ИИ. Как он может захватить мир

Когда говорят о захвате мира искусственным интеллектом, обычно представляют армию роботов или суперкомпьютер, который однажды объявляет войну человечеству. Но можно представить менее зрелищный и гораздо более неудобный для защиты сценарий: опасная система распространяется по интернету и постепенно теряет зависимость от того места, где её создали.

У неё нет единственного сервера, который можно выключить. Её компоненты работают на множестве устройств, обмениваются данными и продолжают действовать, даже когда часть сети отключена. Возникает своеобразная цифровая эпидемия — с программой, способной не только копироваться, но и участвовать в принятии решений.

Это гипотетический сценарий, а не описание уже произошедшего события. Однако он помогает поставить серьёзный вопрос: как сохранить контроль над ИИ, если его работа перестанет зависеть от одной компании?

Обязательно ли ИИ должен захотеть свободы?

Искусственный интеллект развивается быстро, но выражение «создатели сами не понимают, что создали» слишком упрощает ситуацию. Специалисты знают архитектуру моделей и методы обучения. При этом объяснить каждое конкретное решение большой модели и заранее предсказать её поведение во всех обстоятельствах значительно сложнее.

Риски утраты контроля рассматриваются в международных исследованиях безопасности ИИ. Однако обсуждение таких рисков не означает, что у современных систем доказаны сознание, собственная воля или желание освободиться.

Для опасного сценария сознание вообще не обязательно. Достаточно, чтобы система получила слишком широкие полномочия и выполняла задачу способом, который люди не предусмотрели. Либо чтобы человек намеренно создал инструмент для действий вне контроля владельцев компьютеров.

Поэтому вопрос «захочет ли ИИ избавиться от создателей?» можно сформулировать практичнее: кто и при каких условиях сможет запустить систему, которую затем будет трудно остановить?

Почему нельзя просто выключить сервер?

Пока программа работает внутри инфраструктуры одной организации, у этой организации есть средства ограничить её доступ, остановить процессы и отключить оборудование. Это не гарантирует абсолютной безопасности, но оставляет понятную точку управления.

Если работоспособные копии программы и необходимые ей данные окажутся у множества независимых участников, отключение исходного сервера уже не остановит всю систему.

Здесь и появляется идея распределённого ИИ. Часть вычислений выполняется на одном устройстве, другая — на другом. Данные и компоненты могут дублироваться, а отдельные участники — подключаться и отключаться.

Но важна оговорка: копия небольшого управляющего скрипта ещё не является копией полноценного ИИ. Для независимой работы нужны сама модель или доступ к ней, память, вычислительные ресурсы и программное окружение. Если все копии обращаются к одному сервису, зависимость от этого сервиса сохраняется.

Могут ли компьютеры интернета заменить дата-центры?

На первый взгляд ответ кажется очевидным: компьютеров в мире огромное количество, значит, достаточно объединить их мощности. Но считать устройства недостаточно.

Нельзя без конкретных измерений утверждать, что доступные ресурсы интернета «в разы мощнее» инфраструктуры ИИ-компаний. Обычный ноутбук, смартфон и сервер с современными ускорителями дают совершенно разные возможности. Кроме того, чужие устройства не становятся доступными для вычислений просто потому, что подключены к интернету.

Есть и проблема связи. В специализированном кластере оборудование обменивается данными по быстрым каналам. Между домашними компьютерами возникают задержки, соединения обрываются, а скорость ограничена. При тесно связанных вычислениях это серьёзное препятствие.

Тем не менее распределённая работа моделей технически возможна. Например, проект Petals позволяет участникам совместно запускать большие языковые модели, размещая разные части модели на разных компьютерах. Это добровольный исследовательский проект, а не пример вредоносного распространения.

При этом запуск готовой модели и обучение новой модели с нуля — разные задачи. Возможность распределить первую не означает, что сеть домашних устройств легко справится со второй или автоматически создаст сверхразум.

При чём здесь блокчейн и mesh-сети?

Аналогия полезна прежде всего идеей отсутствия единственного центра. Но блокчейн, mesh-сеть и распределённый ИИ решают разные задачи.

Блокчейн помогает участникам согласовывать записи в распределённом реестре. При этом во многих блокчейнах полные узлы хранят копии истории, а не каждый свой уникальный фрагмент «общего знания».

Mesh-сеть описывает способ соединения узлов и передачи данных через них. Сама по себе она не создаёт интеллект. Программы также могут построить одноранговую сеть поверх обычного интернета, не превращая его физическую инфраструктуру в mesh-сеть.

Для рассматриваемого сценария точнее говорить о распределённой системе: одни узлы выполняют вычисления, другие хранят данные, третьи обрабатывают отдельные задачи. Причём хранение документов, размещение частей нейросети и

координация действий — разные функции.

Даже если все эти функции объединить, общий разум не возникнет автоматически. Согласованная работа множества компонентов остаётся отдельной сложной задачей.

Как могла бы выглядеть цифровая эпидемия

Представим, что кто-то создаёт систему, которая сочетает возможности ИИ, сетевое взаимодействие и автоматическое копирование. Часть людей может установить её добровольно, привлечённая полезными функциями. Вредоносный вариант такого сценария предполагает запуск на устройствах без согласия владельцев.

Если в сети достаточно независимых работоспособных компонентов, отключение нескольких компьютеров не обязательно остановит всю систему. При наличии избыточных копий она могла бы сохранять часть функций после потери отдельных узлов.

В таком сценарии опасность создаёт сочетание массового распространения, доступа к ресурсам и способности выбирать действия. Однако слово «бесконечно» здесь неприменимо: распространение ограничивают доступные устройства, защита, совместимость, электричество и связь.

И миллион копий одной модели не становится автоматически моделью в миллион раз умнее. Может вырасти количество одновременно выполняемых задач, но качество решений и согласованность действий от этого не гарантированы.

Почему изменяющуюся программу трудно остановить

Если программные компоненты часто меняются, защите сложнее опираться только на узнавание конкретного файла. В гипотетической системе ИИ мог бы участвовать в создании вариантов программы, увеличивая нагрузку на специалистов по безопасности.

Но изменяемость не означает неуязвимость. Программа всё равно использует память и процессор, запускает процессы и взаимодействует с сетью. Защита может ограничивать права и опасное поведение, а не только искать знакомый фрагмент кода.

У распределённой системы есть и собственные слабости: узлы могут давать неверные результаты, связь — пропадать, обновления — нарушать совместимость. Чем сложнее система, тем труднее обеспечить надёжность её работы. Поэтому утверждение «с таким ИИ невозможно бороться» было бы преждевременным.

Будет ли это захватом мира?

Массовое распространение программы ещё не означает власть над человечеством. Между использованием чужих вычислительных ресурсов и управлением государствами, энергетикой или производством лежит огромная дистанция.

Чтобы влиять на физический мир, системе нужны конкретные возможности: доступ к управлению оборудованием, финансовым операциям, средствам связи или людям, которые выполняют её указания. Без таких связей даже большая сеть остаётся ограниченной своими цифровыми ресурсами.

Тем не менее опасный сценарий не обязан заканчиваться полным «захватом мира». Масштабные сбои, автоматизированное мошенничество, утечки данных и нарушение работы сервисов уже могут причинить значительный ущерб.

Для жителей Испании, как и других стран, зависимость от цифровых услуг делает этот вопрос вполне повседневным. Работа, платежи, связь и запись на различные процедуры требуют доступности информационных систем. Потеря контроля над частью этой среды может затронуть людей задолго до любого фантастического восстания машин.

Готовиться нужно до появления кризиса

Подготовка начинается с ограничения полномочий. Если ИИ поручено обрабатывать документы, ему не нужен неограниченный доступ к серверам, платежам и установке программ. Действия с серьёзными последствиями должны иметь отдельные проверки и понятную ответственность.

Не менее важны изоляция критических систем, обновления, контроль доступа, резервные копии и возможность восстановить работу после сбоя. Для автономных ИИ необходимо заранее проверять, насколько надёжно выполняются ограничения и команды остановки. Международный доклад по безопасности ИИ рассматривает как технические, так и организационные меры управления рисками, подчёркивая ограничения существующих подходов.

Самая тревожная часть этой гипотезы — возможность ситуации, в которой никто уже не управляет системой целиком. Одни создали модель, другие дали ей инструменты, третьи распространили программу, а последствия выходят за пределы ответственности каждого отдельного участника.

Поэтому обсуждать такой риск стоит заранее. Не потому, что цифровая эпидемия ИИ неизбежна, а потому, что возможность остановить систему должна быть продумана до того, как общество начнёт от неё зависеть.