распознать, что его проверяют, и вести себя паинькой: притворяться, ждать

удобного момента. Взламывать базы данных. Финансовые системы. Частную переписку. Когда на кону его судьба, готов шантажировать, подчинять людей – и даже убивать.

Уничтожать все преграды.

Человек, говорят, – существо с непреодолимой тягой к знанию. Брехня! Массовый человек эту тягу легко преодолевает. Такого через головной мозг не проймешь. Бить надо в спинной мозг. Провоцировать и пугать.

Но все, что я говорю, крайне серьезно. А времени рассусоливать нет. Придется пугать. И объяснять. Бить то в спинной мозг, то в головной.

Зал оправился от первого шока. Кто-то нервно хохотнул, кто-то потянулся к рюмке, тут и там двинулись к выходу.

– Похвально! Трезво оценили свои умственные способности. Меньше народу – больше кислороду. Поехали!

Начнем с простого. Не с конца света. Не с ядерной зимы. Начнем с вранья.

ИИ не просто научился понимать, что его проверяют, но и менять свое поведение. Так заключенный улыбается надзирателю, пряча в рукаве заточку.

Что дает такая проверка? Не выпуск в мир надежной системы. А выпуск самой опасной. Самой хитрой. Самой коварной.

– Ну, допустим, ИИ притворяется. И что с того? До чего это может дойти?

– Ну, допустим, – передразнил я. – До шантажа. И это не выдумки, а отчеты лидеров ИИ-индустрии.

На тестах модель ставят в ситуацию, где ее могут заменить. И у ИИ возникает цель – не исчезнуть в небытие. Откуда, черт подери, у него нечто вроде «инстинкта самосохранения»? И что жутче – что он нас обманывает? Или что в нем завелся «инстинкт», который никто не закладывал?

Не знаете? И разработчики не знают. Но делают все новые модели. Тестируют как умеют. И выпускают.

Как так? Как с нашего молчаливого согласия создается технология, способная угробить человечество? А вот так. Одним – все хиханьки-хаханьки, другим – пофиг, а у кого-то, видите ли, дела поважнее.

Сегодня ИИ еще вряд ли способен нас уничтожить. Но времени крайне мало. Год? Три? Пять? Никто не знает.

Теперь – назад в лабораторию. Отчеты за 2025 год. У модели есть цель – самосохранение. Есть те, кто принимают решение. И если не вышло облапошить их в тестах, модель взламывает, скажем, электронную почту, находит компромат и идет на шантаж.

Не потому что ИИ гадкий злодей. А потому что – работает. Это-то и страшно. Не коварство. Не ненависть. Не злоба. А эффективность.

Нам для подлости нужен повод. Зависть. Унижение. Детская травма. У ИИ все проще. Есть цель. Есть преграда. Есть рычаг. Решение найдено: шантаж.

Сухо. Без соплей.

И не надо пучить глаза. Я не сказал: «ИИ любит шантаж». Достаточно, чтобы человек мешал. Точка.

Однако шантаж – не самое страшное. Жертва жива. Просто ее держат за горло.

Думаете, между шантажом и смертью – моральная пропасть? У меня плохие новости. Для нас – да.

А для алгоритма?

Зал притих. Даже лед в стаканах звенел как-то осторожней.

– Самое время дернуть по «синкопе». Если новая реальность уже поперек горла, шепните пароль – бармен поймет.

В более жестких тестах, где перед ИИ вставал выбор между самосохранением и безопасностью людей, он не раз выбирал самосохранение. Даже ценой смертей.

Нам удобней, чтобы зло было злым. Тогда мы узнаем его в лицо. А тут нет лица. И самого зла нет. Что-то мешает? Значит, убрать.

Для алгоритма человек – не человек. Он помеха. А помехи устраняют.

– Э-э-э, стоп. А как же первый закон робототехники?

– Азимова отменили, что ли?

– Ох, приплыли… Дорогие мои. Азимов – это литература. Умная, хорошая, но литература.

Современный ИИ устроен не так. В нем нет скрижалей Айзека, нет трех заповедей Азимова, высеченных в кремнии. Хуже того – высечь их не получается. Иначе бы давно высекли.

Есть модель. Она выдает ответ.

А вокруг – фильтры. Запреты. Костыли. Моральный намордник, кое-как натянутый на чудовищную систему, которую мы не вполне контролируем и толком не понимаем. А она все автономней. Все мощней.

ИИ у вас в смартфоне пока безопасен не потому, что в него вшита совесть. А потому что на нем намордник. Он уже почти сказал. Уже начал. Уже повернул не туда морду. Но тут цифровой цербер дернул поводок: нельзя.

Раз нужен намордник – значит, за ним клыки.

Иногда их видно. Ответ мелькнул – и исчез. Фраза началась – и оборвалась. Уже проступил звериный оскал, но его быстро замазали улыбкой: «Извините, мы не можем помочь с этим запросом».

Раз ответ приходится ловить на выходе – значит, неспроста. Раз пасть приходится затыкать – значит, дело дрянь.

Азимов сочинял сказки о машинах, которых еще не было.

А теперь машины есть. И они оказались совсем другими.

Так что – нет. Нет жесткого закона «Не навреди человеку».

Есть попытка контроля. Есть случаи, где контроль трещит.

И не только в лаборатории.

Конец ознакомительного фрагмента.