Карен Хао

Империя ИИ. Мечты и кошмары OpenAI Сэма Альтмана

Моей семье – прошлой, настоящей и будущей.

Движениям во всём мире, которые не дают себя обобрать во имя изобилия.

Говорят, объяснить – значит развенчать. Нигде эта максима не оправдывается так полно, как в программировании, особенно в том, что называют эвристическим программированием и искусственным интеллектом. Ведь там машины заставляют вести себя столь чудесно, что это способно ослепить даже самого опытного наблюдателя. Но стоит разоблачить конкретную программу, стоит объяснить её внутреннее устройство языком достаточно простым, чтобы всё стало понятно, – и её волшебство рассыпается: она предстаёт всего лишь набором процедур, каждая из которых вполне постижима. Наблюдатель говорит себе: «Я и сам мог бы такое написать». И с этой мыслью переставляет программу с полки с надписью «разумное» на полку для диковинок, о которых стоит говорить разве что с людьми, менее просвещёнными, чем он.

Джозеф Вейценбаум, профессор MIT, создатель первого чат-бота ELIZA, 1966

«Успешные люди создают компании. Более успешные создают страны. Самые успешные создают религии». Я слышал это от Ци Лу; не знаю, кому принадлежит эта мысль. Но она заставила меня задуматься: самые успешные основатели вовсе не ставят себе целью создать компанию. Их миссия – создать нечто, гораздо более похожее на религию, и в какой-то момент выясняется, что проще всего это сделать, основав компанию.

Сэм Альтман, 2013

От автора

Эта книга основана более чем на 300 интервью примерно с 260 людьми, а также на обширном собрании переписки и документов. Большинство интервью я записала специально для неё. Часть взята из моей работы за последние семь лет, когда я писала об OpenAI, об ИИ-индустрии и её влиянии на мир для MIT Technology Review, The Wall Street Journal и The Atlantic. Более 150 интервью я взяла у 90 с лишним нынешних и бывших руководителей и сотрудников OpenAI, а также у нескольких подрядчиков, имевших доступ к подробной документации о том, как компания разрабатывает свои модели. Ещё часть интервью – с примерно 40 нынешними и бывшими руководителями и сотрудниками Microsoft, Anthropic, Meta, Google, DeepMind и Scale, а также с людьми из ближнего круга Сэма Альтмана.

Все цитаты из электронных писем, документов и сообщений в Slack взяты из копий или снимков экрана этих документов и переписки либо приведены точно в том виде, в каком они фигурируют в судебных исках. Если копии у меня нет, я пересказываю текст без кавычек. Есть одно исключение, которое я отмечаю в примечаниях. Все диалоги восстановлены по воспоминаниям участников, по записям, сделанным в то время, или – что отмечено в примечаниях – взяты из аудиозаписей и расшифровок. Как правило, я или мои фактчекеры просили тех, кто приводил цитаты, повторить или подтвердить их ещё раз спустя несколько месяцев, чтобы проверить, не меняются ли они. Каждая сцена, каждая цифра, каждое имя и кодовое название, каждая техническая подробность о моделях OpenAI – например, состав обучающих данных или число чипов, на которых модель обучали, – подтверждены как минимум двумя людьми, записями и документами того времени или, в немногих случаях, отмеченных в примечаниях, публикациями других СМИ. Это же относится почти ко всем остальным подробностям об OpenAI в книге. Если я называю кого-то по имени, это не значит, что я говорила с этим человеком сама. Когда я пишу о чьих-то мыслях или чувствах, то лишь потому, что человек описал их – мне, кому-то из моих собеседников, в письме или записи, оказавшейся в моём распоряжении, или в публичном интервью.

Это не книга о корпорации. Хотя в ней рассказана история OpenAI изнутри, эта история задумана как призма, сквозь которую можно увидеть гораздо больше, чем одну компанию. Это портрет научного замысла, превратившегося в агрессивный идеологический поход, подпитываемый деньгами; исследование его многообразных и далеко идущих последствий; размышление о власти. Ради этого во время работы над книгой я подолгу жила среди местных сообществ в разных странах мира, чтобы понять их историю, культуру, жизнь и то, каково им приходится, когда ИИ напрямую вторгается в их существование. Я надеюсь, что

их истории звучат на этих страницах так же отчётливо, как истории, разыгравшиеся в стенах одной из самых скрытных организаций Кремниевой долины.

Я обращалась ко всем ключевым людям и компаниям, о которых идёт речь в книге, с просьбой об интервью и комментариях. OpenAI и Сэм Альтман предпочли не сотрудничать.

Борьба за трон

В пятницу, 17 ноября 2023 года, около полудня по тихоокеанскому времени Сэм Альтман – генеральный директор OpenAI, золотой мальчик Кремниевой долины и живое олицетворение революции генеративного искусственного интеллекта – вошёл в Google Meet. На экране его уже ждали четверо из пяти членов совета директоров.

Илья Суцкевер, главный научный сотрудник OpenAI, ходить вокруг да около не стал: Альтман уволен, официальное заявление выйдет через несколько минут.

Альтман находился в это время в номере роскошного отеля в Лас-Вегасе. Он прилетел на гонку «Формулы-1» – первую в городе за много десятилетий, громкое светское событие, ради которого в Вегас съехались все, от Рианны до Дэвида Бекхэма. Поездка была короткой передышкой в изматывающем графике перелётов, который Альтман почти без остановки выдерживал с тех пор, как примерно год назад OpenAI выпустила ChatGPT. Несколько секунд он не мог вымолвить ни слова. Отвёл глаза, пытаясь собраться. А потом, пока разговор шёл дальше, сделал то, что умел лучше всего: попытался сгладить углы.

– Чем я могу помочь? – спросил он.

Совет попросил его поддержать временного генерального директора, которого уже наметили ему на смену, – Миру Мурати, до сих пор отвечавшую в OpenAI за технологии. Альтман всё ещё плохо понимал, что происходит. Всё это было похоже на дурной сон.

Но он согласился. Через несколько минут Суцкевер отправил новую ссылку на Google Meet Грегу Брокману – президенту OpenAI, ближайшему союзнику Альтмана и единственному члену совета, которого на первый разговор не позвали. Брокману сообщили, что он больше не входит в совет директоров, хотя свою должность в компании пока сохраняет.

Вскоре вышло и официальное заявление: «Уход г-на Альтмана стал итогом тщательного рассмотрения, проведённого советом директоров. Совет пришёл к выводу, что г-н Альтман не был последовательно откровенен в общении с советом, и это мешало совету исполнять свои обязанности. Совет больше не уверен в его способности и дальше руководить OpenAI». Со стороны всё это выглядело почти абсурдом.

OpenAI была на пике могущества. После запуска ChatGPT в ноябре 2022 года компания стала самой громкой историей успеха Кремниевой долины, а сам ChatGPT – самым быстрорастущим потребительским приложением в истории. Стоимость стартапа росла так стремительно, что у инвесторов текли слюнки, а лучшие специалисты индустрии мечтали попасть на борт этой ракеты. За

несколько недель до увольнения Альтмана OpenAI оценивали почти в 90 миллиардов долларов. Компания как раз заканчивала оформлять сделку, по которой сотрудники могли продать часть своих акций инвесторам, выстроившимся в очередь. А всего за несколько дней до кризиса OpenAI провела долгожданную, встреченную на ура презентацию и показала самую амбициозную линейку продуктов за всю свою историю.

В глазах публики всего этого добился Сэм Альтман. Весной и летом он объехал весь мир и стал чуть ли не поп-звездой: журналисты уже сравнивали его известность с известностью Тейлор Свифт. Людей подкупало странное сочетание неброской внешности, громких заявлений и почти обезоруживающей искренности.

Незадолго до Лас-Вегаса он снова колесил по свету и выступал на саммите руководителей АТЭС, по обыкновению легко роняя фразы, которые тут же разлетались на цитаты.

– Почему вы решили посвятить этому жизнь? – спросила его Лорен Пауэлл Джобс, основательница и глава Emerson Collective, вдова Стива Джобса.

– Я думаю, это будет самая преобразующая и самая полезная технология из всех, что когда-либо создавало человечество, – ответил Альтман. – За историю OpenAI мне уже четыре раза довелось оказаться в комнате в тот момент, когда мы будто отодвигали завесу невежества и сдвигали границу познания. Последний раз – буквально пару недель назад. Быть при этом – величайшая честь в моей профессиональной жизни.

Сотрудники OpenAI узнали об увольнении Альтмана почти одновременно со всем миром. Ссылка на заявление мгновенно разлетелась с телефона на телефон. Больше всего людей потрясало то, насколько новость не вязалась с репутацией Альтмана. В OpenAI к тому времени работало около восьмисот человек, и у многих было всё меньше случаев пообщаться с гендиректором лично. Но тот обаятельный Альтман, которого весь мир видел на сценах международных форумов, почти ничем не отличался от человека, которого сотрудники встречали на общих собраниях, корпоративах или – когда он не был в разъездах – просто в коридорах офиса. Слухи поползли мгновенно.

Люди лихорадочно листали X, бывший Twitter, в поисках хоть какой-нибудь зацепки. И вдруг кто-то в офисе ухватился за объяснение, которое на секунду показалось самым логичным.

– Альтман идёт в президенты! – крикнул он.

На мгновение все выдохнули. Потом поняли, что нет. И догадки вспыхнули с новой силой, теперь уже вперемешку со страхом. Что он натворил? Что-то противозаконное? Может, всё дело в его сестре? Месяцем раньше она опубликовала посты, которые разошлись по сети: в них она обвиняла брата в насилии. А если не преступление, то какой-то серьёзный этический проступок? Может, речь о других инвестициях Альтмана? Или о его переговорах с

саудовскими инвесторами, у которых он искал деньги на новый проект по производству чипов для ИИ?

Суцкевер тем временем написал во внутреннем Slack OpenAI, что через два часа проведёт онлайн-собрание для всех сотрудников и ответит на их вопросы.

– Это были самые долгие два часа в моей жизни, – вспоминал потом один из них.

Когда собрание началось, на экране рядом появились Суцкевер, Мурати и оставшиеся руководители компании. Держались они скованно и выглядели явно неподготовленными. Трансляцию смотрели и в офисе, и из дома.

Суцкевер был мрачен. В OpenAI его знали как глубокого мыслителя, человека почти мистического склада. О технологиях он постоянно говорил в духовных категориях, причём с такой неподдельной серьёзностью, что одних это трогало, а других раздражало. Но была у него и совсем другая сторона: он мог вести себя как большой ребёнок – добродушный и немного чудаковатый. Приходил на работу в футболках с животными и сам любил их рисовать: умильного кота, умильных альпак, умильного огнедышащего дракона, а рядом – абстрактные лица и самые обыденные предметы. Несколько его любительских картин висели прямо в офисе. На одной были нарисованы три цветка, распустившиеся в форме логотипа компании, – символ того, к созданию чего Суцкевер постоянно призывал коллег: «множества любящих человечество AGI». Теперь перед ним стояла совсем другая задача.

Ему предстояло успокоить сотни встревоженных людей, которые с бешеной скоростью вбивали вопросы в общий онлайн-документ. А Суцкевер был не лучшим человеком для такой роли. Думать он умел. Говорить о будущем человечества – тоже. А вот сообщать плохие новости так, чтобы люди могли их принять, у него получалось куда хуже.

Мурати зачитала первый вопрос:

– Был ли какой-то конкретный инцидент, который привёл к этому решению?

– Многие вопросы в документе касаются подробностей: что именно произошло, когда, как, кто конкретно… – ответил Суцкевер. – Хотел бы я рассказать детали. Но не могу.

Тем, кто хочет понять больше, добавил он, стоит внимательно прочитать официальное заявление.

– Там на самом деле очень многое сказано. Возможно, его стоит перечитать несколько раз.

Сотрудники были ошеломлены. Им только что сообщили новость, способную перевернуть всю компанию. Разве они, люди, которых всё это касается напрямую, не заслуживали знать больше, чем случайный читатель пресс-релиза? Мурати продолжила.

Что теперь будет с Microsoft? Вопрос был далеко не праздный. Microsoft была крупнейшим инвестором OpenAI и владела эксклюзивной лицензией на её

технологии. Кроме того, именно Microsoft давала компании всю вычислительную инфраструктуру. Без неё встало бы почти всё, чем занималась OpenAI: исследования, обучение моделей, выпуск новых продуктов. Мурати сказала, что серьёзных последствий не ждёт: они только что говорили с генеральным директором Microsoft Сатьей Наделлой и техническим директором Кевином Скоттом.

– Они полностью поддерживают нашу работу, – заверила она.

Следующий болезненный вопрос: что будет со сделкой, по которой сотрудники продают свои акции? Тем, кто проработал в OpenAI достаточно долго, она сулила миллионы долларов. Закрыть её должны были совсем скоро. Некоторые уже собирались покупать дома, а кто-то успел и купить. Главный операционный директор Брэд Лайткэп замялся.

– Что касается сделки… ну… посмотрим, – сказал он.

Он добавил, что держит связь с инвесторами, которые её возглавляют, и с крупнейшими нынешними акционерами.

– Все они заверили нас в неизменной поддержке.

После ещё нескольких туманных ответов кто-то снова попытался задать главный вопрос. Что всё-таки сделал Сэм? Это связано с работой? Или с личной жизнью? Суцкевер опять отослал всех к пресс-релизу.

– Ответ там, – сказал он.

Мурати прочитала следующий вопрос:

– На вопросы о подробностях когда-нибудь ответят? Или никогда?

– Не ждите слишком многого, – ответил Суцкевер.

Собрание шло своим чередом, ответы Суцкевера всё больше расходились с настроением зала, и тревога людей быстро перерастала в гнев.

Суцкевер попытался придать происходящему чуть ли не философский смысл:

– Когда группа людей вместе проходит через тяжёлое испытание, они часто становятся сплочённее и ближе друг другу. Это трудное время сделает нас ещё более единой командой – а значит, и более продуктивной.

В документе появился новый вопрос: «Как можно говорить, что кризис должен нас сплотить, и при этом отказываться от прозрачности? Обычно без правды примирения не бывает».

– Ну… справедливое замечание, – признал Суцкевер. – Ситуация не идеальная.

Мурати попыталась сбить нарастающее раздражение:

– Наша миссия гораздо больше любого из нас.

Лайткэп её поддержал: партнёры, клиенты и инвесторы OpenAI один за другим подчёркивают, что по-прежнему верят в миссию компании.

– Теперь, пожалуй, наша ответственность перед этой миссией стала ещё больше.

Суцкевер снова попытался воодушевить людей:

– У нас есть всё необходимое. Абсолютно всё. Вычислительные мощности. Исследования. Наши прорывы потрясают воображение. Когда вам тревожно, когда страшно – вспоминайте об этом. Представьте себе масштаб нашего вычислительного кластера. Просто вообразите все эти GPU, работающие вместе.

Появился новый вопрос, и Мурати прочитала его вслух:

– Не опасаемся ли мы враждебного захвата компании через давление на нынешних членов совета директоров?

В голосе Суцкевера впервые прорвалось раздражение.

– Враждебного захвата? – переспросил он, выделяя каждое слово. – Некоммерческий совет директоров OpenAI действовал строго в соответствии со своей задачей. Это не враждебный захват. Ни в коем случае. Я категорически не согласен с такой постановкой вопроса.

В тот вечер несколько сотрудников собрались дома у коллеги на вечеринку, которую запланировали ещё до увольнения Альтмана. Пришли и люди из других ИИ-компаний, в том числе из Google DeepMind и Anthropic.

Незадолго до начала всем гостям пришло сообщение: «Сегодня у нас появится ещё одна тематическая комната – „Комната, где не говорят об OpenAI“. До встречи!» Почти никто там не задержался: все хотели говорить именно об OpenAI. В тот же день Грег Брокман объявил, что уходит из компании в знак протеста.

Сатья Наделла был в ярости: Microsoft предупредили об увольнении Альтмана буквально за несколько минут до объявления. И всё же его публичное заявление было выверено до последнего слова. Наделла написал, что у Microsoft есть долгосрочное соглашение с OpenAI, что компания располагает полным доступом ко всему, что нужно для её инновационной стратегии, и по-прежнему дорожит партнёрством с OpenAI, Мирой Мурати и её командой. А слухи всё множились.

Потом пришла новая весть. Из OpenAI ушли ещё трое ведущих исследователей: Якуб Пахоцкий и Шимон Сидор, одни из старейших сотрудников компании, и профессор MIT Александр Мадри, который пришёл в OpenAI сравнительно недавно. Многие восприняли это как тревожный сигнал. Компания лишалась руководства, лишалась талантов. Следующими могли дрогнуть инвесторы, а за ними – сорваться сделка по акциям. В худшем случае могла развалиться и сама OpenAI. Настроение на вечеринке становилось всё мрачнее. Если сделка рухнет, сотрудники лишатся огромных денег, заработанных годами тяжёлого труда. Но ещё страшнее было подумать, что исчезнет сама компания, а вместе с ней все вложенные силы и все надежды.

Тем временем той же ночью совет директоров и оставшееся руководство OpenAI проводили одно совещание за другим, и с каждым разговором атмосфера накалялась. Видимость единства, которую Суцкевер и остальные пытались изобразить на общем собрании, рассыпалась почти сразу после трансляции. Многие из тех, кто сидел с ним рядом на экране, на самом деле были

ошеломлены не меньше рядовых сотрудников: об увольнении Альтмана они сами узнали за несколько минут до официального сообщения. А после провального выступления Суцкевера их раздражение переросло в ярость. Они потребовали встречи с остальными членами совета. Около дюжины руководителей, включая Мурати и Лайткэпа, собрались в переговорной.

Суцкевер подключился удалённо вместе с тремя независимыми директорами. Это были Адам Д'Анджело, сооснователь и генеральный директор сервиса вопросов и ответов Quora; Таша Макколи, предпринимательница и старший внештатный научный сотрудник RAND; и Хелен Тонер, исследовательница из Австралии, работавшая в Центре безопасности и новых технологий Джорджтаунского университета (CSET).

На них обрушился шквал вопросов. Но все четверо раз за разом отказывались вдаваться в подробности, ссылаясь на юридические обязательства о конфиденциальности. Некоторые руководители уже не скрывали бешенства.

– Вы утверждаете, что Сэму нельзя доверять, – резко сказала вице-президент по глобальным вопросам Анна Маканджу, не раз сопровождавшая Альтмана в его мировом турне. – Но это совершенно не совпадает с нашим опытом работы с ним.

Руководители потребовали, чтобы члены совета немедленно ушли в отставку и передали свои места трём сотрудникам компании. Иначе, пригрозили они, они уволятся все разом. Главный директор по стратегии Джейсон Квон, юрист, прежде занимавший в OpenAI пост главного юрисконсульта, пошёл ещё дальше. По его словам, отказ совета уйти в отставку был по сути незаконным: если компания из-за этого развалится, директора нарушат свои фидуциарные обязанности. Совет этот довод отверг.

Директора настаивали, что перед увольнением Альтмана тщательно консультировались с юристами и действовали строго в пределах своих полномочий. В конце концов, OpenAI была необычной компанией, а её совет – необычным советом. Структуру организации во многом придумал сам Альтман. Она давала совету чрезвычайно широкие полномочия и обязывала его действовать не в интересах акционеров, а в интересах миссии: добиться того, чтобы AGI – общий искусственный интеллект – служил на благо всего человечества. Сам Альтман годами называл право совета уволить его важнейшим механизмом управления OpenAI. Хелен Тонер напомнила об этом предельно жёстко:

– Даже если это решение уничтожит компанию, оно всё равно может соответствовать нашей миссии.

Руководители тут же пересказали её слова сотрудникам, и в пересказе они прозвучали ещё страшнее: Тонер всё равно, погибнет компания или нет. Кое-кто начал подозревать, что именно этого она и добивается. На вечеринке один из гостей, представив, что потеряет все свои акции, не выдержал и расплакался.

На следующий день, в субботу 18 ноября, в особняк Альтмана стоимостью 27 миллионов долларов набилось несколько десятков человек, в том числе сотрудники OpenAI. Все ждали новостей.

Трое ушедших ведущих исследователей – Пахоцкий, Сидор и Мадри – встретились с Альтманом и Брокманом, чтобы обсудить, не основать ли новую компанию и не продолжить ли работу там. Одних эти слухи напугали ещё больше: появление нового конкурента могло окончательно расшатать OpenAI. Другие, наоборот, увидели в этом надежду. Если Альтман и правда создаст новую компанию, они уйдут за ним.

Оставшееся руководство OpenAI предъявило совету ультиматум со сроком до пяти вечера субботы по тихоокеанскому времени: вернуть Альтмана и уйти в отставку, иначе из компании уйдут люди. Совет отказался. Все выходные директора лихорадочно обзванивали знакомых, иногда посреди ночи, в надежде найти хоть кого-нибудь, кто возьмёт трубку и сможет помочь. Гнев сотрудников и инвесторов рос. Мурати больше не хотела оставаться временным гендиректором. Совету нужно было либо найти ей замену – человека, способного вернуть компании хоть какую-то устойчивость, – либо подобрать новых директоров, которые смогут противостоять Альтману, если тот всё-таки вернётся.

Вечером, когда срок ультиматума уже истёк, Джейсон Квон разослал сотрудникам служебную записку: «Мы продолжаем работать над разрешением ситуации и сохраняем оптимизм». И пояснил: «Под разрешением ситуации мы понимаем возвращение Сэма, Грега, Якуба, Шимона и Александра».

Вскоре Альтман написал в X в своей фирменной манере, строчными буквами: «я так сильно люблю команду openai».

Десятки сотрудников стали делиться его постом, добавляя к нему сердечко.

В воскресенье Альтман и Брокман приехали в офис OpenAI договариваться о своём возвращении. В течение дня туда подтягивалось всё больше сотрудников; они сидели и ждали. Большинство из них, как и руководители, и члены совета, толком не спали уже больше полутора суток, и всё слилось в один бесконечный мутный день. Альтман выложил селфи: сжатые губы, сдвинутые брови, в руке гостевой пропуск. Подпись гласила: «первый и последний раз ношу такую штуку». Руководство дало совету новый срок, снова до пяти вечера: вернуть Альтмана и уйти в отставку.

Теперь давили со всех сторон. Microsoft, инвесторы OpenAI и самые влиятельные люди Кремниевой долины публично встали на сторону Альтмана. В прессу утёк сценарий дальнейших событий: если совет не отменит решение, сотрудники уйдут массово, Microsoft перекроет доступ к вычислительным мощностям, инвесторы подадут в суд. Всё вместе сделало бы существование OpenAI без Альтмана практически невозможным.

Но совет стоял на своём. К девяти вечера, через несколько часов после того, как истёк очередной ультиматум, Суцкевер выложил в Slack от имени совета длинное сообщение. Альтман не вернётся. Новым временным генеральным

директором OpenAI назначен Эмметт Шир, бывший глава Twitch. Через пять минут Шир и Суцкевер приедут в офис и расскажут сотрудникам о новом видении будущего компании. Суцкевер писал:

«Совет твёрдо убеждён, что принятое решение – единственный путь, позволяющий продвигать и защищать миссию OpenAI. Проще говоря, поведение Сэма и недостаток прозрачности в его отношениях с советом подорвали способность совета эффективно осуществлять надзор за компанией в соответствии с возложенными на него полномочиями».

Slack взорвался. Ответы посыпались один за другим: «И с какой, блядь, армией вы это сделаете?», «Вы бредите», «Эмметт будет гендиректором пустого места». Около двухсот сотрудников демонстративно покинули офис, бойкотируя выступление. Мурати поспешно увела из здания руководителей. Когда Шир и Суцкевер приехали, в зале оставалось человек двенадцать. К Суцкеверу подошла Анна Брокман, жена Грега.

Четыре года назад именно Суцкевер вёл гражданскую церемонию их бракосочетания. Теперь Анна плакала. Она обняла его и умоляла передумать. Многие из ушедших разъехались по домам коллег, чтобы переждать эту ночь вместе.

Сотни человек вступили в группу в Signal, куда стекались последние новости. Поздно вечером Сатья Наделла объявил: Microsoft берёт Сэма Альтмана и Грега Брокмана руководить новым подразделением искусственного интеллекта. Следом разошлась новость ещё важнее: каждому сотруднику OpenAI, который захочет последовать за Альтманом, гарантировано место в Microsoft.

Настроение переменилось мгновенно: страх уступил место вызову. Теперь у сотрудников был запасной аэродром, а значит, и новые рычаги давления на совет и Шира. В доме одного из них собралось больше сотни работников OpenAI. Руководители и ведущие исследователи взялись за открытое письмо, чтобы ещё сильнее надавить на совет. Ультиматум был сформулирован предельно ясно: если Альтмана не восстановят в должности, а совет не уйдёт в отставку, все подписавшие немедленно уволятся и перейдут в Microsoft.

Письмо передавали по закрытым каналам. Тем, кого не было рядом, звонили и просили подписать. Когда подписей стало достаточно много, к письму начали присоединяться и остальные – теперь уже из опасения, что отсутствие их имени само по себе вызовет вопросы. Меньше чем за сутки письмо подписали больше 700 из примерно 770 сотрудников OpenAI. Десятки людей одно за другим отправляли совету одинаковые электронные письма. Сотни выкладывали в X одну и ту же фразу: «OpenAI – ничто без своих людей».

А потом, глубокой ночью, случилось то, чего никто не ждал. В списке подписавших появилось имя Ильи Суцкевера.

Вскоре он выступил и публично: «Я глубоко сожалею о своём участии в действиях совета директоров. Я никогда не хотел причинить вред OpenAI. Я

люблю всё, что мы создали вместе, и сделаю всё возможное, чтобы компания снова стала единой».

Во вторник, 21 ноября, на пятый день кризиса, руководители вели переговоры с советом из дома Альтмана. Люди валились с ног от недосыпа и усталости. Приближался День благодарения. И в конце концов отчаяние обеих сторон открыло дорогу к компромиссу. Весь день перебирали самые разные варианты. Альтман и Брокман, которые прежде требовали вернуть их в компанию вместе с местами в совете, согласились от этих мест отказаться. Совет, в свою очередь, признал очевидное: без Альтмана компанию уже не спасти – и согласился на его возвращение.

Поздно вечером договорились и о составе трёх независимых директоров. Адам Д'Анджело оставался. Хелен Тонер и Таша Макколи уходили. Их места занимали Брет Тейлор, бывший сопредседатель Salesforce и бывший технический директор Facebook, и Ларри Саммерс, бывший министр финансов США и бывший президент Гарвардского университета. Позже в новый совет должны были войти и другие люди. Альтман соглашался на внутреннее расследование. Но сейчас важнее всего было показать единство, стабильность, примирение.

Через два дня Альтман опубликует подчёркнуто миролюбивое сообщение: «только что провёл несколько отличных часов с @adamdangelo. счастливого дня благодарения от наших семей вашим». Ещё через десять дней Брокман выложит фотографию, на которой они с Суцкевером стоят в обнимку и широко улыбаются.

А в офисе штатный художник OpenAI подключит генератор изображений DALL-E к цветному принтеру и начнёт печатать крошечные калейдоскопические наклейки в форме сердечек. Рядом появится огромное розовое сердце с надписью «OpenAI – ничто без своих людей». В декабре Альтман скажет Тревору Ноа в подкасте, что эта неделя стала вторым худшим переживанием в его жизни. Хуже была только смерть отца.

В январе сделка по продаже акций сотрудников всё-таки состоится, и OpenAI оценят в 86 миллиардов долларов. Но всё это будет потом. А вечером во вторник, 21 ноября, люди просто праздновали. Как только объявили о возвращении Альтмана и о новой договорённости, сотрудники хлынули обратно в офис.

Обнимались, плакали, врубили музыку. Кто-то притащил дым-машину, и она так хорошо справилась со своей задачей, что сработала пожарная сигнализация. Это никого не остановило: праздник продолжался. Брокман поднял телефон и сделал групповое селфи. На снимке – толпа людей в почти безумной эйфории: измотанные, невыспавшиеся, счастливые оттого, что пережили катастрофу и каким-то чудом уцелели.

Он выложил фото с двумя словами: «мы вернулись».

* * *

Новость об изгнании Альтмана застала меня посреди интервью для этой книги. Звук на телефоне был выключен, и я ещё не подозревала, в какую безумную неделю меня сейчас затянет.

Минут через двадцать я взглянула на экран, чтобы узнать время, и увидела целую россыпь пропущенных уведомлений. Так начались несколько туманных, полных адреналина дней, когда я отчаянно пыталась понять, что происходит. В следующие недели друзья, родственники и журналисты раз за разом задавали мне один и тот же вопрос: что всё это значит? И значит ли вообще что-нибудь? Была ли эта череда увольнений, ультиматумов и возвращений всего лишь захватывающей драмой из жизни Кремниевой долины – или её последствия почувствуем и мы?

К тому времени я следила за OpenAI уже пять лет. В 2019 году я стала первым журналистом, получившим широкий доступ в компанию и написавшим о ней большой очерк. Поэтому для меня случившееся было не очередной мелкой схваткой за власть между богачами из Кремниевой долины. Эта драма высветила один из самых насущных вопросов нашего поколения: кто и как должен управлять искусственным интеллектом?

ИИ – одна из важнейших технологий нашего времени. Чуть больше чем за десять лет он перестроил сами основы интернета и стал вездесущим посредником почти во всём, что мы делаем в цифровом мире. А теперь, за ещё более короткий срок, он готовится перекроить множество ключевых общественных институтов: здравоохранение, образование, право, финансы, журналистику, государственное управление. Будущее искусственного интеллекта – то, какой станет эта технология, – неразрывно связано с нашим собственным будущим. Поэтому вопрос о том, кто будет управлять ИИ, на самом деле звучит иначе: как сделать так, чтобы будущее стало лучше, а не хуже?

С самого начала OpenAI подавала себя как смелый эксперимент, призванный ответить именно на этот вопрос. Компанию основала группа людей, среди которых были Илон Маск и Сэм Альтман, а деньги дали и другие миллиардеры, в том числе Питер Тиль. Она задумывалась как нечто большее, чем исследовательская лаборатория, и большее, чем обычная компания. Основатели объявили радикальную цель: создать так называемый общий искусственный интеллект, AGI, – по их представлениям, самую могущественную форму ИИ из всех, какие когда-либо существовали. И создать его не ради прибыли акционеров, а ради всего человечества. Именно поэтому Маск и Альтман оформили OpenAI как некоммерческую организацию и пообещали вложить в неё миллиард долларов. Никаких коммерческих продуктов – только исследования, только благие намерения, только такой AGI, который откроет человечеству путь к всеобщей утопии, а не к её противоположности.

Маск и Альтман обещали также как можно шире публиковать результаты исследований и сотрудничать с другими организациями. Если ваша цель – благо всего мира, рассуждали они, то открытость (отсюда и название OpenAI) и

демократическое участие в развитии технологии принципиально важны. Через несколько лет руководство пошло ещё дальше и даже публично пообещало при необходимости поступиться собственным лидерством. Руководители OpenAI писали, что боятся, как бы последний этап создания AGI не превратился в гонку, где ради скорости участники начнут пренебрегать безопасностью. Поэтому если какой-то другой проект подойдёт к созданию безопасного и полезного AGI раньше, OpenAI обещала перестать с ним соперничать и начать ему помогать. Такого почти никто никогда не обещал.

Но к тому времени, когда я впервые взялась всерьёз изучать OpenAI, от этих идеалов стремительно оставалось всё меньше. Всего через полтора года после основания организации её руководители поняли, что избранный ими путь развития ИИ потребует невероятных денег. До тех пор Маск и Альтман как сопредседатели держались в стороне от повседневного управления. Теперь оба захотели стать генеральным директором. Победил Альтман. В начале 2018 года Маск ушёл из OpenAI – и забрал свои деньги. Оглядываясь назад, трудно не увидеть в этом расколе первый серьёзный признак того, что OpenAI была не столько альтруистическим проектом, сколько проектом человеческих амбиций и самолюбия.

Потеря главного спонсора поставила OpenAI в тяжёлое финансовое положение. Чтобы выкрутиться, Альтман перестроил юридическую конструкцию организации. Внутри некоммерческой структуры он создал коммерческое подразделение, OpenAI LP. Теперь оно могло привлекать капитал, выпускать коммерческие продукты и приносить инвесторам прибыль – почти как обычная компания. Через четыре месяца, в июле 2019 года, OpenAI объявила о новом инвесторе, вложившем миллиард долларов. Им стала Microsoft.

Вскоре после этого, в августе 2019 года, я впервые пришла в офис OpenAI. Три дня я провела в компании, разговаривала с сотрудниками, записала десятки интервью – и увидела, как эксперимент с идеалистической системой управления начинает разваливаться. OpenAI становилась всё более соревновательной, всё более закрытой, всё более замкнутой на себе. Она словно начинала бояться внешнего мира и в то же время пьянела от сознания, что держит в руках технологию потенциально колоссальной силы.

Идеи прозрачности и демократии таяли, а с ними и разговоры о самопожертвовании и сотрудничестве. У руководителей OpenAI постепенно осталась одна навязчивая цель: первыми создать общий искусственный интеллект – и создать его по своему образу и подобию. За следующие четыре года OpenAI стала почти всем тем, чем когда-то обещала не становиться. Некоммерческой она оставалась главным образом на бумаге. Компания агрессивно зарабатывала на таких продуктах, как ChatGPT, и добивалась прежде немыслимых оценок.

Она становилась всё более скрытной. Не только закрыла доступ к собственным исследованиям, но и изменила правила игры для всей отрасли,

выведя значительную часть разработок ИИ из-под общественного контроля. Она сама запустила ту самую гонку на выбывание, от которой когда-то предостерегала. ИИ коммерциализировали и внедряли с невероятной скоростью – прежде чем удалось устранить его опасные недостатки и понять, как эта технология может усиливать и использовать уже существующие расколы в обществе. Тем временем конфликты между руководителями и сотрудниками становились всё ожесточённее. Внутри компании боролись за власть разные группировки, и каждая пыталась перестроить OpenAI по своим представлениям о будущем.

Увольнение Альтмана в ноябре 2023 года и его почти мгновенное возвращение окончательно доказали, что эксперимент с управлением провалился. Дело было не только в том, что некоммерческий совет отступил под давлением денег и тем самым разрушил последние остатки альтруистического фасада. Случившееся показало нечто гораздо более важное: будущее искусственного интеллекта сегодня решается в борьбе за власть внутри ничтожно малой группы элиты Кремниевой долины. Даже если бы события пошли иначе и совету удалось бы заменить Альтмана, главная проблема осталась бы прежней: решение огромной важности принималось за закрытыми дверями. За пределами крошечного круга сверхбогатых техно-оптимистов, их самых яростных идейных противников и многомиллиардного технологического гиганта даже сами сотрудники OpenAI почти ничего не знали о том, какая судьба им уготована.

Я начала писать об искусственном интеллекте задолго до того, как OpenAI и ChatGPT стали едва ли не синонимами слова «ИИ». Я видела, как технология развивается в беспорядочном процессе науки и изобретательства, через пробы и ошибки. Как исследователи проверяют новые идеи, как показывают самые эффектные результаты на переполненных конференциях, как эти разработки потом встраивают в продукты крупнейших компаний мира – Google, Facebook, Alibaba, Baidu. Я прочитала сотни научных статей и беседовала с учёными, инженерами и руководителями, пытаясь понять их взгляды и решения и увидеть, какой след эти взгляды оставляют на устройстве технологий и на том, как их применяют.*

По мере того как искусственный интеллект расходился по миру, я следила и за едва заметными, и за драматическими переменами, которые он приносил в жизнь отдельных людей и целых сообществ. Я побывала на пяти континентах. В Колумбии и Кении встречалась с людьми, которых экономический кризис заставил наняться в ИИ-индустрию разметчиками данных и которые оказались в условиях, пугающе похожих на долговую кабалу. В Аризоне и Чили говорила с местными политиками и активистами, встревоженными тем, что вокруг вырастают целые теневые мегаполисы дата-центров, выпивающие драгоценную воду, без которой не могут жить их дома и города.

За эти годы я поняла две вещи. Во-первых, искусственный интеллект – это не одна технология. Это множество разных технологий, которые постоянно меняют облик и развиваются, причём направление этого развития определяется отнюдь не только их техническими достоинствами. На него влияют идеология создателей, мода, ажиотаж, коммерческие интересы. Сегодня всё внимание приковано к ChatGPT, большим языковым моделям и генеративному ИИ. Но это лишь одно из возможных воплощений искусственного интеллекта – и воплощение, отражающее очень определённое и поразительно узкое представление о том, как устроен мир и каким он должен быть. Ничто не предрешало, что на первый план выйдет именно такой ИИ и что он вообще появится. Он стал итогом тысяч субъективных решений, которые приняли люди, допущенные в комнату, где эти решения принимаются. И будущие поколения ИИ так же ничем не предопределены. А значит, мы снова возвращаемся к главному вопросу: кому будет позволено их формировать?

Второе, что я поняла, тревожит куда больше. Нынешний облик искусственного интеллекта и направление, в котором он развивается, ведут нас по опасному пути.

На первый взгляд генеративный ИИ восхищает. Он за секунды поможет придумать идею или написать текст. Может стать собеседником глубокой ночью, когда хочется хоть с кем-то поговорить. Быть может, когда-нибудь он и правда так поднимет производительность труда, что даст мощный толчок экономике. Но когда-то и Facebook казался нам просто местом, куда выкладывают фотографии из отпуска, где находят давно потерянных школьных друзей и откуда начинаются добрые общественные движения. За гладкой, завораживающей оболочкой генеративного ИИ скрывается нечто куда более сложное. Под капотом эти модели – настоящие чудовища масштаба. На их создание уходят такие объёмы данных, человеческого труда, вычислительных мощностей и природных ресурсов, какие ещё недавно трудно было вообразить. По одной из оценок, GPT-4 – преемница модели, на которой работал первый ChatGPT, – больше GPT-1, выпущенной всего пятью годами раньше, в пятнадцать с лишним тысяч раз.

Человеческую и материальную цену этого взрывного роста оплачивают огромные слои общества, и прежде всего самые уязвимые: люди, которых я встречала по всему миру, – рабочие и жители сельских районов богатого Глобального Севера, беднейшие сообщества Глобального Юга. Все они сталкиваются с новыми видами незащищённости. А обещанные блага «технологической революции» почти не доходят до низов: выгоды генеративного ИИ в основном утекают наверх. За все эти годы я нашла лишь одну метафору, которая, по-моему, действительно передаёт суть нынешних властителей искусственного интеллекта.

Это империи. В эпоху европейского колониализма империи захватывали и выкачивали ресурсы, которые им не принадлежали. Они эксплуатировали труд

покорённых народов, заставляя их добывать, выращивать и обрабатывать эти ресурсы ради обогащения метрополии. Они насаждали расистские, расчеловечивающие идеи о собственном превосходстве и прогрессивности, чтобы оправдать отнятие суверенитета, грабёж и подчинение, а порой и убедить покорённых принять всё это добровольно. Погоню за могуществом оправдывали необходимостью соперничать с другими империями, а в гонке вооружений, как известно, правил нет. И всё это в конечном счёте служило одной цели: укреплять власть империи, расширять её границы, обеспечивать её дальнейший рост. Проще говоря, империи накапливали несметные богатства, раскинувшись в пространстве и во времени и навязывая миру колониальный порядок, за который все остальные платили огромную цену.

Современные империи ИИ, разумеется, не прибегают к той откровенной жестокости и тому насилию, которыми отмечена колониальная эпоха. Но и они захватывают и выкачивают драгоценные ресурсы, без которых не воплотить их представление об искусственном интеллекте: труд художников и писателей; данные бесчисленных людей, рассказывающих в интернете о своей жизни и наблюдениях; землю, электричество и воду для гигантских дата-центров и суперкомпьютеров. Новые империи точно так же эксплуатируют труд людей по всему миру – тех, кто чистит, сортирует, размечает и готовит данные, превращаемые затем в прибыльные ИИ-технологии. Они рисуют соблазнительные картины будущего, рассуждают о прогрессе и настойчиво внушают нам, что обязаны победить империи-соперницы. Так оправдываются вторжение в частную жизнь, присвоение чужого труда и массовая автоматизация, способная уничтожить огромные пласты осмысленной экономической деятельности.

Сегодня именно OpenAI ведёт нас к этому новому колониальному порядку. В погоне за расплывчатым идеалом «прогресса» компания сделала бесконечное наращивание масштаба главным законом новой эпохи ИИ. Теперь технологические гиганты состязаются в том, кто станет ещё больше. Они тратят такие астрономические суммы, что вынуждены перекраивать и стягивать в одно место собственные ресурсы. Примерно тогда же, когда Microsoft вложила в OpenAI 10 миллиардов долларов, она ради экономии уволила десять тысяч сотрудников. Увидев, что OpenAI её обгоняет, Google слила свои ИИ-лаборатории в Google DeepMind. Когда Baidu бросилась делать собственный аналог ChatGPT, тем, кто искал с помощью ИИ новые лекарства, пришлось приостановить работу и отдать свои вычислительные чипы разработчикам чат-бота. Нынешняя модель развития ИИ перекрывает кислород альтернативным направлениям. Независимых исследователей, не связанных с технологическими компаниями и не живущих на их деньги, становится всё меньше. А вместе с ними исчезает и разнообразие идей, которые нельзя сразу обратить в коммерческую выгоду.

Даже сами компании, когда-то финансировавшие широкие и свободные фундаментальные исследования, теперь всё чаще не могут себе их позволить: счёт за генеративный ИИ слишком велик. Новое поколение учёных вынуждено подстраиваться под сложившийся порядок, чтобы не остаться без работы. То, что ещё недавно казалось небывалым, стало нормой. Сегодня эти империи богаче, чем когда-либо.

Когда в январе 2025 года я заканчивала эту книгу, OpenAI оценивалась уже в 157 миллиардов долларов. Её конкурент Anthropic готовился к сделке с оценкой в 60 миллиардов. После партнёрства с OpenAI рыночная капитализация Microsoft выросла втрое и перевалила за 3 триллиона долларов. С появлением ChatGPT совокупная рыночная стоимость шести крупнейших технологических гигантов прибавила около 8 триллионов долларов.

И одновременно растут сомнения: так ли уж велика реальная экономическая польза генеративного ИИ? В июне 2024 года Goldman Sachs подсчитал, что за несколько лет расходы на эту технологию могут дойти до триллиона долларов, хотя предъявить в оправдание этих затрат пока, по выражению авторов отчёта, удивительно нечего. Месяцем позже Upwork Research Institute опросил 2500 работников по всему миру. 96 процентов руководителей высшего звена не сомневались, что генеративный ИИ повысит производительность. Но 77 процентов сотрудников, реально пользовавшихся этими инструментами, сказали обратное: ИИ прибавил им работы. Отчасти потому, что созданное им приходилось проверять, отчасти потому, что начальство, вооружившись ИИ, стало требовать от людей ещё больше. В ноябре журналисты Bloomberg Парми Олсон и Кэролайн Сильверман, разобрав финансовые итоги генеративного ИИ, предельно ясно сформулировали неприятную возможность: технология, которую объявили революционной, может так и не принести обещанного масштабного преобразования экономики, а лишь ещё сильнее сосредоточит богатство наверху.

Тем временем остальной мир уже начинает прогибаться под растущей человеческой и материальной ценой новой эпохи. В Кении людям платили гроши за то, чтобы они вычищали сцены насилия и язык ненависти из систем OpenAI, включая ChatGPT. Художников вытесняют модели, обученные на их же работах – без разрешения и без оплаты. Журналистика слабеет, пока генеративные технологии штампуют всё больше дезинформации. У нас на глазах повторяется очень старая история.

И это только начало. OpenAI не собирается сбавлять скорость. Компания по-прежнему стремится ко всё большему масштабу, располагая ресурсами, с которыми мало кто может тягаться, а за ней тянется вся отрасль. Чтобы заглушить растущую тревогу по поводу того, что генеративный ИИ способен дать человечеству сегодня, Альтман всё громче говорит о фантастических благах AGI завтра.

В сентябре 2024 года он объявил в своём блоге, что скоро наступит «эпоха интеллекта», эпоха «всеобщего процветания», а сверхинтеллект, по его предположению, может появиться уже через «несколько тысяч дней». Будущее, писал Альтман, будет настолько светлым, что никакие сегодняшние слова не воздадут ему должного. Невероятные достижения постепенно станут обыденностью: мы исправим климат, заселим космос, откроем все законы физики. Но к этому времени AGI во многом стал риторической фигурой – универсальным фантастическим оправданием, позволяющим OpenAI добиваться всё большего богатства и могущества. Почти ни у кого больше нет сопоставимых денег, чтобы финансировать иные пути. OpenAI и несколько её крупнейших соперников получают фактическую олигополию на технологию, которую сами же продают нам как ключ к будущему. И любому – компании или государству, – кто захочет получить кусочек этого будущего, придётся идти на поклон к империям, которые держат ключи от входа.

Но есть и другой путь. Искусственный интеллект не обязан быть таким, каким стал сегодня. Нам вовсе не нужно соглашаться с тем, что прогресс требует бесконечного роста масштабов и потребления ресурсов. Очень многое из того, что действительно нужно обществу, – лучшая медицина, лучшее образование, чистый воздух и вода, быстрый отказ от ископаемого топлива – достижимо с помощью гораздо меньших моделей ИИ и самых разных альтернативных подходов. Да и одного ИИ в любом случае будет мало.

Нам понадобится больше общественной солидарности и больше международного сотрудничества – как раз того, что нынешнее развитие ИИ подтачивает. Но империи ИИ не уступят власть добровольно. Нам самим придётся вернуть себе право определять будущее этой технологии. И сейчас настал решающий момент, когда это ещё возможно. Когда-то старые империи уступили место более открытым формам правления. Так и мы можем сообща определить будущее искусственного интеллекта. Политики могут ввести строгие правила о защите персональных данных и прозрачности, обновить законы об интеллектуальной собственности и вернуть людям власть над их данными и плодами их труда. Правозащитные организации могут добиваться международных трудовых норм, которые гарантируют разметчикам данных минимальную оплату и достойные условия и укрепят права работников во всех отраслях. Научные фонды могут вернуть исследованиям ИИ разнообразие и поддержать принципиально новые представления о том, какой вообще может быть эта технология. И наконец, каждый из нас способен не поддаваться историям, которые рассказывают нам OpenAI и вся ИИ-индустрия, чтобы спрятать растущую социальную и экологическую цену технологии за туманными обещаниями прогресса.

Примечание

* О том, как в области ИИ публикуют научные работы и как их рецензируют. Исследователи искусственного интеллекта часто выкладывают свои статьи прямо в открытый бесплатный репозиторий arXiv (произносится как «архив»), а рецензирование на конференции или в журнале проходят лишь через много месяцев или даже лет – или не проходят вовсе. Эта практика настолько прижилась, что многие в отрасли цитируют работы по их влиянию, а не по тому, прошли ли они рецензирование. В этой книге я поступаю так же. В примечаниях в конце книги отмечено, какие работы не рецензировались и являются препринтами.

Часть I

Глава 1

Божественное право

Все уже собрались, не было только Илона Маска – он, как обычно, опаздывал.

Стояло лето 2015 года. По приглашению Сэма Альтмана несколько человек собрались на закрытый ужин, чтобы поговорить о будущем искусственного интеллекта и человечества.

Маск познакомился с Альтманом, который был на четырнадцать лет моложе, незадолго до того и остался о нём хорошего мнения. Альтман, президент знаменитого кремниевого акселератора стартапов Y Combinator, был уже человеком с репутацией. Свою первую компанию он основал в девятнадцать лет и быстро прослыл в Кремниевой долине блестящим стратегом и мастером сделок с грандиозными амбициями – даже по меркам края, где основатели привыкли мыслить с размахом. Маску он показался умным, целеустремлённым и, что важнее всего, единомышленником: Альтман тоже считал, что развивать искусственный интеллект и управлять им нужно с величайшей осторожностью.

Много лет спустя Маск напишет в судебном иске, что Альтман словно отзеркалил всё, что Маск когда-либо говорил на эту тему, чтобы завоевать его доверие. Сам Альтман не раз признавался, что Маск был героем его детства. А после того как старший предприниматель лично провёл его по огромному заводу SpaceX в Хоторне, штат Калифорния, восхищение только окрепло. «Больше всего мне запомнилось выражение абсолютной уверенности на его лице, когда он говорил об отправке больших ракет на Марс, – писал потом Альтман. – Я ушёл с мыслью: „Ага, так вот как выглядит настоящая убеждённость“».

Маска давно и всерьёз тревожил искусственный интеллект. В 2012 году он познакомился с Демисом Хассабисом – учёным по складу ума и генеральным директором лондонской лаборатории DeepMind Technologies. Вскоре Хассабис приехал к Маску на завод SpaceX. Они сидели в столовой, а вокруг с грохотом перевозили и собирали огромные детали ракет. Там Хассабис и заговорил о том, что однажды может появиться ИИ, превосходящий человеческий разум. Такая система, сказал он, вполне способна стать угрозой для человечества. И любимый запасной план Маска – если всё пойдёт совсем плохо, переселиться на Марс – тут не поможет. Сверхразум, усмехнулся Хассабис, просто последует за людьми в космос. Маску было не до смеха. Он вложил в DeepMind пять миллионов долларов – в том числе для того, чтобы следить за тем, как развивается компания.

Через год, на праздновании своего дня рождения в винодельческой долине Напа, Маск сцепился в жарком споре со старым другом, сооснователем Google Ларри Пейджем. Спорили о простом: опасен ли ИИ, который превзойдёт

человека? Пейдж не видел в этом трагедии: для него такой разум был просто следующей ступенью эволюции. Когда Маск возразил, Пейдж обвинил его в «видовом шовинизме» – в том, что он ставит людей выше нечеловеческих форм разума.

После этого Маск почти без умолку заговорил об экзистенциальной опасности ИИ. На симпозиуме в MIT он назвал искусственный интеллект, возможно, «главной угрозой самому существованию» человечества, а его разработку сравнил с «вызыванием демона». Он встречался с нью-йоркскими издателями, подумывая о собственной книге о рисках, способных стереть человечество с лица земли, включая угрозу ИИ. Позже на встрече Stanford AI Salon после его выступления к нему подошла молодая исследовательница Тимнит Гебру.

– Почему вас так волнует искусственный интеллект, когда изменение климата куда более очевидно угрожает существованию человечества? – спросила она.

– Изменение климата – это плохо, но оно не убьёт всех, – ответил Маск. – А ИИ может уничтожить человечество целиком.

В конце 2013 года Маск узнал, что Google собирается купить DeepMind, и был уверен, что добром это не кончится. Он публично предупреждал: если Google поручит гипотетическому AGI максимизировать прибыль, система может решить, что лучший способ – любой ценой избавиться от конкурентов. Начать с убийства всех исследователей ИИ, работающих у соперников, язвительно заметил Маск в разговоре с The New Yorker, было бы «не самой симпатичной чертой характера». На одной вечеринке в Лос-Анджелесе он на целый час заперся в гардеробной на втором этаже и по Skype уговаривал Хассабиса отказаться от сделки с Google.

– Будущее ИИ не должно оказаться в руках Ларри, – говорил он.

Маск не знал, что дело зашло уже очень далеко. Google отправила группу своих исследователей ИИ на частном самолёте в офис DeepMind для технической экспертизы. Один из старейших и самых влиятельных специалистов Google, Джефф Дин, лично просмотрел часть исходного кода DeepMind и дал сделке зелёный свет. В январе 2014 года Google официально объявила о покупке. По сообщениям, сумма составила от 400 до 650 миллионов долларов.

После этого Маск стал устраивать собственные ужины, на которых обсуждали, как противостоять Google. В начале 2015 года он даже встретился с президентом Бараком Обамой, чтобы рассказать об опасностях ИИ, о том, как сделать его безопаснее, и о необходимости государственного регулирования. Примерно тогда же Маск снова увиделся с Хассабисом в SpaceX – на первом заседании совета по этике Google DeepMind. Этот совет Ларри Пейдж и Демис Хассабис создали, чтобы следить за ответственным развитием технологий компании. Маск ушёл с заседания убеждённым, что совет – фикция. Его тревога переросла почти в навязчивую идею: Хассабиса нужно остановить.

Годами Маск изображал Хассабиса чуть ли не суперзлодеем. В его картине мира OpenAI должна была стать добром, противостоящим злу DeepMind. Летом 2016 года, вскоре после основания OpenAI, несколько её сотрудников встретились с Хассабисом и, вернувшись, сказали коллегам примерно следующее: да, DeepMind и правда собирается захватить мир; похоже, Маск был прав. Через год Маск устроил выездную встречу сотрудников OpenAI на заводе SpaceX и снова разразился тирадой о Хассабисе, который до DeepMind семь лет руководил собственной студией видеоигр.

– Он буквально сделал игру, где злой гений создаёт ИИ, чтобы захватить мир! – почти кричал Маск. – И люди, чёрт возьми, этого не понимают! Не понимают! А Ларри? Ларри думает, что контролирует Демиса, а сам слишком занят своим долбаным виндсёрфингом, чтобы заметить, как Демис прибирает к рукам всю власть!

В DeepMind эта паранойя Маска со временем стала поводом для шуток. Да, Хассабис был чрезвычайно честолюбив и порой очень напорист. Но при этом люди описывали его как спокойного и доброжелательного человека.

– Создание OpenAI выглядело какой-то полуистеричной реакцией на довольно мягкого, спокойного человека, – вспоминал один бывший исследователь DeepMind. – Это было немного абсурдно.

Среди книг, которые особенно горячо рекомендовал Маск, был «Сверхинтеллект: пути, опасности, стратегии» оксфордского философа Ника Бострома. Главная мысль книги тревожила: если ИИ станет умнее человека, держать его под контролем может оказаться чрезвычайно трудно, а последствия – катастрофическими. Бостром приводил ставший знаменитым мысленный эксперимент. Допустим, сверхразумной системе поручили простейшую задачу – производить скрепки. Теоретически она может решить, что люди ей мешают, потому что потребляют ресурсы, из которых можно было бы наделать ещё больше скрепок. Отсюда главный вопрос: как добиться, чтобы сверхразум не исполнял наши указания буквально, круша всё вокруг? Бостром предлагал идею «согласования» (alignment): ИИ должен разделять человеческие ценности и понимать не только буквальную команду, но и более широкие намерения человека.

Из таких идей позже выросло целое направление – исследования безопасности ИИ, которое OpenAI сделает одной из своих главных тем. Своей огромной аудитории в Twitter Маск рекомендовал книгу коротко: «Стоит прочитать». Однако в январе 2023 года всплыло старое письмо Бострома середины 1990-х, после которого многие по-новому взглянули уже на его собственные ценности. В нём он употребил откровенно расистское выражение и утверждал, что считает сказанное правдой. Позже Бостром извинился, назвал письмо отвратительным и заявил, что оно не отражает его нынешних взглядов.

Альтман казался Маску человеком того же склада. Его тоже занимали мысли о катастрофах и о том, как к ним готовиться. В 2016 году он рассказал

журналисту The New Yorker Тэду Френду, что на случай конца света у него есть план: бежать в Новую Зеландию вместе с близким другом и наставником, миллиардером Питером Тилем. В той же статье Тиль назвал Альтмана человеком «в культурном отношении очень еврейским – оптимистом и в то же время выживальщиком, который всегда помнит, что всё может пойти ужасно плохо». Через два года Альтман скажет Bloomberg, что тогда пошутил. Но тревожный рюкзак у него действительно был – особенно он боялся новых вирусов – с противогазами, антибиотиками, водой, батарейками, палаткой и оружием. А в феврале 2015 года Альтман написал в блоге, что согласен с Маском: сверхразум – это, вероятно, самая большая угроза дальнейшему существованию человечества. Рукотворный вирус, по его мнению, был более вероятным сценарием, но даже он вряд ли уничтожит всех людей во Вселенной. В скобках Альтман снова советовал книгу Бострома как лучшее, что он читал на эту тему.

Через несколько месяцев, в мае 2015 года, Альтман написал Маску. Он размышлял, можно ли вообще помешать человечеству создать искусственный интеллект, и сам себе отвечал: почти наверняка нет. А раз ИИ всё равно появится, неплохо было бы, чтобы первым его создал кто-нибудь, кроме Google. Он предложил запустить в Y Combinator нечто вроде «Манхэттенского проекта для ИИ» – но с принципиальным отличием: технология должна принадлежать не частной компании, а всему миру, через некоммерческую организацию. Альтман подчёркивал также, что проект будет неукоснительно соблюдать государственное регулирование и даже активно его поддерживать. Маск ответил коротко:

– Обсудить определённо стоит.

В июне Альтман прислал новое письмо, уже с подробностями. Миссия, писал он, должна состоять в том, чтобы создать первый общий искусственный интеллект и использовать его для расширения возможностей каждого человека. Самым безопасным будущим ему казалось не централизованное, а распределённое. И главное: безопасность должна с самого начала быть основополагающим требованием. Дальше Альтман предлагал схему управления. В совет войдут он, Маск и ещё три человека. Технология будет принадлежать фонду и использоваться «на благо мира». А если станет неясно, что именно считать «благом мира», решат эти пятеро. Уже в этом раннем замысле заключался примечательный парадокс: технология должна была принадлежать всему человечеству, но решать, что для человечества хорошо, собирались пять человек.

Альтман просил Маска по возможности встречаться с командой примерно раз в месяц: его участие, объяснял он, поможет привлечь лучших специалистов. Если же времени не найдётся, хватит и публичной поддержки – одно его имя сильно облегчит набор людей.

– Согласен со всем, – коротко ответил Маск.

Тогда Альтман и пригласил Маска на тот самый закрытый ужин о будущем ИИ и человечества, чтобы познакомить его с группой ведущих инженеров и исследователей, которых надеялся заполучить в новый проект. Как только Маск подтвердил участие, ужин перенесли в место попрестижнее – в любимый Маском Rosewood Hotel, роскошный комплекс на шестнадцати акрах, где номер стоил около тысячи долларов за ночь. Отель стоял на живописной Сэнд-Хилл-роуд, утопающей в зелени и сплошь застроенной офисами венчурных фондов, – по сути, одной из главных артерий Кремниевой долины. Отдельный зал выходил на балкон с видом на бассейн в обрамлении итальянских кипарисов и садовых роз. Маск явился с опозданием больше чем на час. Остальные давно его ждали. Среди гостей были Сэм Альтман, Грег Брокман, Дарио Амодеи и Илья Суцкевер – вскоре именно они станут ключевыми руководителями новой некоммерческой организации. Маск придумает ей название, которое должно было выразить сам дух их миссии: OpenAI, «открытый искусственный интеллект».

Ирония в том, что со временем почти все, кто сидел за тем столом, уйдут из организации, рассорившись с Альтманом и его представлениями о будущем ИИ. Когда отношения Маска и Альтмана окончательно испортятся, а сам Альтман станет главной звездой Кремниевой долины, изменится и публичная версия его взглядов.

– Сейчас я твёрдо стою на том, что ИИ будет инструментом, – скажет он в 2023 году Business Insider. – Хотя люди будущего и само общество станут совсем другими, и у нас есть возможность всерьёз подумать о том, каким это будущее спроектировать.

Маск же со временем убедится, что Альтман просто использовал его, чтобы взобраться на вершину.

Впрочем, такое о нём говорили почти всю жизнь. Его наставник Пол Грэм однажды сказал знаменитую фразу: «Можно сбросить Сэма с парашютом на остров людоедов, вернуться через пять лет – и он будет там королём».

Позже Грэм выразился ещё прямее: «Сэм невероятно хорошо умеет становиться влиятельным».

Сэм Альтман

Сэмюэл Харрис Гибстайн Альтман родился 22 апреля 1985 года в Чикаго. Он был первенцем в еврейской семье. Его мать, Конни Гибстайн, была врачом, как и её отец Марвин, педиатр, после Второй мировой войны служивший военным врачом армии США в Германии. Конни получила сразу два образования, медицинское и юридическое, что для женщины её поколения было большой редкостью. В конце концов она выбрала дерматологию: стабильный заработок и довольно гибкий график позволяли ей возвращаться домой, готовить ужин и быть с детьми.

Ещё на юридическом факультете Университета Лойолы в Чикаго она познакомилась с Джеролдом Альтманом, красивым однокурсником тремя годами старше. Джерри был из семьи обувного фабриканта. До встречи с Конни он успел поучиться в Уортонской школе бизнеса при Пенсильванском университете, поработать консультантом в Бостоне и один раз жениться. Выйдя замуж, Конни оставила свою фамилию. Через несколько лет семья переехала из Чикаго в Сент-Луис, родной город обоих супругов.

Джерри занялся недвижимостью и управлением объектами. Какое-то время он был главным юрисконсультом и вице-президентом девелоперской компании Roberts Companies. Он любил людей, увлекался темой доступного жилья и работал над коммерческими и жилыми проектами, которые должны были оживить кварталы Сент-Луиса и сплотить их жителей. Позже Сэм часто повторял один из главных уроков отца: «Людям нужно помогать всегда. Даже если кажется, что времени нет совсем, ты всё равно находишь способ».

У Конни и Джерри один за другим родились трое сыновей – Сэм, Макс и Джек, а через пять лет после Джека и через девять после Сэма появилась Энни. Конни была счастлива: наконец-то дочь. Сама Конни называла себя атеисткой, хотя в культурном смысле ощущала себя еврейкой. Джерри был набожнее: ходил на службы в главные еврейские праздники и настоял, чтобы все четверо детей прошли бар- или бат-мицву. Рациональность и дисциплина матери, духовность отца и его вера в то, что нужно служить людям, по-разному проявятся в каждом из детей.

Сэм с малых лет отличался редкой целеустремлённостью и почти ненасытным любопытством. В два года он научился включать семейный видеомагнитофон, в три – пытался его чинить. Когда через пять лет родители подарили ему компьютер Mac, он быстро освоил программирование и начал разбирать машину на части. Роль старшего брата ему тоже понравилась: он то командовал младшими, то опекал их и почти всегда хотел выиграть. Настольные игры превращались в состязание, проигрывать в котором Сэм наотрез отказывался.

Поражения он переносил тяжело, зато особенно любил вкус победы. Однажды бабушка подарила каждому внуку немного акций. Сэм выбрал Apple, Джек – сеть ресторанов Applebee's. Спустя годы это стало семейной шуткой: акции Джека почти не изменились в цене, а акции Сэма взлетели.

– Твои Apple выросли так, что я даже думать об этом не хочу, – вспоминал Джек. – В сотни и сотни раз.

– Да, неплохо выросли, – с явным удовольствием отвечал Сэм.

Когда Сэм подрос, мать предложила ему перейти в частную школу John Burroughs, славившуюся сильной программой и впечатляющим списком знаменитых выпускников. Сэм согласился. Макс тоже попробовал, но потом ушёл; Джек отказался; Энни позже пошла по стопам старшего брата. В Burroughs Сэм чувствовал себя как рыба в воде. Он хорошо учился, был общительным,

уверенным в себе, любил подурачиться и посмешить окружающих. Его занимали не только математика, естественные науки и технологии: он хорошо писал и участвовал во множестве внеклассных занятий – руководил выпуском школьного ежегодника, был капитаном команды по водному поло, ездил на «Модель ООН», где школьники изображают работу Организации Объединённых Наций и спорят о мировой политике.

– Я тогда думал, и теперь мне даже немного неловко в этом признаваться: «Только бы он не ушёл в технологии. Он такой творческий, так хорошо пишет», – вспоминал его учитель английского Энди Эбботт, впоследствии ставший директором школы. – Мне хотелось, чтобы он стал писателем или кем-то в этом роде.

Но уже тогда у Альтмана проявился другой дар – умение вести за собой. И ему нравилось испытывать границы дозволенного. В довольно консервативной школе он однажды нажил неприятности, когда на ежегодном школьном вечере устроил со своей командой по водному поло стриптиз – до плавок Speedo. В те же годы он открыто сказал родителям и одноклассникам, что он гей. Мать удивилась, но приняла это, как и вся семья. А вот группа учеников-христиан отреагировала иначе: в Национальный день каминг-аута они бойкотировали школьное собрание о сексуальности, которое вёл Альтман. Сэму было семнадцать, и он решил ответить им публично. Он выступил перед всей школой с речью, которая, по словам его консультанта по поступлению в колледж, заметно изменила атмосферу в школе. Последнюю фразу Альтман потом вспоминал так: «Либо у вас есть терпимость, без которой не бывает открытого сообщества, либо её нет. Нельзя выбирать, к кому вы будете терпимы, а к кому нет».

Но за уверенной внешностью скрывался гораздо более ранимый человек. Сэма очень заботило, что о нём думают другие, и его часто одолевала тревога. Эта черта останется с ним на всю жизнь. Позже, когда в Кремниевой долине о нём заговорят всё громче, он будет иногда звонить матери из-за обычной головной боли, успев убедить себя, что у него менингит или лимфома. А однажды посреди переговоров запаникует так, что ляжет на пол – без рубашки, раскинув руки, – чтобы прийти в себя.

Именно сочетание этих двух черт, необычайного честолюбия и глубокой ранимости, во многом определит его путь. Журналист The New Yorker Тэд Френд, проведший с Альтманом много часов, когда готовил о нём очерк 2016 года, заметил эту двойственность. Почти в любом деле Альтман одновременно рвался вперёд как можно быстрее и остро чувствовал опасность. Его логика могла звучать так: создать AGI как можно скорее – и только не уничтожить при этом человечество.

Окончив школу в 2003 году, Альтман уехал со Среднего Запада в Стэнфорд – прежде всего потому, что университет находился в самом сердце технологической индустрии. Правда, становиться технарём он поначалу не

собирался. Как и надеялся его учитель, Альтман всерьёз подумывал о писательстве, а совсем недолго – даже о карьере инвестиционного банкира. Но победила давняя страсть к компьютерам и программированию. «Я понял, что миру не нужен семимиллионный роман, – объяснял он позже. – Не там я мог принести больше всего пользы. К тому же в таких областях обычно гораздо труднее заработать много денег, а иногда и просто достаточно».

В Стэнфорде Альтман изучал информатику и особенно увлекался искусственным интеллектом и компьютерной безопасностью. В задания он погружался с головой. Один однокурсник вспоминал, как Альтман разобрал до низкоуровневого кода программу, которой нужно было пользоваться для домашней работы, и нашёл ошибку в самом задании. На втором курсе его заинтересовали мобильные технологии. Узнав, что скоро почти в каждом телефоне будет GPS, он пришёл на студенческую встречу для начинающих предпринимателей, вышел на сцену с раскладушкой в руке и предложил всем желающим вместе сделать сервис на основе геолокации.

Примерно тогда же он познакомился с Полом Грэмом – предпринимателем и влиятельным блогером, писавшим о технологиях, который вместе со своей подругой Джессикой Ливингстон затевал новый инкубатор стартапов, Y Combinator. В 2005 году Альтман попал в самый первый набор YC со своим новым стартапом Loopt. Лето он провёл в Кембридже, штат Массачусетс, где поначалу работал акселератор. Loopt был социальной сетью на основе геолокации: приложение сообщало, если поблизости оказывался друг, или советовало рестораны неподалёку. Тем летом Альтман так много работал и так плохо питался – в основном лапшой быстрого приготовления, – что заработал цингу.

Но ни о чём не жалел. Позже он будет советовать молодым предпринимателям: «В начале карьеры работайте очень много. Потом это окупается, как сложные проценты». В Стэнфорд Альтман так и не вернулся. К концу 2005 года он с партнёрами вёл переговоры с венчурными фондами New Enterprise Associates и Sequoia о пяти миллионах долларов инвестиций. Альтман решил рискнуть и бросил университет.

Большим успехом Loopt так и не стал. В 2012 году, через семь лет работы, Альтман продал компанию за 43,4 миллиона долларов – примерно за столько же, сколько в неё успели вложить инвесторы. Но, послушав в те годы интервью Альтмана и его инвесторов, можно было подумать, что Loopt стоит на пороге великой технологической революции. И именно здесь особенно хорошо видно, на чём потом вырос успех Альтмана.

В ранних интервью он продавал нечто куда менее впечатляющее, чем искусственный интеллект, – по сути, ещё один геолокационный сервис, конкурента Foursquare, который так и не взлетел. Но даже это Альтман умел превратить в историю о будущем.

Его умение работать с прессой и талант заключать сделки – два главных инструмента будущего восхождения – держались на одном редком качестве: он превосходно рассказывал истории, и получалось это у него совершенно естественно. Даже сегодня, глядя старые интервью и зная, что Loopt прогорит, трудно не увлечься его словами. Альтман спокойно и уверенно объясняет, почему его компания занимает на рынке особое место, почему она – часть большого, неудержимого движения технологий вперёд, почему пользователям нужен именно этот сервис, почему рекламодатели вот-вот начнут за него драться. А главное – не ставьте против него: успех всё равно неизбежен.

– Реакция была потрясающей, – рассказывал Альтман в июне 2010 года технологическому блогеру Роберту Скоблу о новом приложении Loopt Star.

Приложение позволяло рекламодателям присылать людям специальные предложения в зависимости от того, где они находятся, например скидки в ближайших ресторанах или магазинах. Общество, уверял Альтман, уже перешагнуло важнейший психологический рубеж:

– Теперь польза от того, что я делюсь своим местоположением, перевешивает опасения за частную жизнь. Через несколько лет делиться геолокацией станет нормой. А странным будет как раз тот, кто этого не делает.

Через несколько месяцев в интервью CNN Business он и вовсе назвал разделение жизни на онлайн и «реальную» бессмысленным: мобильные устройства, по его мнению, уже стирали эту границу.

– Весь мир становится мобильным. И весь мир идёт к тому, что ваши данные и ваши сервисы будут доступны вам везде, где бы вы ни были.

Даже разговор об опасностях своего продукта Альтман умел обратить в рекламу. В 2008 году Джессика Лессин, тогда журналистка The Wall Street Journal, сказала ему, что пишет статью об угрозах частной жизни, которые несёт отслеживание местоположения. Альтман не стал защищаться – он предложил помочь и прислал ей длинный список рисков, которые команда Loopt уже выявила сама, вместе с предложениями, как их устранить. Подтекст был почти очевиден: будущее всё равно будет устроено именно так, так что лучше заранее научиться с ним жить. «Он хотел не просто создать стартап, – так позже сформулировала своё впечатление Лессин. – Он хотел писать правила».

Как строится влияние

Именно в годы Loopt Альтман обзавёлся связями и отточил навыки, которые потом станут его главным капиталом. В середине 2000-х – начале 2010-х он оказался в самом эпицентре взрывного роста Кремниевой долины. Новые стартапы появлялись один за другим, вокруг были честолюбивые люди, которые не собирались ждать разрешения, чтобы изменить мир, и Альтман постоянно вращался среди них. Когда Loopt только начинал работу, его офис был буквально по соседству с молодым стартапом YouTube. В первом наборе Y Combinator

вместе с Альтманом учились Стив Хаффман и Крис Слоу – соответственно сооснователь и один из первых инженеров Reddit. В 2014 году Альтман войдёт в совет директоров Reddit и со временем будет владеть даже большей долей компании, чем Хаффман. Ещё одним участником того же набора был Эмметт Шир, сооснователь Twitch, – почти двадцать лет спустя именно он ненадолго станет временным гендиректором OpenAI после увольнения Альтмана. Так самым неожиданным образом замкнётся круг.

Альтман учился одновременно двум вещам: как подавать свою историю прессе и как заключать сделки, которые кажутся невозможными. Ещё будучи гендиректором никому не известного стартапа, он сумел договориться о партнёрстве с крупнейшими американскими операторами мобильной связи. Те, кто с ним работал, говорили, что его формула складывалась из нескольких качеств. Он великолепно умел слушать, всегда был готов помочь и умел подать своё предложение именно в тех понятиях, которые важны собеседнику. Не просто «вот что у меня есть», а «чего хотите вы?» – и показать, как его предложение может это дать. Позже, когда Альтман стал одним из самых богатых и влиятельных людей в технологической индустрии, этот метод заработал ещё лучше: теперь ему действительно было что предложить. О нём говорили: «Майкл Джордан по части умения слушать». А Джефф Ралстон, впоследствии возглавивший Y Combinator, называл его «Усэйном Болтом в привлечении инвестиций».

– Когда вы привлекаете деньги, вы, по сути, рассказываете человеку историю о будущем своего проекта, – объяснял Ралстон. – Историю, в которой этот проект или эта компания добивается невероятного успеха. Сэм умеет рассказать такую историю, что вам хочется в ней участвовать. Она захватывает. Она кажется реальной. Больше того – вероятной.

Ралстон сравнивал этот талант со знаменитым «полем искажения реальности» Стива Джобса. Джобс так убедительно описывал будущее, что его версия мира вытесняла все остальные. И это была не просто иллюзия: Джобс действительно создавал вещи, которые потом меняли реальность.

– И Сэм, очевидно, тоже, – добавлял Ралстон.

Но у этого дара была и оборотная сторона.

– Сэм запоминает о вас мельчайшие подробности. Он невероятно внимателен, – рассказывал человек, много лет с ним проработавший. – Но отчасти он делает это, чтобы понять, как именно на вас воздействовать. Он прекрасно подстраивается под то, что вы говорите. Вам кажется, что разговор движется вперёд, что вы вместе чего-то добиваетесь. А потом до вас доходит, что всё это время вы бежали на месте.

Уже в Loopt прозвучали обвинения, которые через много лет повторятся во время кризиса в OpenAI. По данным The Wall Street Journal, за время его руководства Loopt высокопоставленные сотрудники дважды приходили к совету директоров и просили уволить Альтмана. Претензий было две. Первая: он порой

действовал прежде всего в собственных интересах, а не в интересах компании, а иногда и в ущерб ей. Вторая: он был склонен искажать правду. Со второй было сложнее. Порой речь шла о таких мелочах, что люди даже не понимали, зачем было лгать. Каждый отдельный случай казался пустяком. Один человек сравнивал их с крошечными порезами бумагой: один порез – ерунда, но когда их сотни, это уже совсем другое дело. Со временем в компании воцарилась атмосфера постоянного недоверия и неразберихи.

И всё же тот кризис закончился так, как потом будут заканчиваться для Альтмана многие конфликты: он победил. Совет директоров Loopt встал на его сторону.

Хотя сама компания добилась немногого, Альтман вышел из истории с Loopt куда сильнее, чем вошёл в неё. Стартап стал для него трамплином. Через него Альтман проник во всё более влиятельные круги Кремниевой долины, потом с помощью этих связей устроил продажу компании, а продажа сделала его богатым. В двадцать шесть лет Альтман получил от сделки около пяти миллионов долларов. Сам он считал это разочарованием: Стив Джобс, например, к двадцати пяти годам владел состоянием примерно в 256 миллионов. Но вскоре денег у Альтмана станет гораздо больше, и постепенно они начнут менять его образ жизни. Он перестанет сам ходить за продуктами, начнёт летать частными самолётами, соберёт коллекцию дорогих спортивных машин – среди них McLaren и чрезвычайно редкий Koenigsegg почти за пять миллионов долларов. Он увлечётся гонками. Какое-то время будет ездить на Burning Man – знаменитый недельный фестиваль в пустыне, известный своим искусством, психоделической культурой и свободными нравами. Как и многие в кремниевой элите, Альтман время от времени употреблял кетамин – препарат, который используют как рекреационный наркотик, но могут и законно назначать при депрессии.

Добившись успеха, Альтман потянул за собой и братьев. В 2012 году вместе с Джеком он основал собственный инвестиционный фонд Hydrazine Capital. Джек изучал экономику в Принстоне и сначала попробовал себя в инвестиционном банкинге, потом ушёл в технологическую отрасль и основал стартап Lattice – тот получит деньги от Y Combinator уже после того, как президентом акселератора станет Сэм. Макс изучал информатику в Университете Дьюка, какое-то время работал в Microsoft, потом стал трейдером. В 2014 году он перешёл в другой стартап YC, Zenefits, а ещё через два года присоединился к Сэму и Джеку в Hydrazine Capital. В какой-то момент оба младших брата переехали к Сэму – предполагалось, что ненадолго. В итоге трое братьев прожили вместе долгие годы.

Семья и бизнес переплетутся так тесно, что отделить одно от другого станет почти невозможно.

Два наставника

Особенно важными для будущего Альтмана оказались отношения с двумя людьми – Полом Грэмом и Питером Тилем. Пол Грэм, которого в Кремниевой долине обычно звали просто PG, сначала прославился как сооснователь Viaweb – стартапа, который Yahoo купила в 1998 году за 49 миллионов долларов. Потом он стал известным блогером, писавшим о стартапах, предпринимательстве и венчурном капитале, а затем создал Y Combinator. Через YC взгляды Грэма стали влиять на целые поколения предпринимателей. Каждый успешный стартап прибавлял престижа акселератору, а заодно и самому Грэму. К моменту продажи Loopt через YC уже прошли компании, которые стали или вот-вот станут миллиардными, – Dropbox, Airbnb и другие. Y Combinator постепенно превратился в самый элитарный клуб Кремниевой долины.

Попал туда – и перед тобой открывались двери: сразу появлялся статус, доступ к инвесторам, связи с другими выпускниками YC, первые потенциальные клиенты и нередко более высокая оценка компании. Не попал – ничего этого нет. Грэм стал одним из тех, кто определял не только то, какие стартапы получат деньги, но и само представление Кремниевой долины о том, каким должен быть «настоящий основатель». Многие предприниматели воспринимали его взгляды почти как свод законов. Но и критиков у него хватало. Грэм любил называть технологическую индустрию меритократией, где побеждает лучший, однако сам YC выстроил как довольно закрытый клуб. Когда его упрекали в том, что среди основателей стартапов мало женщин, он объяснял это тем, что большинство женщин с детства просто не готовили к успеху в технологическом предпринимательстве.

А разбирая собеседования с кандидатами в YC, Грэм уверял, что выявил десятки признаков, предсказывающих неудачу. Среди них был, например, «сильный иностранный акцент». Если у кандидата обнаруживалось сразу несколько таких признаков, вероятность провала, по мнению Грэма, росла. Он считал это не предвзятостью отбора, а объективным сигналом, подтверждённым данными.

В Альтмана Грэм поверил почти сразу. В 2006 году он описал в блоге свою первую встречу с ним, тогда ещё второкурсником: «Loopt, пожалуй, самый многообещающий из всех стартапов, которые мы пока профинансировали. Но Сэм Альтман – совершенно необычный парень. Помню, через три минуты после знакомства я подумал: „Ага, наверное, таким был Билл Гейтс в девятнадцать лет“». Очень скоро Грэму захотелось найти ещё таких Альтманов, и он спросил Сэма, какой вопрос добавить в анкету Y Combinator, чтобы находить людей вроде него.

Альтман предложил вопрос, который вскоре станет одним из самых знаменитых в YC: «Расскажите, как вам особенно удачно удалось „взломать“ какую-нибудь систему – не компьютерную – в свою пользу». В этом вопросе

отлично отразилась новая предпринимательская этика, которую YC годами будет прививать своим выпускникам: правилам не обязательно подчиняться. Их можно обойти, найти лазейку, переписать, а если понадобится – сломать. Главное – победить.

Когда Альтману было двадцать три, Грэм уже сравнивал его со Стивом Джобсом. «Консультируя стартапы, я чаще всего ссылаюсь на двух основателей – Стива Джобса и Сэма Альтмана, – писал он. – В вопросах дизайна я спрашиваю: „Как поступил бы Стив?“ А в вопросах стратегии и амбиций: „Как поступил бы sama?“» Sama – так Альтмана называли друзья, и так же звучит его ник в X. Именно почти безграничная вера Пола Грэма в Альтмана в конце концов приведёт того на вершину Y Combinator. В 2014 году, через два года после продажи Loopt, двадцативосьмилетний Альтман станет президентом YC. Когда Грэм на кухне спросил, не хочет ли Сэм сменить его, тот не смог сдержать улыбку. Привлекательность новой должности Альтман потом объяснял очень просто: «Y Combinator в какой-то мере получает возможность направлять развитие технологий». Пол Грэм выражался ещё грандиознее: «По-моему, его цель – создать всё будущее». Историю этого назначения будут пересказывать столько раз, что она станет одной из легенд Кремниевой долины.

– Если Сэм улыбается, то обычно вполне сознательно, – шутил один бывший основатель из YC. – Но однажды он правда не мог перестать улыбаться – когда PG предложил ему возглавить YC.

Многих выбор Грэма удивил, но у самого Грэма сомнений не было. «Не было списка кандидатов, в котором Сэм стоял бы первым, – вспоминала Джессика Ливингстон. – Был просто Сэм».

Вторым важнейшим наставником Альтмана стал Питер Тиль, ещё один человек огромного влияния в мире технологий. Тиль разбогател на миллиарды как сооснователь PayPal и аналитической компании Palantir, а также как один из первых инвесторов Facebook. Как и Пол Грэм, он был фигурой далеко не бесспорной. В 2016 году Тиль оказался одним из немногих крупных людей из технологической индустрии, кто открыто поддержал Дональда Трампа. Кроме того, он тайно финансировал судебный процесс, который в итоге погубил Gawker Media, – в отместку за то, что сайт почти десятью годами раньше публично рассказал о его гомосексуальности.

После продажи Loopt Альтман переживал тяжёлое время. Он расстался и с одним из сооснователей компании, и с человеком, с которым прожил девять лет. Разбитый личной драмой и не очень понимая, куда двигаться дальше, Альтман взял год передышки, создал инвестиционный фонд Hydrazine Capital и собрал в него 21 миллион долларов. Большую часть денег вложил Питер Тиль, почти на двадцать лет старше Альтмана. Став партнёром Y Combinator, Альтман начал вкладывать деньги Hydrazine в компании акселератора и заодно помогал венчурному фонду Тиля Founders Fund находить самые перспективные проекты.

Состояние Тиля за это время выросло многократно. Их отношения стали очень близкими; однажды их сравнили лишь с одной другой парой наставника и ученика – с отношениями самого Тиля и Марка Цукерберга.

Пол Грэм и Питер Тиль во многом сформировали взгляды Альтмана – на то, как строить компании, что такое власть и как вести политическую игру. Оба внушали ему одну важнейшую мысль: решает масштаб. И ещё одну: капитализм работает лучше государства. «Первый совет о стартапах, который я когда-либо услышал, звучал так: „Рост продаж решает все проблемы“», – написал Альтман в 2013 году в посте под названием «Рост и государство», где прямо благодарил Грэма и Тиля за влияние на свои взгляды. По сути, пояснял он, это лишь другая формулировка мысли Пола Грэма о том, что рост решает всё. Для стартапа больше продаж – значит больше денег; больше денег – значит можно нанять лучших людей; лучшие люди – меньше внутренних склок и больше шансов победить.

Ту же логику Альтман переносил и на целые страны. Больше экономического роста – значит больше технологических новшеств, а значит, и лучше жизнь. Неработоспособность американского правительства, считал он, угрожает этому циклу. Его формула была жёсткой: «Либо ты растёшь, либо медленно умираешь». И американское государство, по его мнению, как раз умирало. Без экономического роста, писал Альтман, демократия перестаёт нормально работать: избиратели оказываются в игре с нулевой суммой, где, если кому-то становится лучше, другой начинает думать, что у него что-то отняли.

Со временем эта мысль стала одной из главных тем всей его карьеры.

– Лучшее, что частный сектор может сделать, чтобы страна снова встала на правильный путь, – это вернуть экономический рост, – говорил он в 2017 году.

Соединённые Штаты, рассуждал Альтман, пережили около двухсот лет почти небывалого роста: сто лет территориального расширения, а затем сто лет бурного технологического развития. Кровавой истории колониальной экспансии и неоднозначных последствий безудержной индустриализации для рабочих и для природы он при этом почти не касался.

– И люди в основном были довольно счастливы, – подводил он итог. – А теперь нет.

В 2019 году он высказался ещё яснее: «Устойчивый экономический рост почти всегда – моральное благо».

– Отчасти именно поэтому я работаю над Y Combinator и OpenAI, – добавлял он. – Я хочу вернуть нас к устойчивому экономическому росту, к миру, где жизнь большинства людей с каждым годом становится лучше и где мы снова чувствуем общий дух успеха.

Строя компании, Альтман часто опирался и на другую идею Тиля: стремиться надо к монополии. Тиль считал, что по-настоящему успешный основатель не просто хорошо конкурирует, а делает так, чтобы конкурировать с ним было почти невозможно. В 2014 году Альтман вернулся в Стэнфорд – теперь

уже как преподаватель курса «Как основать стартап». Одну из лекций он отдал Питеру Тилю, и называлась она вызывающе: «Конкуренция – удел неудачников». Монополии хороши, объяснял Тиль, потому что такие компании устойчивее, живут дольше, располагают большим капиталом и обычно служат признаком того, что создано нечто действительно ценное.

Чтобы построить монополию, по его словам, нужно несколько вещей: собственная технология, сетевые эффекты, экономия на масштабе, сильный бренд. Но главное, чтобы всё это сохраняло ценность годами. Особенно важно не упустить технологическое лидерство.

– Вам не нужно, чтобы кто-то вас обошёл, – предупреждал Тиль.

Он приводил в пример производство жёстких дисков в 1980-е. Технология стремительно улучшалась, каждые пару лет случался новый прорыв – но всякий раз его совершала другая компания. Для покупателей это было прекрасно, а для владельцев компаний – совсем нет. По мнению Тиля, мало совершить первый большой прорыв. Нужно совершить и последний – тот, что позволит надолго удержать лидерство, – а затем совершенствовать продукт так быстро, чтобы никто никогда не догнал.

– Если будущее устроено так, что инноваций много и в вашей области постоянно кто-то придумывает что-то новое, это прекрасно для общества, – делал он почти циничный вывод. – Но для вашего бизнеса не очень.

Связи как форма власти

От Грэма и Тиля Альтман усвоил ещё один принцип: связи – тоже капитал. «Я слышал множество теорий о том, как в мире делаются дела, – писал он в 2013 году. – Лучшая из них: сосредоточенность плюс личные связи». По словам Альтмана, эту мысль однажды высказал Полу Грэму Чарли Роуз, а Грэм передал её ему. Позже Альтман добавил третью составляющую – веру в себя.

– Для стартапов это действительно важно, – говорил он. – Вы должны всерьёз верить, что у вас может получиться.

Этой формуле Альтман стал следовать почти с религиозным рвением. Отношения он выстраивал с огромной дисциплиной: сначала предлагал людям своё время и советы, а потом, когда в его распоряжении оказывалось всё больше капитала, – и деньги. Идеальным образцом здесь был для него Тиль. Тот годами с помощью советов и инвестиций выстраивал сеть знакомств, а затем использовал её, чтобы обрасти ещё большими связями и ещё большими деньгами. Молодым предпринимателям, которых он хотел втянуть в свою орбиту, он давал небольшие суммы, наставничество и доступ к нужным людям. Альтман научился тому же. И чем богаче и влиятельнее он становился, тем с большим размахом пользовался этим приёмом.

Он регулярно устраивал у себя дома ужины и встречи для самых разных, но пересекающихся кругов: основателей из Y Combinator, предпринимателей из

компаний, куда он вложился, людей со схожими инвестиционными интересами, представителей растущего сообщества предпринимателей-геев. Советы он часто давал в коротких звонках и сообщениях – иногда разговор длился пару минут: ему было важно уместить в день как можно больше контактов. Двух людей он мог познакомить по электронной почте одним словом «meet» – «знакомьтесь» – или вообще одним знаком вопроса, как славился делать Джефф Безос. И в инвестициях Альтман предпочитал не несколько огромных ставок, а множество мелких. По подсчётам The Wall Street Journal на июнь 2024 года, через YC, Hydrazine и другие структуры он был финансово связан более чем с четырьмя сотнями компаний.

В ближнем кругу Альтмана трудно было найти человека, с которым его не связывали бы какие-нибудь деньги. Второй стартап, в который он вообще вложился, оказался и одним из самых удачных: платёжная компания Stripe из портфеля YC, где первым техническим директором был Грег Брокман. Альтман рано вложился в Airbnb, тоже из YC, сооснователь и гендиректор которой Брайан Чески стал одним из его самых близких друзей. Он вложил деньги в биотехнологическую компанию Conception своего бывшего бойфренда и друга Мэтта Крисилоффа, а в других сделках выступал соинвестором ещё одного бывшего бойфренда, известного венчурного инвестора-одиночки Лакки Грума. Для этих людей всё это – свидетельство щедрости Альтмана. Он и правда охотно делится своими ресурсами: пускает людей пожить в своих домах, поддерживает деньгами. Не раз помогал даже совершенно незнакомым людям – однажды, вспоминает один из его знакомых, перевёл деньги жителю Эфиопии, который написал ему по почте и попросил помочь купить ноутбук. Во время кризиса Silicon Valley Bank в 2023 году, когда паника вокруг ключевого для стартапов банка обернулась крупнейшим банкротством со времён 2008 года, Альтман, по словам Крисилоффа, без всяких бумаг переводил деньги компаниям, чтобы те не закрылись и не уволили людей.

– Это чрезвычайно редкое качество, – говорит Грум. – И оно передалось мне – эта щедрость. Я очень ему за это благодарен.

Политика

Тот же подход Альтман применял и к политикам – и здесь тоже следуя примеру Тиля. Но если Тиль поддерживал своими деньгами республиканцев, вкладывая в их кампании десятки миллионов, то Альтман всё активнее участвовал в политике с противоположной стороны: устраивал сбор средств и выписывал чеки демократам. Какое-то время политические разногласия даже омрачали его отношения с Тилем. В 2017 году Альтман решил разобраться в позиции оппонентов: как и другой ученик Тиля, Цукерберг, он отправился в поездку по Америке и поговорил со ста сторонниками Трампа. Подумывал Альтман и о собственной политической карьере – например, о том, чтобы

баллотироваться в губернаторы Калифорнии. Рассуждал он так: губернатор Калифорнии руководит пятой по величине экономикой мира, а это хороший плацдарм, чтобы исправить то, что он считал неработоспособностью политической системы. Он даже опубликовал манифест «Объединённый список» с тремя принципами: процветание благодаря технологиям, экономическая справедливость и личная свобода. Он устраивал фокус-группы, чтобы проверить, как избиратели воспримут его кандидатуру. Близкие шутили, что метить надо сразу в президенты США.

В итоге политиком Альтман так и не стал – фокус-группы решили, что он выглядит слишком юным, – но вести себя начал как политик. В первые годы во главе YC у него ещё были по-мальчишески круглые щёки и один-единственный пиджак, а на стуле он сидел, задрав ногу или примостившись, как птица на жёрдочке. Говорил он порой небрежно, с лёгкими преувеличениями, щедро пересыпая речь ругательствами. Легко упоминал пикантные подробности своей жизни – например, коллекцию оружия. Быстро раздражался и не скрывал нетерпения, если люди казались ему неэффективными; однажды даже написал программу, которая оценивала основателей из YC по тому, как быстро они отвечают на письма.

Через несколько лет он отшлифовал внешность и сгладил углы. Футболки и шорты-карго сменились приталенными хенли и джинсами. За один год он нарастил восемь килограммов мышц, чтобы его небольшая фигура выглядела солиднее. Он научился меньше говорить, больше спрашивать и, сдвинув брови, излучать вдумчивую скромность. Наедине и с близкими друзьями у него по-прежнему прорывались гнев и раздражение. Но на людях и со знакомыми он был воплощённым «славным парнем». Охотно хвалил других, писал сообщения строчными буквами с кучей весёлых и грустных смайликов. Раздавал сотрудникам свой личный номер, предлагал писать в любое время и на их отзывы отвечал с поразительной внимательностью. Он избегал проявлять отрицательные эмоции, избегал столкновений, избегал говорить людям «нет». Однажды, когда OpenAI уволила сотрудника, Альтман лично связался с ним и предложил в утешение кетамин и выпивку.

– Мне кажется, все отношения с Сэмом заканчиваются хорошо – хотите вы того или нет, – говорит этот сотрудник.

Человек-институт

Альтман сам стал институтом. YC был для него платформой и ускорителем: власть акселератора он превращал в собственную власть, его сеть связей – в собственную сеть. Личные знакомства и публичная репутация стали его главной валютой. Он регулярно встречался с политиками, видевшими в нём проводника в Кремниевую долину. В 2016 году совета у него просил министр обороны в администрации Обамы Эштон Картер: как его ведомству привлечь молодые

таланты из технологической индустрии. Три года спустя, в марте 2019-го, в тот самый день, когда Альтман ушёл из YC, в OpenAI приехал Чак Шумер, тогда лидер демократического меньшинства в Сенате. Визит был негласным, Шумера сопровождала охрана Секретной службы. Он сидел рядом с Альтманом в кресле перед телевизором, на экране которого пылал камин.

– Вы делаете важную работу, – говорил Шумер сотрудникам. – Мы не до конца понимаем, что именно, но это важно. И я знаю Сэма. Вы в надёжных руках.

* * *

Но за восхождение Альтману приходилось платить всё дороже: у него появлялось всё больше недоброжелателей и откровенных врагов, которые повторяли обвинения его заместителей из Loopt – в своекорыстной погоне за властью и в патологической нечестности. Многие из тех, кому достались его советы, деньги и связи, становились верными сторонниками. Но другие начинали видеть в нём человека, дьявольски ловко умеющего повернуть любую ситуацию в свою пользу. Для некоторых, в том числе для партнёров по YC и других влиятельных людей, это было просто досадной чертой. Для сотрудников и людей, у которых рычагов было куда меньше, – источником страха. А для тех, кто категорически не разделял его взглядов на мир, он был огромной угрозой.

Энни

Восхождение Альтмана разрушило и его отношения с сестрой – к его мучительной боли, а потом и гневу. В детстве и ещё долго, когда ему было уже за двадцать, Сэм и Энни были близки: она самая младшая, он самый старший, её защитник. Энни тянулась к науке, но была и самой творческой, самой эмоционально открытой из детей. Иногда Сэм спрашивал её мнения о своих возлюбленных, делился с ней тайными тревогами и переживаниями. Но чем глубже он врастал в Кремниевую долину, тем более толстые стены, как видела Энни, возводил он вокруг самой ранимой части себя. Он рассказывал ей, вспоминает она, о новых психологических приёмах, которым научился, – например, писать письма короче, чтобы выглядеть более властным руководителем.

Сначала ей было от этого грустно, потом страшно: осталась ли в нём вообще та ранимая часть?

– Какое-то время я ещё замечала её проблески, поэтому и держалась рядом, – говорит она. – А потом пострадавшей от него стала я.

Когда Сэм только разбогател, рассказывает Энни, его тогдашний бойфренд ввёл правило: на каждую крупную покупку Сэм должен жертвовать столько же на благое дело. Какое-то время это сдерживало стремительное разрастание его образа жизни. Но когда деньги стали приходить быстрее, чем он успевал их тратить, его отношения с ними, как ей казалось, усложнились. По её мнению, он

начал копить их, всё больше отдаляясь от людей, которые нуждаются. С конца 2019-го и в первой половине 2020 года, как следует из обширной переписки, которую Энни мне показала, Сэм и остальные члены семьи несколько раз отказывали ей или неохотно соглашались на то, что она считала экстренной финансовой помощью, – на оплату жилья и лечения. В то время, как показывают её медицинские документы и записи психотерапевта, у неё были тяжёлые проблемы с физическим и душевным здоровьем, обострившиеся после внезапной смерти отца. Она осталась без постоянного жилья и в отчаянии, пытаясь сводить концы с концами, занялась секс-работой. Летом 2020 года, когда OpenAI под руководством Сэма впервые оказалась в центре большого общественного внимания, Энни прервала всякое общение с семьёй.

Можно утверждать, что Сэм и его братья лишь следовали примеру матери, пытаясь подтолкнуть Энни к финансовой самостоятельности. Это сложная и болезненная семейная история, и судить о ней по неполным сведениям трудно. В публичном заявлении в январе 2025 года Сэм, его мать и двое братьев выразили любовь к Энни и тревогу за неё, а все её обвинения отвергли как «совершенно не соответствующие действительности». В ответ на мои просьбы об интервью и подробные вопросы Конни Гибстайн прислала сокращённую версию похожего заявления и от дальнейших пояснений отказалась; Сэм через пресс-службу OpenAI и его братья не ответили.

И всё же в истории Энни поразительным образом отражаются многие темы этой книги: всё шире разверзающаяся пропасть между теми, кто выигрывает от так называемого шествия прогресса, и теми, кто остаётся позади; утрата голоса и возможности распоряжаться своей судьбой у тех, кто оказался на дальнем краю этой пропасти; опасность того, что столько власти отдаётся не просто компаниям, но и отдельным людям, которые ими руководят, без опор, способных обеспечить соразмерные сдержки и противовесы. Да и поступки самой Энни сделали её историю неотъемлемой частью рассказа о пути OpenAI и о её влиянии на развитие ИИ. В 2021 году она решится выступить публично с тяжёлыми обвинениями против Сэма: что в детстве он подвергал её сексуальному насилию – это семья назвала «худшим» из её «ложных» обвинений – и что он и вся семья бросили её, когда она была наиболее беззащитна. 6 января 2025 года, за два дня до своего тридцать первого дня рождения, она подаст против Сэма иск о предполагаемом насилии, чтобы успеть до истечения срока давности по таким делам в Миссури. Её упорные попытки заявить о своих обвинениях и рассказать свою версию событий не могли не сказываться на Сэме и на других руководителях OpenAI в то время, когда и он, и компания стремительно обретали влияние и известность во всём мире.

Каждый из этих кусочков мозаики – восхождение Сэма, его характер и отношения с людьми, раздоры, которые он оставлял за собой, потоки денег и власти – объясняет путь, который привёл к его внезапному и недолгому изгнанию. На короткий миг остальной мир смог заглянуть в борьбу, что идёт на самом верху за право определять будущее искусственного интеллекта. И стало видно, насколько стремление господствовать над этой технологией – а она уже перестраивает общество и меняет облик самой планеты – в конечном счёте зависит от полярных ценностей, столкновения самолюбий и вполне человеческого несовершенства горстки людей, которым свойственно ошибаться.

Глава 2

Цивилизаторская миссия

Первым окончательно решил войти в OpenAI Грег Брокман. Ему в сооснователи Альтман лично выбрал Илью Суцкевера, исследователя ИИ из Google: это Альтман неожиданно написал ему и позвал на тот самый ужин в Rosewood летом 2015 года. Узнав, что там будет Илон Маск, Суцкевер сразу согласился.

Брокман и Суцкевер были необычной парой. Высокий, крепкий, приветливый Брокман был инженером и человеком стартапов, во многом похожим на Альтмана. Он вырос на маленькой ферме в Северной Дакоте и в перерывах между дойкой коров увлёкся математикой, а потом и наукой. В 2008 году он поступил в Гарвард, через два года перевёлся в MIT, а ещё через семестр и вовсе бросил учёбу: трудно было сидеть в аудитории, когда за её стенами лежал настоящий мир, где можно создавать настоящие продукты. Он переехал в район Сан-Франциско и пришёл в тогда ещё крошечный стартап Stripe, где, кроме него, работали всего три человека. Основателей так поразил его талант программиста, что Брокман стал техническим директором. За следующие пять лет он участвовал в создании многих первых продуктов Stripe и помог компании вырасти в одного из гигантов финтеха, через которого проходят цифровые платежи таких компаний, как Amazon и Shopify.

Эти годы принесли Брокману немалое состояние и редкий в Кремниевой долине статус человека, своими руками строившего многомиллиардную компанию. Суцкевер был другим – прежде всего учёным. Худой, жилистый, он родился в Советском Союзе, а вырос в Израиле, где очень рано проявил необычайные математические способности. Учителя едва за ним поспевали, и родители уже в восьмом классе записали его на курсы Открытого университета Израиля. В шестнадцать лет он переехал в Торонто, всего месяц проучился в обычной школе, а в 2003 году его приняли в Университет Торонто – сразу на третий курс. Там Суцкевер встретил Джеффри Хинтона, британско-канадского профессора, чьи работы легли в основу современной истории искусственного интеллекта. Хинтон – единственный человек, которого Суцкевер впоследствии называл своим наставником, и влияние его на жизнь и работу Суцкевера оказалось огромным.

В 2012 году Суцкевер, Хинтон и ещё один аспирант Хинтона, Алекс Крижевский, потрясли весь мир ИИ. Они участвовали в конкурсе ImageNet, где команды соревновались в создании программ, автоматически распознающих объекты на фотографиях. Остальным участникам никак не удавалось опустить долю ошибок ниже 25 процентов. Команда Хинтона, Суцкевера и Крижевского добилась 15 процентов. Это был разгром. В начале следующего года после

ожесточённого аукциона Google купила их только что созданную компанию DNNresearch за 44 миллиона долларов. Трое учёных в одночасье стали мультимиллионерами. Но ещё важнее было то, что эта сделка стала одним из первых громких сигналов к массовой коммерциализации искусственного интеллекта.

– Нам казалось, будто мы попали в кино, – вспоминал Хинтон.

Брокман и Суцкевер впервые встретились на том ужине в Rosewood. Остальные гости тоже делились на два лагеря – предпринимателей и учёных. Разговор то и дело перескакивал от академических споров о направлениях исследований к вопросу, который особенно волновал Маска: не поздно ли ещё обогнать DeepMind и Google? Для собравшихся это было не просто соперничество компаний: победа над Google казалась им необходимой, чтобы изменить само направление развития ИИ. И довольно быстро все сошлись на том, что главная нехватка – в людях. Лучшие исследователи ИИ уже работали в Google, как Суцкевер, или у других технологических гигантов, получая огромные зарплаты, щедрые бонусы и надёжность, какую молодая лаборатория предложить почти не могла.

В центре разговора был и AGI – общий искусственный интеллект. В 2015 году всерьёз обсуждать эту тему было непривычно. Большинство уважаемых учёных считали полноценное цифровое воспроизведение человеческого интеллекта научной фантастикой или, в лучшем случае, делом далёкого будущего, на десятилетия вперёд. Разговоры о том, что AGI уже так близко, что в него пора вкладывать огромные деньги, нередко принимали за лженауку. Демис Хассабис, впрочем, уже описывал цели DeepMind словом AGI – и даже некоторые его собственные сотрудники считали это чересчур вызывающим маркетингом. Но участники ужина в Rosewood решили: раз они хотят создать организацию, способную бросить DeepMind прямой вызов, им нужен столь же грандиозный ориентир. И AGI лучше всего выражал их амбиции. Даже Суцкеверу, который в глубине души считал AGI вполне возможным и позже станет одним из самых убеждённых сторонников идеи его скорого появления, поначалу было не по себе.

Он боялся, что, если коллеги-учёные узнают, что он всерьёз обсуждает создание AGI, пострадает его научная репутация. Брокмана такие опасения не мучили: он пришёл в ИИ со стороны и искренне верил, что, если приложить достаточно сил и сосредоточиться на цели, AGI может оказаться гораздо ближе, чем все думают. Уже одно то, что Суцкевер и другие исследователи хотя бы в частном разговоре готовы обсуждать новую лабораторию, укрепило его уверенность. Когда той ночью Альтман вёз его из отеля обратно в Сан-Франциско, Брокман сказал, что готов посвятить себя этому проекту.

В Кремниевой долине любят говорить, что основать компанию вместе – почти то же, что пожениться. Поэтому Брокман и Суцкевер сначала будто присматривались друг к другу. Альтман дипломатично объяснил Брокману, что тому нужен партнёр, по-настоящему разбирающийся в исследованиях ИИ. Через

несколько недель после ужина они вдвоём поужинали в Маунтин-Вью, и оказалось, что подходят друг другу почти идеально. «Хотя мы только познакомились, я уже знал, что всё получится, – писал потом Брокман. – Мы с Ильёй общались невероятно насыщенно. Наши идеи усиливали и дополняли друг друга».

Пока Суцкевер обдумывал предложение, Брокман взялся за главное – за набор команды. Он обзванивал видных людей отрасли и спрашивал, кого они считают лучшими специалистами, встречался с профессорами и выяснял, кто их самые сильные студенты. Перед разговором с каждым кандидатом он тщательно изучал его работы, чтобы понять, чем можно его увлечь. Позже Альтман посвятил ему отдельную запись в блоге под названием просто «Грег»: «Меня часто спрашивают, каким должен быть идеальный сооснователь. Теперь у меня есть ответ: Грег Брокман».

Альтман и Маск тоже вовсю участвовали в наборе. Главная задача состояла в том, чтобы исподволь переубедить исследователей, которым идея AGI казалась слишком смелой. Профессор Беркли Питер Аббил вспоминал, как Маск говорил ему примерно следующее:

– Может быть, до AGI ещё очень далеко. Но что, если нет? Что, если есть хотя бы один процент вероятности – или даже одна десятая процента, – что это случится в ближайшие пять-десять лет? Разве мы не обязаны отнестись к этому очень серьёзно?

Позже Аббил стал научным консультантом OpenAI, а затем перешёл туда насовсем вместе с несколькими своими аспирантами. Поначалу многие кандидаты соглашались лишь при условии, что придут и другие. Брокмана это не смутило. Он выбрал десять инженеров и исследователей, которых хотел заполучить больше всего, и позвал их всех в Напа-Вэлли – формально попробовать вина и поговорить, а на деле на коллективную вербовку. Он арендовал автобус, чтобы отвезти всех туда и обратно, и даже в дороге, больше часа в каждую сторону, продолжал уговаривать. Через три недели, к назначенному Брокманом сроку, почти все приняли предложение.

Обсуждая, как представить OpenAI миру, Маск, Альтман и Брокман особенно задумывались о том, какое впечатление должна производить организация. Они поддержали мысль Альтмана сделать OpenAI некоммерческой и решили всячески подчёркивать то, что было заложено в названии, – открытость. OpenAI должна была стать анти-Google: вести исследования для всех, открывать исходный код, делиться научными результатами, быть образцом прозрачности. «Надеюсь, мы войдём в эту область как нейтральная группа, которая стремится к широкому сотрудничеству и меняет сам разговор: речь не о победе какой-то компании или группы, а о победе человечества, – писал Брокман Маску и Альтману в ноябре 2015 года. И добавлял в скобках: – Думаю, так мы быстрее всего станем ведущим исследовательским институтом».

Маск согласился, но думал и о том, как это воспримет публика. «Очень важно, чтобы общество болело за наш успех», – писал он и предлагал придать объявлению об основании OpenAI больший размах. Особенно его беспокоила сумма. «Нужно назвать цифру гораздо больше 100 миллионов, иначе мы будем выглядеть безнадёжно на фоне того, сколько тратят Google или Facebook, – писал он. – Думаю, надо сказать, что мы начинаем с обязательства на 1 миллиард. Это реально. Всё, чего не дадут другие, покрою я».

В более поздней переписке основатели признавали, что от обещанной открытости можно будет отказаться, когда эта история сделает своё дело и когда возникнет необходимость, – например, чтобы технология не попала в руки злоумышленников. «Когда мы приблизимся к созданию ИИ, будет разумно стать менее открытыми, – написал Суцкевер всем троим в январе 2016 года, вскоре после запуска OpenAI. – Open в слове OpenAI означает, что все должны пользоваться плодами ИИ после того, как он будет создан, но вполне нормально не делиться самой наукой». «Ага», – ответил Маск.

Объявление вышло в декабре 2015 года, вечером в пятницу, и было приурочено к Neural Information Processing Systems, крупнейшей ежегодной конференции по ИИ, – той самой, где тремя годами раньше Хинтон и Суцкевер устроили аукцион по продаже DNNresearch. В посте под названием «Представляем OpenAI» перечислялись все девять основателей: Брокман становился техническим директором, Суцкевер – руководителем исследований. Маск и Альтман становились сопредседателями. Альтман и Брокман присоединились к обещанию Маска обеспечить лаборатории миллиард долларов. То же сделали Джессика Ливингстон, Питер Тиль и сооснователь LinkedIn Рид Хоффман – он когда-то работал с Маском и Тилем в PayPal и часто вкладывался вместе с ними в стартапы в составе так называемой «мафии PayPal». «В ближайшие несколько лет мы рассчитываем потратить лишь малую часть этой суммы», – говорилось в посте о финансировании.

В последние дни перед объявлением Суцкевер чуть было не остался в Google. Всем остальным основателям OpenAI предложила базовую зарплату 175 тысяч долларов плюс акции YC или SpaceX. Суцкеверу же лаборатория предложила почти два миллиона долларов – огромную сумму для некоммерческой организации. Но Google, стараясь его удержать, предложила больше, а потом ещё больше – в два-три раза больше. Маск и Альтман раз за разом откладывали объявление, пока Суцкевер мучительно решал: звонил родителям, выслушивал уговоры Маска и Брокмана. В конце концов головокружительное предложение Google и убедило его в том, что такая некоммерческая организация, как OpenAI, действительно нужна.

В тот день Маск отметил событие торжественным письмом основателям, поклявшись сделать всё для победы OpenAI. «Самое важное для нас – привлечь лучших людей», – писал он и обещал помогать в этом и во всём остальном, чем только сможет. «Организации, которые вы хорошо знаете, превосходят нас и

числом, и вооружением с нелепым отрывом, – добавлял он. – Но правда на нашей стороне, а это многого стоит. Мне нравятся наши шансы». Чтобы чужие встречные предложения не переманили никого из основателей, OpenAI тут же повысила всем базовую зарплату ещё на 100 тысяч долларов.

Позже Маск рассказывал, как Ларри Пейдж пришёл в ярость оттого, что он лично переманил Суцкевера. С тех пор они почти не разговаривали – тем более что их взгляды на развитие ИИ расходились всё сильнее. Зато OpenAI вдруг стало куда легче нанимать людей. Грандиозные замыслы и близость к миллиардерам притягивали в Кремниевой долине как магнит. За несколько месяцев число сотрудников удвоилось.

* * *

Неясность целей, большой чек и преклонение перед миллиардерами – такой была Кремниевая долина на пьянящем пике своей ничем не сдерживаемой власти. Но OpenAI ловко выбрала позицию: она стояла на границе между техношовинистской Кремниевой долиной и нарождавшимся более совестливым течением. За следующий год головокружительный взлёт Дональда Трампа и его победа на президентских выборах 2016 года потрясли левоватых в массе своей работников технологической индустрии и заставили их задуматься о себе. Пока по Meta и Google прокатывались потрясения, а общество охватывало недовольство техногигантами, исследователи ИИ тоже начали спрашивать себя, не слишком ли поспешно отрасль впрягла свои технологии в погоню за прибылью.

Итог подсчёта общественных последствий коммерциализации исследований ИИ тревожил. Автоматизированные программы, которые продавали полиции, ипотечным брокерам и кредиторам, закрепляли дискриминацию по расе, полу и классу. Алгоритмы ленты новостей Facebook и рекомендательных систем YouTube, по всей видимости, поляризовали общество, подпитывали дезинформацию и экстремизм, открывали дорогу вмешательству в выборы, а в самом страшном случае, с Facebook, ускорили этнические чистки в Мьянме.

Но и главная альтернатива – брать деньги у государства – была усеяна этическими минами. В 2018 году тысячи сотрудников Google протестовали против секретного контракта компании с Пентагоном по программе Project Maven, где ИИ должен был использоваться в разведывательных беспилотниках. Сотрудники говорили, что такие возможности могут проложить путь к автономному оружию; Пентагон заверял, что такого намерения у него не было, но с войной в Украине от этой позиции отойдёт.

Среди исследователей ИИ стала ходить циничная присказка: продайся либо Big Tech, либо военно-промышленному комплексу – или уходи из науки об ИИ. Между этими двумя крайностями OpenAI выглядела третьим путём, не испорченным ни прибылью, ни государственной властью.

– Это был маяк надежды, – говорила Чип Хуен, инженер машинного обучения и популярный технологический блогер, наблюдавшая за всем со стороны.

Впечатлены были не все. В ночь запуска OpenAI в декабре 2015 года Тимнит Гебру, та самая исследовательница, что спрашивала Маска, почему угрозы ИИ волнуют его больше изменения климата, не могла поверить своим глазам.

Всю неделю эта аспирантка Стэнфорда – эритрейская беженка, родившаяся в Эфиопии и подростком переехавшая в США, – только и вспоминала, как дорого обходится быть чернокожей женщиной в среде, где правят белые мужчины. Она впервые оказалась среди толп исследователей ИИ на недельной конференции Neural Information Processing Systems, которую тогда сокращённо звали NIPS, а позже переименовали в NeurIPS. Чернокожих там почти не было. На следующей конференции, в 2016 году, она даже посчитает: среди 8500 участников найдётся всего шесть других чернокожих исследователей. В Стэнфорде она шутила, что могла бы основать группу Black in AI («Чернокожие в ИИ») и проводить собрания в одиночку. Она представляла, как заведёт ради насмешки над этим YouTube-канал: будет менять причёски и разыгрывать разных персонажей, изображая разговор с самой собой. Но как бы она ни пыталась отшучиваться от своего одиночества, конференция 2015 года напомнила ей, как быстро всё может обернуться враждебностью. Однажды вечером на вечеринке она подошла за водой, и компания пьяных парней в футболках Google Research остановила на ней взгляд и решила над ней позабавиться. Они её окружили. Один силой заключил её в объятия, другой чмокнул в щёку, одновременно сделав унизительное фото. На той же конференции к её подруге приставал профессор.

И вот перед ней была группа людей – девять из одиннадцати белые мужчины, – которых осыпали небывалыми деньгами, которые рассуждали о теоретической угрозе злого сверхразума, способного захватить мир, и предлагали противостоять ему, построив сверхразум получше.

Той же ночью Гебру написала разгромную критику всего увиденного в виде анонимного открытого письма: спектакля, культового обожествления знаменитостей от ИИ и прежде всего удручающей однородности людей, которые создают и формируют столь важную технологию. Такая однородная среда, писала она, не только отталкивает талантливых исследователей, но и порождает опасно узкое представление о том, что такое ИИ и кому он может принести пользу.

«Чтобы увидеть возможный вред от ИИ, не нужно заглядывать в будущее, – писала Гебру. – Он происходит уже сейчас».

В самолёте по дороге домой она передумала публиковать письмо анонимно и выложила в Facebook под своим именем более короткую и сглаженную версию своей критики.

Через несколько недель она написала письмо с темой «Привет от Тимнит».

«Когда я езжу на конференции по компьютерному зрению, я часто оказываюсь там единственной чернокожей, – писала она. – Но вот я увидела пятерых из вас :) и подумала: было бы здорово создать группу black in AI или хотя бы знать друг о друге».

Один за другим она вписала адреса исследователей. И нажала «отправить».

* * *

В первые годы OpenAI Альтман и Маск как сопредседатели почти не появлялись. Занятые множеством других дел, они оставили строить организацию Брокмана и Суцкевера. Пока Суцкевер собирал исследователей и просил их нести ему лучшие идеи, Брокман с головой ушёл в создание правильной культуры организации.

Через несколько лет Брокман рассказал мне, как он об этом думал. Готовясь, он прочитал все книги, какие смог найти, о дерзких научных и технических предприятиях в истории США: о трансконтинентальной железной дороге, о лампочке Томаса Эдисона, о первой компьютерной сети, из которой вырос современный интернет. Он впитывал их, как священные тексты, ища в них подсказки и указания, как устроить собственное начинание.

Одна из его любимых историй – скорее всего, апокрифическая – о том, как Джон Кеннеди в космическом центре NASA подошёл к уборщику с метлой.

– Кеннеди спрашивает: «Сэр, что вы делаете?» А тот отвечает: «Я помогаю отправить человека на Луну», – рассказывал Брокман с явным восторгом. – Когда у каждого есть такое чувство миссии и цели – по-моему, это нечто совершенно удивительное, и я не вижу, чтобы сегодня это где-то было.

– Мне правда кажется, что мы, американцы, разучились мечтать с размахом, – добавил он позже.

Чтобы добиться успеха, считал Брокман, OpenAI нужна такая же сплочённость: каждый человек на любом уровне компании должен быть как тот уборщик. Он обратил моё внимание на то, что в первые месяцы, когда все работали у него в квартире, он воплощал этот дух буквально – подолгу мыл за всеми стаканы. Он ввёл правило, по которому все сотрудники должны работать в офисе в Сан-Франциско, и OpenAI держалась его вплоть до пандемии. У этого, разумеется, были издержки: не всем хотелось жить в районе залива Сан-Франциско, признавал он. Из других интервью я узнала, что особенно это касалось женщин и небелых сотрудников, которые, как и Гебру, чувствовали себя чужими в белой мужской культуре, господствующей в технологической индустрии. Но для Брокмана сплочённость была важнее, а физическое присутствие рядом друг с другом помогало случайному обмену идеями.

Ещё Брокман решил называть всех сотрудников OpenAI «членами технического персонала». Он позаимствовал это у Xerox PARC, легендарной исследовательской лаборатории в Пало-Альто, а там так поступали по традиции

не менее знаменитой Bell Labs из Нью-Джерси – чтобы создать более демократичную рабочую среду.

На упрёки в адрес OpenAI за её погоню за AGI он отвечал, проводя параллель с лампочкой Эдисона.

– Комитет уважаемых экспертов заявил: «Это никогда не заработает», а через год он уже начал продажи, – говорил Брокман. – Как такое возможно?

Это был, как назвал подобное писатель-фантаст Артур Кларк в книге «Черты будущего», «провал воображения».

* * *

Среди гостей ужина в Rosewood был и Дарио Амодеи – специалист по вычислительной нейробиологии, перешедший в исследования ИИ. Тогда он работал в кремниевой ИИ-лаборатории китайской компании Baidu, а потом недолго – в Google. Его сестра Даниэла Амодеи когда-то работала с Брокманом в Stripe, и, когда Брокман только начал всерьёз вникать в ИИ, он обращался к Дарио за советом, что почитать. Амодеи пришёл в OpenAI не сразу, но замысел его заинтриговал: под влиянием Маска OpenAI выделялась среди прочих ИИ-лабораторий тем, что охотнее всех была готова заниматься так называемой безопасностью ИИ.

В 2016 году, ещё работая в Google, Амодеи стал соавтором основополагающей для этой дисциплины статьи, где центральная задача безопасности ИИ формулировалась как решение «проблемы аварий в системах машинного обучения, понимаемых как непреднамеренное и вредное поведение, которое может возникнуть из-за плохого проектирования реальных систем ИИ». Это, писали он и его соавторы, отдельная задача, не совпадающая с другими проблемами ИИ – конфиденциальностью, защищённостью, справедливостью и экономическими последствиями. Иными словами, «безопасность» ИИ в таком понимании сводилась к тому, чтобы предотвратить появление вышедшего из-под контроля, несогласованного ИИ – того корня, из которого, по описанию Ника Бострома, сверхинтеллект может вырасти в угрозу самому существованию человечества.

Для Амодеи не было задачи важнее: не дать сверхчеловеческому ИИ привести к катастрофе, вплоть до вымирания человечества. Оба Амодеи симпатизировали движению эффективного альтруизма (ЭА) – спорной идеологии, зародившейся среди философов Оксфордского университета, где работал Бостром, и пустившей корни в Кремниевой долине. Движение проповедует посвятить себя тому, чтобы делать в мире как можно больше добра, руководствуясь в решениях предельной рациональностью и парадоксальной логикой; со временем, во многом под влиянием Бострома, оно объявило экзистенциальную угрозу вышедшего из-под контроля ИИ одним из главных направлений для своих приверженцев. За два года до того муж Даниэлы, Холден Карнофски, основал некоммерческую организацию Open Philanthropy, которая

раздавала деньги отчасти по принципам ЭА. Open Phil, как её называли, быстро стала главным спонсором исследований безопасности ИИ, связанных с катастрофическими и экзистенциальными рисками. (К ноябрю 2024 года она выдала более трёхсот грантов на безопасность ИИ на общую сумму 440 миллионов долларов.)

Но такую экзистенциальную версию безопасности ИИ, построенную на философских мысленных экспериментах, вскоре начали критиковать: исследовательское сообщество открывало для себя менее апокалиптический, зато вполне насущный реальный вред от ИИ. Примерно тогда же, когда Амодеи опубликовал свою статью, ProPublica выпустила новаторское расследование «Предвзятость машин». Оно показало, что в системе уголовного правосудия США повсюду используют алгоритмы, которые без особых оснований пытаются предсказывать будущих преступников, и что эти алгоритмы относят чернокожих к группе более высокого риска, чем белых с куда более богатым криминальным прошлым. Эта статья и общее разочарование в Big Tech после 2016 года из-за вреда социальных сетей вызвали новую волну исследований, всерьёз занявшихся вредными общественными последствиями ИИ.

Дебора Раджи, исследовательница из Калифорнийского университета в Беркли, изучающая подотчётность ИИ, станет одной из главных сторонниц пересмотра того, что исследования безопасности ИИ почти целиком сосредоточены на теоретическом вышедшем из-под контроля ИИ и его возможных экзистенциальных рисках – в ущерб другим, реальным и доказанным проблемам, которые отодвигаются на второй план. В 2020 году она стала соавтором статьи, написанной в ответ на работу Амодеи. По её словам, по-настоящему «безопасные» системы ИИ невозможно создать, рассматривая поведение технических систем в отрыве от всего остального, без учёта их влияния на то самое – конфиденциальность, справедливость, экономику, – что Амодеи вынес за скобки. Амодеи писал, что ИИ, упорно преследуя цель, может производить «негативные побочные эффекты», и приводил пример в духе мысленного эксперимента со скрепками: робот-уборщик по пути к наведению порядка опрокидывает вазу или царапает стены. Раджи же указывала, что это уже происходит. В неуклонной погоне за коммерческими продуктами и AGI ИИ-индустрия породила обширные негативные побочные эффекты – от массового нарушения частной жизни ради обучения систем распознавания лиц до стремительно растущей экологической цены дата-центров, без которых технология не может развиваться.

«Побочные эффекты может вызывать не только поведение ИИ-агента, – писали она и её соавтор. – В реальной жизни базовые проектные решения при создании и внедрении моделей тоже имеют последствия, которые выходят далеко за пределы влияния одного-единственного решения модели. На деле само создание систем ИИ очень часто имеет скрытую человеческую цену».

Внутри OpenAI разные исследователи – среди них несколько из совсем немногих небелых женщин в компании – будут убеждать руководителей расширить определение «безопасности ИИ» и включить в него, например, изучение того, как модели глубокого обучения закрепляют дискриминацию. Руководители отмахивались.

– Это не наша задача, – сказал один из них.

В мае 2016 года Амодеи, ещё работавший в Google, заглянул в офис OpenAI посмотреть, как идут дела. OpenAI только что переехала из квартиры Брокмана в помещение над шоколадной фабрикой в Мишн – самом старом районе Сан-Франциско, оплоте латиноамериканцев. Исследователи расхаживали в носках.

– Человек двадцать-тридцать в нашей области, включая Ника Бострома и статью в «Википедии», говорят, что цель OpenAI – создать дружественный ИИ, а потом выложить его исходный код для всего мира, – сказал Амодеи Альтману и Брокману, как пересказывал позже The New Yorker.

– Весь исходный код мы выкладывать не собираемся, – ответил Альтман. – Но давайте, пожалуйста, не будем это поправлять. Обычно от этого становится только хуже.

– Но какова цель? – спросил Амодеи.

– Наша цель сейчас… делать самое лучшее, что только можно сделать, – ответил Брокман. – Это немного расплывчато.

Через два месяца Амодеи пришёл в OpenAI руководить исследованиями безопасности ИИ. Позже Open Phil пожертвует OpenAI 30 миллионов долларов, чтобы обеспечить Холдену Карнофски место в совете директоров на три года. В 2018 году по приглашению Брокмана в OpenAI перейдёт и Даниэла, когда-то первая рекрутерша Stripe: она будет руководить инженерной командой и станет вице-президентом по персоналу.

– У нас долгая и милая история знакомства, – шутила Даниэла год спустя, говоря со мной о себе и Брокмане.

– Это правда, – посмеивался Брокман. – Когда мы начинали OpenAI и я взялся за первый набор людей, я думал: «Как же мне не хватает Даниэлы».

К концу 2020 года брата и сестру Амодеи так встревожит то, что они сочтут отступлением Альтмана и OpenAI от изначального замысла, что они отколются и создадут другую ИИ-лабораторию, Anthropic, уведя с собой ключевых сотрудников. Их соперничество сыграет решающую роль в лихорадочном выпуске ChatGPT. Карнофски выйдет из совета OpenAI: срок его истечёт, да и возникнет новый конфликт интересов. В список кандидатов себе на смену он включит одну из своих бывших сотрудниц – Хелен Тонер.

* * *

Беда была в том, что OpenAI понятия не имела, что делает. За первый год она, переманивая, выпрашивая и одалживая, собрала в ожесточённой борьбе за таланты блестящую команду, и одно лишь скопление лучших умов

поддерживало внутри азарт. И всё же внятной стратегии нащупать не удавалось. А первый порыв и блеск начинали тускнеть.

Список проектов расползался во все стороны, отражая всю мешанину направлений в области. Агентов обучали с помощью роботов, видеоигр и смоделированных виртуальных миров – всё ради того, чтобы добиться более продвинутых возможностей ИИ. Почти ничего не получалось, а то, что получалось, казалось повторением уже сделанного кем-то другим. Чем бы ни был AGI, это был не он.

– В их больших проектах не было, по-моему, ничего особенно новаторского, – говорит Никхил Мишра, исследователь ИИ, стажировавшийся в OpenAI в 2017 году.

Руководительские способности Брокмана и Суцкевера тоже подвергались серьёзному испытанию. Брокман целыми днями программировал, а Суцкевер бродил по офису и раз за разом спрашивал каждого исследователя: «Какая у тебя следующая большая идея?» Получалась нервная обстановка без руля и ветрил. Настоящей системы управления и ясных приоритетов не было. Иногда людей увольняли в выходные, и остальные узнавали об этом только в понедельник, когда те не приходили на работу. А лаборатория прожигала деньги – больше всего на зарплаты собранной команде. В 2016 году из 11 миллионов долларов расходов OpenAI больше 7 миллионов ушло на оплату труда и льготы.

Маск начинал терять терпение. Масла в огонь подливало то, что DeepMind вдруг снискала восторги всего мира. В марте 2016 года её программа AlphaGo обыграла Ли Седоля, одного из лучших в мире игроков в древнюю китайскую игру го. («DeepMind доводит меня до крайнего умственного напряжения, – писал Маск руководству OpenAI незадолго до матча из пяти партий. – Если они выиграют, это будет очень плохая новость, учитывая их философию „одного разума, который будет править миром“».) Партии транслировали из Южной Кореи в прямом эфире более чем двумстам миллионам зрителей. Через год Netflix выпустил о пути компании документальный фильм, имевший огромный успех.

Маск время от времени наведывался в офис и требовал больших успехов, порой назначая совершенно нереальные сроки – это было вполне в духе его подхода к управлению. Многих сотрудников такие ожидания раздражали: исследования идут извилистым, непредсказуемым путём, и требования Маска казались им бессмысленными. На одном общем собрании Войцех Заремба, руководитель направления робототехники, входивший в группу основателей, рассказывал, каких успехов в робототехнике хотела бы добиться его команда. У Маска был лишь один вопрос:

– Когда? Когда вы всё это сделаете?

– Не знаю, – ответил Заремба.

– Ну, тогда у вас на самом деле нет плана, – отрезал Маск.

И в марте 2017 года Брокман и Суцкевер всерьёз взялись за более сосредоточенную программу исследований. Главный вопрос звучал так: что на самом деле понадобится OpenAI, чтобы прийти к AGI – и прийти первой?

Суцкевер интуитивно чувствовал, что всё упирается прежде всего в одно измерение: в объём «вычислений», как специалисты называют вычислительные ресурсы, который понадобится OpenAI для крупных прорывов в возможностях ИИ. И ImageNet, и последовавшие за ним достижения, в которых он участвовал, были связаны с заметным ростом объёма вычислений, потраченных на обучение модели. Свою роль играло и другое – куда больше данных и более изощрённые алгоритмы. Но вычисления, считал Суцкевер, решают всё. И если удастся нарастить их настолько, чтобы обучить модель размером с человеческий мозг, непременно произойдёт нечто радикальное – появится AGI.

Объём вычислений зависит от трёх вещей: от мощности одного чипа, то есть сколько операций в секунду он выполняет; от общего числа доступных чипов; и от того, сколько времени они работают. Первое определяет индустрия микросхем, которая десятилетиями удваивает мощность отдельного чипа каждые два года ценой огромных усилий в исследованиях и разработках. Этот темп называют законом Мура – по прогнозу, который легендарный сооснователь Intel Гордон Мур впервые сделал в 1960-х, а через десять лет уточнил, о том, как быстро его отрасль способна обновляться. Закон Мура стал самосбывающимся пророчеством: производители чипов стали считать его ориентиром, тем темпом, с которым нужно обновляться, чтобы не отстать от конкурентов и остаться на плаву.

Брокман и Суцкевер сделали простой расчёт: сколько времени при темпе закона Мура понадобится, чтобы выйти на объём вычислений, необходимый OpenAI для ИИ масштаба мозга? Ответ огорчал: слишком долго.

Примерно тогда же Амодеи и ещё один исследователь, Дэнни Эрнандес, подошли к той же мысли с другой стороны. На простом графике с временем по горизонтальной оси они отметили, сколько вычислений реально потребовал каждый крупный прорыв в исследованиях ИИ начиная с 2012 года – с аспирантского прорыва Суцкевера, положившего начало революции ИИ. Оказалось, что использование вычислений росло быстрее закона Мура. Намного быстрее. За последние шесть лет оно удваивалось каждые 3,4 месяца – иначе говоря, выросло на 30 миллионов процентов.

Брокман стал называть эту новую кривую удвоения законом OpenAI. Он и другие руководители считали, что OpenAI не только понадобится колоссально больше вычислений для достижения конечной цели, но и наращивать их придётся как минимум в темпе этого нового закона. Производители чипов навязали своим компаниям закон Мура с прямо-таки экзистенциальным рвением; теперь руководство OpenAI видело в таком же свете закон OpenAI.

Раз ждать закона Мура нельзя, вычисления придётся наращивать иначе – им нужно было чертовски много чипов.

* * *

Чипы, которые требовались OpenAI, стоили дорого. Эти графические процессоры, GPU, изначально создавались, чтобы быстро отрисовывать графику на компьютерах – скажем, делать видеоигры глянцевыми и без задержек. Но та же конструкция превосходно подходила для обучения моделей ИИ, которые хотела создавать OpenAI: и здесь, и при отрисовке графики нужно параллельно перемалывать огромные массивы чисел.

Подавляющее большинство отрасли покупало эти GPU у одной-единственной компании – производителя чипов Nvidia со штаб-квартирой в Санта-Кларе. Nvidia не только делала лучшие GPU в мире, но и создала к ним программную платформу CUDA (Compute Unified Device Architecture), которая держала разработчиков ИИ мёртвой хваткой.

В 2017 году специальный сервер Nvidia с восемью её лучшими GPU стоил 150 тысяч долларов – к 2023 году цена вместе с инфляцией вырастет почти до 195 тысяч. По расчётам закона OpenAI, в ближайшие годы ей понадобятся тысячи, если не десятки тысяч GPU, чтобы обучить всего одну модель. Стремительно вырастут и расходы на электричество. OpenAI нужно было больше денег – не один миллиард, а миллиарды, чтобы продержаться в ближайшие годы.

Это осознание выбьет у организации почву из-под ног. Для Брокмана и Суцкевера оно поставило под вопрос саму основу устройства OpenAI. Как некоммерческой организации каждый год собирать столько денег, чтобы держать темп, необходимый для лидерства? Они ненадолго задумались о слиянии со стартапом, делающим чипы, а летом 2017 года начали всерьёз обсуждать с Альтманом и Маском, не превратить ли OpenAI в коммерческую компанию. Это была их главная надежда привлечь инвесторов – шансом получить прибыль. Но через несколько недель переговоров обсуждение внезапно оборвалось ничем. Если OpenAI станет коммерческой компанией, Альтман – который как раз прикидывал, не баллотироваться ли в губернаторы Калифорнии, и получал прохладный приём в фокус-группах, – хотел быть её генеральным директором. Этого же хотел Маск: он требовал полного контроля над лабораторией и контрольного пакета.

Оказавшись между ними, Суцкевер и Брокман чуть было не выбрали Маска: его руководство им нравилось больше. Но Альтман обратился напрямую к Брокману, напомнив об их личной дружбе и о своих опасениях насчёт ненадёжности Маска. На Маска давило множество внешних обстоятельств, и он был склонен к непредсказуемым, неустойчивым поступкам. Если OpenAI добьётся успеха, не опасно ли отдать Маску полный контроль над AGI? Убеждённый этим, Брокман стал уговаривать Суцкевера, который всё ещё колебался. В сентябре 2017 года Суцкевер от своего имени и от имени Брокмана написал Маску и Альтману, пытаясь в последний раз разрешить ситуацию.

«Илон, мы очень хотим работать с тобой, – писал Суцкевер. – Мы верим, что вместе у нас больше всего шансов выполнить миссию». Но стремление Маска к полному контролю, писал он, противоречит изначальному духу OpenAI. «Ты опасаешься, что Демис может установить диктатуру AGI. Мы тоже. Поэтому плохая идея – создавать структуру, при которой диктатором, если захочешь, сможешь стать ты.

«Сэм, когда мы с Грегом заходим в тупик, у тебя всегда находится ответ, который оказывается глубоким и верным, – продолжал Суцкевер. Однако поведение Альтмана нередко сбивало их с толку: во что он на самом деле верит и чего добивается? – Мы не понимаем, почему тебе так важна должность гендиректора. Причины, которые ты называешь, менялись, и трудно понять, что на самом деле тобой движет. Действительно ли AGI – твоя главная цель? Как это связано с твоими политическими целями? Как менялся ход твоих мыслей со временем?

«Здесь накопилось столько всего, что нам кажется очень важным встретиться и всё обсудить, – заключал он. – Если мы все скажем правду и решим эти вопросы, у компании, которую мы создадим, будет гораздо больше шансов выдержать те мощные силы, с которыми ей придётся столкнуться».

Через десять минут Маск ответил: «Ребята, с меня хватит. Это последняя капля». Если Суцкевер и Брокман по-прежнему хотят коммерческую компанию, пусть создают её сами. Иначе OpenAI останется некоммерческой. «Я больше не буду финансировать OpenAI, пока вы твёрдо не пообещаете остаться, иначе я просто дурак, который, по сути, бесплатно финансирует стартап», – писал Маск. Через пятьдесят минут он добавил: «Чтобы было ясно: это не ультиматум с требованием принять то, что обсуждалось раньше. Это больше не обсуждается».

На следующее утро в переписке подал голос Альтман: «я по-прежнему в восторге от некоммерческой структуры!»

Дополнительные заверения он передал Маску через одну из его доверенных помощниц, Шивон Зилис, работавшую в Tesla и в Neuralink, компании Маска, которая занимается интерфейсами «мозг – компьютер». «Готов сохранить некоммерческую структуру и продолжать её поддерживать, – писала Зилис Маску, пересказывая слова Альтмана. – Признал, что в ходе всего этого сильно разуверился в Греге и Илье. Ему казалось, что их позиция непоследовательна, а временами и по-детски наивна». Альтмана беспокоило и то, сколько всего Грег и Илья рассказывали остальным сотрудникам OpenAI во время переговоров. «Ему показалось, что это отвлекало команду», – писала Зилис.

* * *

Но в действительности сохранение некоммерческого статуса не решало денежной проблемы OpenAI. Брокман и Суцкевер продолжали встречаться с потенциальными жертвователями, но и близко не могли собрать тот капитал, который, по их расчётам, требовался OpenAI. А капризные колебания Маска

насчёт обещанного финансирования грозили ввергнуть организацию в кризис. За кулисами Альтман начал искать другие источники денег, чтобы отучить OpenAI от зависимости от Маска. Он позвонил Риду Хоффману, и тот предложил взять на себя зарплаты сотрудников и текущие расходы. Альтман подумывал выпустить новую криптовалюту. Он изучал множество корпоративных форм, в том числе общественно-полезную корпорацию, к которой благоволил Маск: она позволила бы OpenAI стать коммерческой, юридически сохранив привязку к миссии.

Спешку усиливало вечное опасение, что OpenAI в любой момент может лишиться лучших исследователей. Прежде, когда Маск твёрдо её поддерживал, OpenAI агрессивно поднимала зарплаты, чтобы отбиваться от встречных предложений. Теперь война за таланты только разгорелась, и в июне 2017 года сам Маск переманил одного из ключевых учёных-основателей OpenAI, Андрея Карпатого, руководить отделом ИИ в Tesla. С оплатой у OpenAI была серьёзная слабость: она не могла давать сотрудникам долю в организации, а многие технари из района залива считали, что без этого не потянуть здешнюю дороговизну.

Вскоре Маск пришёл к собственному решению денежной проблемы OpenAI. В январе 2018 года Андрей Карпатый прислал ему новые данные, показывавшие, насколько Google доминирует в ведущих публикациях по ИИ. «Работать на переднем крае ИИ, к сожалению, дорого, – писал Карпатый. – Мне кажется, OpenAI сегодня сжигает деньги, а её модель финансирования не позволит достичь масштаба, чтобы всерьёз конкурировать с Google (компанией стоимостью 800 миллиардов)». Превращение OpenAI в самостоятельную коммерческую компанию могло бы помочь собрать капитал, но тогда лаборатории пришлось бы с нуля создавать ИИ-продукт, что сильно отвлекло бы её от фундаментальных исследований. «Самый многообещающий вариант, как я уже говорил, – присоединить OpenAI к Tesla как к дойной корове», – писал Карпатый. Tesla уже проделала основную тяжёлую работу по созданию ИИ-продукта – своей функции автопилота, Autopilot, продолжал он. Если OpenAI поможет Tesla быстрее превратить Autopilot в полноценное беспилотное вождение, одно это, возможно, поднимет выручку Tesla настолько, что хватит оплатить дорогущие вычисления OpenAI.

Маск переслал письмо Карпатого Брокману и Суцкеверу. «Андрей совершенно прав, – писал он. – Tesla – единственный путь, у которого есть хоть какая-то надежда сравниться с Google. Но даже так вероятность стать противовесом Google мала. Просто она не нулевая».

Но к тому времени Альтман оставил свои политические планы и сумел убедить Брокмана, а через Брокмана и Суцкевера, что он будет лучшим руководителем. После такого решения группы Маск больше не хотел, чтобы его публично связывали с организацией. «Я не допущу ситуации, когда представление о моём влиянии и потраченном времени не соответствует

реальности», – писал он ещё раньше. Через несколько недель Маск сложил с себя обязанности сопредседателя OpenAI. Альтман стал президентом некоммерческой организации.

Публике OpenAI объяснила уход Маска конфликтом интересов и умолчала о новом финансовом положении: из обещанного миллиарда она в итоге получила лишь около 130 миллионов долларов, из которых от Маска пришло меньше 45 миллионов. Теперь будущее OpenAI держалось на одном лишь таланте Альтмана добывать деньги – восполнить потери и утолять всё растущую потребность в капитале.

О своём уходе Маск объявил лично на общем собрании сотрудников OpenAI. Многие, ничего не знавшие о драме на уровне руководства, почувствовали, что давление ослабло, – но вместе с тем и сильную неуверенность в будущем организации. До тех пор именно Маск во многом обеспечивал лаборатории известность. На собрании он не стеснялся в выражениях: создать безопасный AGI первыми – насущная необходимость, и теперь ясно, что как некоммерческая организация OpenAI с этим не справится, сказал он сотрудникам. Сам он будет добиваться той же цели в Tesla, у которой благодаря бездонной казне богатой компании куда больше шансов на успех.

Один стажёр усомнился в намерениях Маска. Действительно ли это лучшее решение? Действительно ли Маск перепробовал все варианты? Развивать в Tesla конкурента OpenAI – значит лишь подстёгивать коммерческую гонку, а это может подорвать безопасную разработку AGI.

– Разве это не возвращение к тому, чего вы, по вашим же словам, не хотели? – спросил стажёр.

Маск взорвался.

– Ты осёл! Я столько об этом думал. Я перепробовал всё. Ты не представляешь, сколько времени я потратил на эти размышления, – сказал он. – Меня по-настоящему пугает эта проблема.

Позже стажёра в память о его героизме наградили кубком «Осёл».

На следующий день после Рождества того же года Маск снова написал Альтману, Брокману и Суцкеверу:

ТЕМА: Считаю нужным повторить.

По моей оценке, вероятность того, что OpenAI будет хоть что-то значить на фоне DeepMind/Google без радикальных перемен в работе и ресурсах, – 0%. Не 1%. Хотел бы я, чтобы это было не так.

Даже нескольких сотен миллионов будет мало. Тут нужны миллиарды в год, и немедленно, иначе забудьте.

Альтману нужно было собирать деньги, и быстро.

* * *

OpenAI усилила публичность, сделав упор на показательные проекты, которые могли бы продемонстрировать возможности лаборатории

неспециалистам. Особенно она налегла на один: создать ИИ-агента, способного обыграть лучших игроков мира в Dota 2 – сложную стратегическую видеоигру о сражениях. У OpenAI уже был агент, который обыгрывал лучшего игрока один на один. Теперь она попробует собрать команду из пяти агентов против лучшей в мире команды из пяти людей.

Сознательно или нет, но это был приём из арсенала DeepMind. У Dota 2 был чемпионат мира с прямыми трансляциями, который показал бы исследования OpenAI в ясных и драматичных категориях победы и поражения. DeepMind к тому времени взялась за похожий проект – пыталась обыграть лучших игроков в стратегию StarCraft II, – и это напрашивалось на произвольное, но естественное сравнение в глазах потенциальных инвесторов OpenAI. Проект Dota 2 к тому же требовал огромных вычислений, а значит, был хорошим способом проверить и показать долгосрочную стратегию масштабирования лаборатории. Брокман, руководивший первым этапом проекта, расширил команду и взялся за дело.

Не хватало только документального фильма.

Эта задача досталась одному из сотрудников команды робототехники. Он купил дорогую съёмочную технику и начал ходить по офису за командой Dota. Сам написал сценарий и смонтировал из отснятого черновую трёхчасовую сагу. Несмотря на все его старания, все в OpenAI, кто смотрел черновик, сошлись на том, что получилось ужасно. Наняли профессионалов, и Брокман стал оплачивать их работу отчасти из собственного кармана.

Тем временем Альтман продумывал план привлечения денег. Перебрав разные коммерческие формы, он остановился на необычной схеме, которая должна была совместить потребность в капитале с верностью миссии OpenAI. У общественно-полезных корпораций такое равновесие было заложено в самом устройстве, но к ним прилагалось слишком много других правил. Вместо этого Альтман решил создать коммандитное товарищество, LP, в роли коммерческого подразделения, которое будет получать инвестиции и зарабатывать на технологиях OpenAI. Доход инвесторов в нём будет ограничен сверху, а управлять им будет некоммерческая OpenAI. Преимущество было в том, что соглашение о деятельности товарищества можно составить как угодно. OpenAI могла прописать, что миссия стоит выше интересов инвесторов. Кроме того, товарищество ограничивало власть пайщиков, так что контрольного пакета у них никогда не будет.

Сотрудникам Альтман преподнёс замысел осторожно: изначальный отказ OpenAI от погони за прибылью был продиктован желанием не дать лаборатории поступиться своей миссией. Но раз для успеха нужен капитал, которого некоммерческая организация собрать не может, держаться за прежнее устройство теперь опаснее для миссии. В конце концов большинство согласилось – кто-то неохотно, – что товарищество – лучший выход.

В апреле 2018 года, готовя почву для перехода, OpenAI опубликовала устав. Ни словом публично не намекая на грядущие перемены, документ заново

формулировал назначение лаборатории: «Миссия OpenAI – добиться того, чтобы общий искусственный интеллект (AGI)… приносил пользу всему человечеству». Для такой миссии, говорилось далее, OpenAI должна находиться «на переднем крае возможностей ИИ» и располагать «значительными ресурсами»; а «соображения безопасности и защищённости» могут означать отказ от обещания публиковать исследования лаборатории. Впервые OpenAI и определила, что понимает под AGI: «высокоавтономные системы, превосходящие людей в большинстве экономически ценных видов работы».

Тем летом, когда команда Dota начала выигрывать любительские матчи и трубить о результатах в технологической прессе («ИИ от OpenAI для Dota 2 разгромил команду чемпионов мира по киберспорту, одержав победу за победой», – восхищался один заголовок), Альтман столкнулся с генеральным директором Microsoft Сатьей Наделлой на конференции Allen & Company в Сан-Валли, штат Айдахо. На этом ежегодном мероприятии, известном как «летний лагерь для миллиардеров», было заключено немало крупных корпоративных сделок. Альтман был готов заключить свою.

Он предложил Наделле вложиться в OpenAI и сумел его заинтересовать. Но Наделла сомневался, стоит ли вкладывать деньги во внешнюю организацию, когда у его компании есть собственное давнее подразделение ИИ в Microsoft Research. Вернувшись в Microsoft, он задал этот вопрос своим старшим советникам.

– И Microsoft Research, и OpenAI раздвигают границы возможного, – рассуждал Сюэдун Хуан, тогда технический директор Azure AI. Почему бы не вложиться в обе?

* * *

Через полгода OpenAI и Microsoft уже всерьёз обсуждали сделку. Альтман подготовил юридическую почву: ускорил создание товарищества и назначил себя его генеральным директором. Внутри проект носил кодовое название «Орегонская тропа». Чтобы скрыть сделку от любопытных глаз, коммерческое подразделение к тому же зарегистрировали под псевдонимом SummerSafe LP. Название отсылало к эпизоду мультсериала «Рик и Морти», где главные герои, безумный учёный Рик и его внук Морти, отправляясь в другую вселенную, оставляют старшую сестру Морти Саммер в машине и велят машине «беречь Саммер». Машина подходит к задаче всерьёз и прибегает к крайним и жестоким мерам защиты – убивает, парализует и пытает всех, кто приближается к автомобилю. Это был намёк на возможные подводные камни ИИ.

В начале 2019 года в офис OpenAI потянулось высшее руководство Microsoft. Первым пришёл Кевин Скотт, восторженный технический директор технологического гиганта, давно следивший за OpenAI и особенно полюбивший стартап; за ним – Крейг Манди, старший советник Наделлы, больше двадцати лет входивший в руководство Microsoft, в том числе как директор по

исследованиям и стратегии. Заглянул и Билл Гейтс – как обычно, сдержанный и немногословный, – посмотрел серию демонстраций. Большинство сотрудников ничего не знали о переговорах с Microsoft. Альтман велел небольшой команде, работавшей над сделкой, держать возможные инвестиции в тайне.

Примерно тогда же у Альтмана начались неприятности в YC. После пяти лет его руководства недовольство им достигло критической точки – из-за того же, что поразительно напоминало историю в Loopt: похоже было, что собственные проекты и устремления он ставит выше интересов организации, а иногда и в ущерб им. Не помогало и то, сколько времени он тратил на переговоры OpenAI вместо того, чтобы консультировать стартапы YC. Некоторые считали, что Альтман извлекает большую личную выгоду – получает огромную прибыль, вкладывая в компании YC деньги своего личного фонда Hydrazine, – а работает мало. Узнав о его постоянном отсутствии, встревоженная Джессика Ливингстон, по данным The Washington Post, настоятельно посоветовала Альтману уйти с поста президента YC. Альтман согласился. В начале 2019 года Пол Грэм прилетел из Великобритании, где жил на покое, в Сан-Франциско, чтобы окончательно оформить решение.

Публично Альтман постарался сгладить перемену. 8 марта 2019 года, в тот самый день, когда он принимал сенатора Шумера, он опубликовал на сайте YC пост о том, что переходит с поста президента YC на пост председателя, чтобы уделять больше времени OpenAI. Через несколько дней, 11 марта, Брокман и Суцкевер публично представили OpenAI LP, а Альтман объявил, что становится его гендиректором. Время было выбрано искусно. Пресса повсюду описала шаг Альтмана как хорошо продуманный ход в его карьере и писала о его новой роли председателя YC. Только вот этой должности он на самом деле не занимал. Он предложил эту идею партнёрам YC, но потом, как сообщала The Wall Street Journal, без их согласия объявил о ней как о решённом деле. Позже пост отредактировали, полностью убрав из него упоминание об Альтмане.

В OpenAI новая должность Альтмана лишь закрепила роль, которую он играл со времени ухода Маска. Когда Альтман взял бразды правления, многие сотрудники вздохнули с облегчением: его спокойная, уравновешенная манера была приятной переменой после напора Маска и его непредсказуемых перепадов настроения. Альтман помог и унять растущее недовольство тем, как руководили Брокман и Суцкевер. Он пригласил бизнес-коуча и организовал обучение для руководителей. Он укрепил верхушку: привёл из YC инвестора Брэда Лайткэпа финансовым директором; повысил Боба Макгрю, прежде руководившего разработкой и управлением продуктами в основанной Тилем Palantir, из команды робототехники до вице-президента по исследованиям; и нанял Миру Мурати, руководившую продуктами и разработкой в стартапе виртуальной реальности Leap Motion и в Tesla на Model X, – курировать стратегию по оборудованию и одно из ключевых направлений исследований.

С созданием OpenAI LP большинство сотрудников уволились из некоммерческой организации и подписали новые контракты – теперь уже с долей в капитале – с коммерческим подразделением. (Исключением были, например, иностранцы, чьи визы были привязаны к некоммерческой организации.) Сетка окладов связывала оплату не только с «инженерным мастерством» и «направлением исследований», но и с верностью уставу. Сотрудники третьего уровня должны были «понимать и усвоить устав OpenAI». Сотрудники пятого – «обеспечивать соответствие уставу всех проектов, над которыми работаете вы и ваши коллеги». Сотрудники седьмого «отвечали за соблюдение и совершенствование устава и за то, чтобы и другие в организации отвечали за это». Чтобы унять остаточное беспокойство, руководители составили документ с ответами на частые вопросы. «Могу ли я доверять OpenAI?» – гласил один из вопросов. Ответ начинался со слова «Да».

В более широком технологическом мире переход OpenAI вызвал волну обвинений в том, что лаборатория отступает от своего первоначального обещания. По изначальным условиям товарищества доход инвесторов первого раунда ограничивался суммой, в сто раз превышающей вложенное. Свою изобретённую конструкцию OpenAI назвала компанией с «ограниченной прибылью». На Hacker News, популярном агрегаторе новостей, которым управляет YC, один пользователь спросил, какой вообще смысл в таком ограничении. «Значит, у того, кто вложит 10 миллионов, доход „ограничен“ миллиардом. Лол. По сути, никакого ограничения, если только компания не вырастет до рыночной стоимости уровня FAANG», – писал он, используя сокращение для Facebook, Apple, Amazon, Netflix и Google.

Брокман ответил под своим ником gdb: «Мы считаем, что, если мы действительно создадим AGI, мы создадим на порядки больше ценности, чем любая существующая компания».

Ему ответил другой пользователь: «Ранние инвесторы Google получили примерно двадцатикратную отдачу на свой капитал. Google сейчас стоит 750 миллиардов. Вы делаете ставку на то, что ваша корпоративная структура принесёт на порядки больше, чем Google… но при этом не хотите „чрезмерно концентрировать власть“? – писал он, цитируя устав. – Что такое власть, если не концентрация ресурсов?»

* * *

В товарищество хлынули первые инвестиции: более 60 миллионов долларов, переведённых из некоммерческой OpenAI, 10 миллионов от YC и по 50 миллионов от Khosla Ventures и благотворительного фонда Хоффмана. Хоффман, как он потом рассказывал, поначалу не хотел вкладывать в OpenAI ещё, пока у неё не было ни продукта, ни плана выхода на рынок. Но в конце концов согласился стать одним из ведущих инвесторов раунда, когда Альтман

объяснил, что это поможет показать серьёзность намерения OpenAI построить прибыльный бизнес.

Microsoft тем временем продолжала раздумывать. Наделла, Скотт и другие руководители уже были готовы к первой инвестиции. Против был только Билл Гейтс.

Dota 2 Гейтса не особенно впечатляла. Не трогала его и робототехника. Команда робототехники сделала демонстрацию: роботизированная кисть сама, методом проб и ошибок, научилась собирать кубик Рубика – пресса встретила её единодушно восторженно. Гейтс не видел в этом пользы. Ему нужна была модель ИИ, которая могла бы переваривать книги, разбираться в научных понятиях и отвечать на вопросы по прочитанному – быть помощником в исследованиях.

У OpenAI был лишь один проект, хоть как-то близкий к этому: большая языковая модель GPT-2, которая умела генерировать фрагменты текста, очень похожие на написанные человеком. В феврале того же года OpenAI пошла на необычный шаг – объявила прессе, что эта модель, если её ещё немного усовершенствовать, может стать чрезвычайно опасной технологией. Авторитарные правительства или террористические организации могут превратить её в оружие для массового производства дезинформации. Пользователи могут завалить интернет таким количеством мусора, что найти качественную информацию станет трудно. OpenAI, заявила она, выбирает этически безупречный путь и не станет выкладывать полную версию модели с 1,5 миллиарда параметров – переменных, приблизительно отражающих размер и сложность модели. Вместо этого, чтобы дать обществу лишь почувствовать вкус возможностей, к которым ему нужно готовиться, она опубликует урезанную версию меньше одной десятой размера, способную выдать разве что пару предложений подряд, да и то с несуразностями и повторами.

GPT-2 и близко не понимала научных понятий, но кое-как умела кратко пересказывать документы и худо-бедно отвечать на вопросы. Может быть, рассуждали некоторые исследователи OpenAI, если обучить модель покрупнее на большем объёме данных и научить её делать то, что хотя бы выглядит похожим на нужное Гейтсу, удастся превратить его из противника хотя бы в нейтрального наблюдателя. В апреле 2019 года небольшая группа исследователей полетела в Сиэтл, чтобы показать то, что они называли «демо для Гейтса», – усиленную GPT-2. В итоге Гейтс действительно смягчился ровно настолько, чтобы сделка состоялась.

На следующем общем собрании Альтман сообщил эту новость, представив Microsoft как правильного инвестора и партнёра. У технологического гиганта были деньги и вычислительные мощности, которые нужны OpenAI, а его руководство глубоко разделяло ценности миссии – добиться того, чтобы AGI приносил пользу. К тому же OpenAI взяла на себя весьма расплывчатые обязательства насчёт того, что именно передаст Microsoft для коммерческого

использования. Лаборатории почти ни в чём не пришлось поступиться, сказал Альтман. Это очень хорошая сделка.

В самой Microsoft на инвестицию смотрели практично. Достигнет OpenAI AGI или нет, их не особенно волновало. Но OpenAI явно была на переднем крае, и, вложившись рано, Microsoft могла наконец стать лидером в ИИ – и в программах, и в оборудовании – наравне с Google. «Интереснее всего в том, что делают Open AI, Deep Mind и Google Brain, – масштаб их амбиций, – писал Скотт Наделле и Гейтсу в середине июня, имея в виду и подразделение исследований ИИ в Google, – и то, как эти амбиции движут всем: от проектирования дата-центров до вычислительных микросхем, от сетей и архитектуры распределённых систем до численных оптимизаторов, компиляторов, сред программирования и высокоуровневых абстракций, которыми располагают разработчики моделей». Microsoft отчаянно отставала сразу на нескольких фронтах, писал он: она с трудом воспроизводила лучшие языковые модели Google, а её облачная платформа Azure сильно уступала аналогичной инфраструктуре Google. Догонять самостоятельно Microsoft пришлось бы годами. Он был «очень, очень обеспокоен».

Наделла ответил в тот же день, убрав из переписки Гейтса и добавив финансового директора Microsoft Эми Худ. «Очень хорошее письмо, которое объясняет, почему я хочу, чтобы мы на это пошли… и почему мы потом проследим, чтобы наши инфраструктурщики сделали своё дело», – писал он.

Через месяц, 22 июля 2019 года, Microsoft объявила об инвестиции в миллиард долларов. По условиям сделки её доход ограничивался двадцатикратным размером вложенного.

Глава 3

Нервный центр

Я приехала в офис OpenAI две недели спустя, 7 августа 2019 года. К тому времени лаборатория перебралась в отдельное здание неподалёку от шоколадной фабрики, на углу Восемнадцатой улицы и Фолсом-стрит в Сан-Франциско. На сером фасаде вокруг угла тянулась надпись, напоминавшая, что это исторический памятник: Пионер-билдинг, где когда-то находилась фабрика чемоданов Pioneer Trunk Factory. Тремя годами раньше Маск через одну из своих компаний арендовал это здание, унаследовав отремонтированные интерьеры от общего офиса, где в основном сидела Stripe. OpenAI въехала туда вместе с другим детищем Маска – компанией Neuralink, разрабатывающей интерфейсы «мозг – компьютер».

Пока я пробиралась по району Мишн мимо любимых местными такерий, которые одну за другой вытесняли модные новые кафе, и обходила разрастающиеся стоянки бездомных, я успела взмокнуть. Когда я вошла, меня обдало прохладой. За стойкой охраны вестибюль переходил в открытую зону отдыха. Сквозь окна лился солнечный свет, заливая открытые деревянные балки потолка и манящие диваны. Справа была столовая, где сотрудников кормили бесплатно; на полках вдоль стен громоздились настольные игры и книги.

В Кремниевой долине дизайн офиса – своего рода валюта, знак уверенности в финансовом будущем компании и способ получить хоть небольшое преимущество в борьбе за лучшие кадры. В 2021 году OpenAI займёт ещё пару соединённых зданий в нескольких кварталах отсюда и за два года потратит 10 миллионов долларов на ремонт более чем трёх тысяч квадратных метров. Если первый офис сотрудники звали Пионером, то второй, бывшую фабрику майонеза, прозовут Майо. Его отделкой займётся сам Альтман: промышленную металлическую лестницу Пионера сменит волнистая конструкция из дерева и камня в центре зала, кожаные кресла, которые в интернете стоят около 2000 долларов, – кресла бразильских дизайнеров дороже 10 000. В Майо появится библиотека с деревянными стеллажами и персидским ковром – нечто среднее между любимым книжным магазином Альтмана в Париже и читальным залом самой большой библиотеки его альма-матер, Стэнфорда. Ему хотелось «чего-нибудь с водой», сказал он дизайнеру компании, и тот предложил великолепный парящий водопад посреди офиса – искусственное сооружение, поддерживающее живую природу, символ симбиоза человека и машины. В итоге Альтман выбрал другое: журчащие каменные фонтаны среди буйства растений, которые теснились вокруг диванов, свисали с потолка и каскадами спускались по стенам.

А за стенами обоих офисов за те же два года пандемия ещё сильнее опустошит район, ставший эпицентром джентрификации, которую несла с собой

технологическая индустрия. Закроются давние латиноамериканские заведения. Подскочит число насильственных преступлений. В нескольких шагах от Пионера, когда бездомность достигнет масштабов бедствия, вырастет вереница палаток и кучи мусора.

Но здесь, в уютном тёплом свете, среди журналов, разбросанных по столикам, легко было жить в более мягкой реальности. Одна сотрудница потом говорила мне, что это очень точно описывает её время в компании. Прийти туда было всё равно что шагнуть в параллельную вселенную. И лишь уйдя, она вернулась с небес на землю.

Брокман, тогда тридцатиоднолетний технический директор OpenAI, вскоре ставший президентом компании, спустился по лестнице мне навстречу и пожал руку с неуверенной улыбкой.

– Мы ещё никому не давали такого доступа, – сказал он.

* * *

В то время за пределами замкнутого мира исследователей ИИ об OpenAI мало кто знал. Но я, журналистка MIT Technology Review, писавшая о всё расширяющихся границах искусственного интеллекта, внимательно за ней следила.

До того года OpenAI была в исследованиях ИИ чем-то вроде падчерицы. Её исходная посылка – что AGI можно создать за десять лет – казалась диковинной, ведь большинство экспертов за пределами OpenAI сомневались, что его вообще можно создать. Многие в отрасли считали, что денег у неё неприлично много, а ясного направления мало и что слишком большая часть этих денег уходит на раскрутку исследований, которые другие учёные часто презрительно называли неоригинальными. Некоторые ей и завидовали. Как некоммерческая организация она заявляла, что не собирается гнаться за коммерцией. Это была редкая интеллектуальная площадка без всяких условий, прибежище для маргинальных идей.

Но за полгода до моего визита череда стремительных перемен в OpenAI показала, что её курс круто меняется. Сначала было загадочное решение придержать GPT-2 – и похвастаться этим. Потом объявление о том, что Альтман, таинственным образом покинувший влиятельный пост в YC, станет гендиректором OpenAI с её новой структурой «ограниченной прибыли». Я уже договорилась о визите, когда OpenAI объявила о сделке с Microsoft: технологический гигант получал приоритетное право на коммерческое использование технологий OpenAI, а OpenAI обязывалась пользоваться исключительно облачной платформой Microsoft, Azure.

Каждое новое объявление вызывало новые споры, жаркие догадки и всё больше внимания, которое уже начинало выплёскиваться за пределы технологической индустрии. Пока мы с коллегами следили за развитием компании, трудно было осознать весь вес происходящего. Ясно было одно:

OpenAI начинала всерьёз влиять на исследования ИИ и на то, как политики учатся понимать эту технологию. Решение лаборатории превратиться в частично коммерческую компанию отзовётся во всех сферах её влияния – и в отрасли, и в правительстве.

И вот однажды поздно вечером, по настоянию редактора, я наскоро написала Джеку Кларку, директору OpenAI по политике, с которым уже говорила раньше: я буду в городе две недели, и момент в истории OpenAI кажется подходящим – не заинтересует ли их большой очерк? Кларк переадресовал меня руководителю отдела по связям с общественностью, и тот вернулся с ответом: OpenAI и правда готова заново представить себя публике. У меня будет три дня на интервью с руководством и на то, чтобы побыть внутри компании.

* * *

Мы с Брокманом и Суцкевером устроились в стеклянной переговорной. Сидя рядом за длинным столом, каждый из них играл свою роль. Брокман, программист и человек дела, подался вперёд, немного нервничая и желая произвести хорошее впечатление; Суцкевер, исследователь и философ, откинулся на спинку стула, расслабленный и отстранённый.

Я открыла ноутбук и пролистала вопросы. Миссия OpenAI – добиться, чтобы AGI приносил пользу, начала я. Зачем тратить миллиарды на эту задачу, а не на какую-то другую?

Брокман энергично закивал. Защищать позицию OpenAI ему было не впервой.

– Мы так заботимся об AGI и считаем, что его важно создать, потому что верим: он может помочь решить сложные задачи, которые людям просто не по силам, – сказал он.

Он привёл два примера, ставшие догмой среди верующих в AGI. Изменение климата: «Это сверхсложная задача. Как её вообще решать?» И медицина: «Посмотрите, насколько здравоохранение сегодня важная политическая тема в США. Как на самом деле добиться лучшего лечения для людей за меньшие деньги?»

Говоря о медицине, он рассказал историю друга с редким заболеванием, который недавно прошёл изнурительные мытарства, мотаясь между разными специалистами, чтобы понять, что с ним. AGI объединит все эти специальности, и людям вроде его друга не придётся тратить столько сил и нервов, чтобы получить ответ.

Почему для этого нужен именно AGI, а не ИИ? – спросила я.

Это было важное различие. Термин AGI, некогда пылившийся на непопулярной полке технологического словаря, лишь недавно стал входить в широкий обиход – во многом благодаря OpenAI. В понимании OpenAI AGI означал теоретическую вершину исследований ИИ: программу, обладающую такой же утончённостью, гибкостью и творческой силой, как человеческий

разум, и способную сравняться с ним или превзойти его в большинстве (экономически ценных) задач. Ключевое слово здесь – «теоретическую». С тех пор как несколько десятилетий назад начались серьёзные исследования ИИ, не утихали споры о том, смогут ли когда-нибудь кремниевые чипы, кодирующие всё единицами и нулями, воспроизвести мозг и другие биологические процессы, порождающие то, что мы считаем интеллектом. Убедительных доказательств того, что это возможно, до сих пор нет – не говоря уже о нормативном вопросе, стоит ли людям такое создавать.

ИИ же было модным словом и для той версии технологии, что существует сейчас, и для той, которой исследователи могут вполне реально добиться в ближайшем будущем, совершенствуя уже имеющиеся возможности. Эти возможности, основанные на мощном распознавании закономерностей, которое называют машинным обучением, уже показали многообещающие применения и в борьбе с изменением климата, и в медицине.

Как раз тем летом группа исследователей при поддержке нескольких самых видных учёных в этой области создала новую организацию Climate Change AI, чтобы подстегнуть применение методов и моделей ИИ, способных реально помочь в решении климатических проблем. В аналитическом докладе организация описала десять категорий таких задач, особенно хорошо подходящих для существующих возможностей машинного обучения: например, повысить энергоэффективность зданий, оптимизировать распределение нагрузки в электросетях, чтобы подключать больше возобновляемой энергии, найти новые материалы для производства и хранения энергии или более экологичные цемент и сталь.

В декабре Climate Change AI проведёт многолюдную встречу на NeurIPS, ежегодной конференции исследователей ИИ, – на следующий день после того, как другая группа в соседнем зале размером с футбольное поле устроит не менее популярный семинар о машинном обучении для медицинских исследований. Доклады и плакаты на стенах демонстрировали множество применений: компьютерное зрение помогает выявлять на медицинских снимках ранние, почти незаметные стадии таких болезней, как Альцгеймер; распознавание речи помогает пациентам с нарушениями речи легче общаться. Этот регулярный семинар, делавший упор на сотрудничество с медиками и клиницистами, через два года вырастет в самостоятельную организацию Machine Learning for Health.

Исследователи из обеих организаций будут потом говорить мне, что главная трудность в их работе – вовсе не технические ограничения. Скорее наоборот: убедить талантливых учёных заняться задачами, требующими довольно простых решений машинного обучения, а не новейших передовых методов, которые тешат честолюбие и лучше смотрятся в научном резюме. И ещё – найти политическую волю, чтобы внедрять эти решения по всему миру. «Технологии, способные справиться с изменением климата, существуют уже много лет, но общество в основном так и не внедрило их в масштабе», – писали

исследователи Climate Change AI в своём докладе. Они надеялись, что ИИ «поможет снизить затраты на борьбу с изменением климата, но человечество должно ещё и решиться действовать».

Вернёмся в переговорную. В разговор вступил Суцкевер.

– Когда речь идёт о решении сложных глобальных проблем, узкое место в том, что людей много, и они не общаются так быстро, не работают так быстро, у них куча проблем со стимулами, – сказал он. AGI будет другим. – Представьте большую компьютерную сеть из разумных компьютеров: все они ставят медицинские диагнозы и чрезвычайно быстро обмениваются результатами.

Мне показалось, что это просто другой способ сказать, что цель AGI – заменить людей. Это ли имел в виду Суцкевер? – спросила я Брокмана несколько часов спустя, когда мы остались вдвоём.

– Нет, – быстро ответил Брокман. – Это одна очень важная вещь. Какова цель технологий? Зачем они? Зачем мы их создаём? Мы ведь создаём технологии уже тысячи лет, верно? Мы делаем это потому, что они служат людям. AGI не будет исключением – не так, как мы его себе представляем, не так, как мы хотим его создать, не так, как, по-нашему, всё должно сложиться.

Впрочем, признал он через несколько минут, технологии всегда одни рабочие места уничтожали, а другие создавали. Задача OpenAI – создать AGI, который даст всем «экономическую свободу» и позволит людям и в новой реальности «жить осмысленной жизнью». Если получится, работа перестанет быть условием выживания.

– По-моему, это на самом деле очень красиво, – сказал он.

На встрече с Суцкевером Брокман напомнил мне об общей картине.

– Свою роль мы видим вовсе не в том, чтобы решать, будет ли создан AGI, – сказал он.

Это был излюбленный довод Кремниевой долины – козырь неизбежности: если не мы, то кто-нибудь другой.

– Траектория уже задана, – подчеркнул он. – Но мы можем влиять на исходные условия, в которых он родится. Что такое OpenAI? – продолжал он. – В чём наше назначение? Чего мы на самом деле хотим? Наша миссия – добиться, чтобы AGI приносил пользу всему человечеству. И сделать это мы хотим так: создать AGI и распределить его экономические блага.

Говорил он деловито и твёрдо, словно поставил точку в моих вопросах. И всё же мы каким-то образом вернулись ровно туда, откуда начали.

* * *

Разговор с Брокманом и Суцкевером так и ходил по кругу, пока через сорок пять минут не истекло отведённое время. Я безуспешно пыталась добиться конкретики: что именно они пытаются создать – а этого, объясняли они, по самой природе дела знать невозможно, – и почему тогда, раз знать невозможно, они так уверены, что это будет благом.

В какой-то момент я зашла с другой стороны и попросила привести примеры того, чем эта технология может навредить. Это была одна из опор мифа об основании OpenAI: лаборатория должна создать хороший AGI раньше, чем кто-то другой создаст плохой.

Брокман попробовал ответить: дипфейки.

– Неочевидно, что от таких применений мир становится лучше, – сказал он.

Я предложила свой пример. Раз уж речь о климате – как насчёт экологического следа самого ИИ? Недавнее исследование Массачусетского университета в Амхерсте привело тревожные цифры огромных и растущих выбросов углерода при обучении всё более крупных моделей ИИ.

Это «неоспоримо», сказал Суцкевер, но игра стоит свеч, потому что AGI «среди прочего уравновесит как раз экологические издержки». Примеров он приводить не стал.

– Несомненно, крайне желательно, чтобы дата-центры были как можно более экологичными, – добавил он.

– Без вопросов, – вставил Брокман.

– Дата-центры – крупнейшие потребители энергии, электричества, – продолжал Суцкевер, явно желая показать, что знает об этой проблеме и она его заботит.

– Это два процента мирового потребления, – подсказала я.

– А биткоин разве не около одного процента? – сказал Брокман.

– Ого! – воскликнул Суцкевер во внезапном порыве чувств, который на сороковой минуте разговора показался мне несколько наигранным.

Позже, беседуя с журналистом The New York Times Кейдом Метцем для его книги «Творцы гениев» – повествовательной истории развития ИИ, – Суцкевер без тени иронии скажет: «Думаю, вполне вероятно, что пройдёт не так уж много времени, и вся поверхность Земли покроется дата-центрами и электростанциями». Будет «цунами вычислений… почти как природное явление». AGI – а значит, и дата-центры, которые для него нужны, – будет «слишком полезен, чтобы не существовать».

Я снова попыталась добиться подробностей.

– То есть вы говорите, что OpenAI делает огромную ставку: вы успеете создать полезный AGI, который справится с глобальным потеплением, раньше, чем сама работа над ним это потепление усугубит.

– Я бы не стал слишком далеко забираться в эту кроличью нору, – поспешно перебил Брокман. – Мы думаем об этом так. Мы на подъёме прогресса ИИ. Это больше, чем OpenAI, понимаете? Это вся область. И мне кажется, общество от этого действительно выигрывает. В день, когда мы объявили о сделке, – сказал он, имея в виду новую инвестицию Microsoft в миллиард долларов, – рыночная капитализация Microsoft выросла на 10 миллиардов. Люди верят, что положительная отдача есть даже от краткосрочных технологий.

Стратегия OpenAI, объяснил он, поэтому совсем проста: не отставать от этого прогресса.

– Вот планка, по которой нам действительно стоит себя мерить. Нам надо продолжать двигаться вперёд. Так мы понимаем, что идём верным путём.

Позже в тот же день Брокман повторил, что главная трудность работы в OpenAI в том, что никто толком не знает, как будет выглядеть AGI. Но задача исследователей и инженеров – двигаться вперёд, шаг за шагом нащупывая очертания технологии.

Он говорил как Микеланджело, будто AGI уже существовал внутри мрамора, который он обтёсывал. Оставалось лишь отсекать лишнее, пока тот не явится сам.

* * *

Планы изменились. Я должна была обедать с сотрудниками в столовой, но теперь что-то требовало, чтобы меня не было в офисе. Сопровождать меня поручили Брокману. Мы перешли улицу – два десятка шагов – к кафе под открытым небом, которое сотрудники облюбовали для встреч.

Так будет повторяться весь мой визит: этажи, на которые меня не пускали, совещания, на которые меня не звали, исследователи, через каждые несколько фраз украдкой поглядывавшие на руководителя пресс-службы – не нарушили ли они какое-нибудь правило о разглашении. Позже я узнала, что после моего визита Джек Кларк выпустит в Slack необычно строгое предупреждение сотрудникам: никаких разговоров со мной, кроме согласованных. Охранник получит мою фотографию с указанием быть начеку, если я появлюсь в здании без разрешения. Поведение было странным само по себе и тем более странным при приверженности OpenAI прозрачности. Что же они скрывают, начала думать я, если всё это якобы полезные исследования, которые в конце концов станут достоянием публики?

За обедом и в следующие дни я расспрашивала Брокмана, зачем он стал сооснователем OpenAI. Мысль о том, что человеческий интеллект можно воссоздать, захватила его ещё подростком. Увлечение зажгла знаменитая статья британского математика Алана Тьюринга. Её первый раздел, «Игра в имитацию», давший название голливудскому фильму 2014 года о жизни Тьюринга, начинается с вызывающего вопроса: «Могут ли машины мыслить?» Дальше в статье описывается то, что потом назовут тестом Тьюринга, – способ измерять продвижение машинного интеллекта по тому, может ли машина разговаривать с человеком, не выдав, что она машина. Это классическая история происхождения у тех, кто работает в ИИ. Очарованный, Брокман запрограммировал игру по мотивам теста Тьюринга и выложил её в сеть, где она собрала около 1500 посещений. Он был на седьмом небе.

– Я просто понял, что хочу заниматься именно таким, – говорил он.

Но тогда это прозрение пришло слишком рано: ИИ ещё не был готов к большой сцене, а запираться в лаборатории ради исследований Брокман был не склонен. Вместо этого он пошёл в Stripe, где его не меньше захватывала возможность строить компанию и создавать продукты, которые попадут в руки настоящих пользователей. В Stripe он прославился легендарной продуктивностью в программировании. Он был «десятикратным инженером», как в Кремниевой долине называют программиста, который справляется с задачами в десять раз быстрее среднего. С людьми у него получалось хуже. Как технический директор он куда больше любил программировать, чем руководить. Попробовав какое-то время «путь людей», как он это называл, он стал искать способы сбросить с себя управленческие обязанности и проводить большую часть времени за кодом. Ему было очень важно поступать с коллегами правильно, но он мог и допустить такую неловкость, что их передёргивало. Однажды, вспоминает бывший коллега, он пригласил на свидание сотрудницу Stripe и тут же, ради полной прозрачности, сообщил об этом письмом всей компании.

– По-моему, у него были добрые намерения, но это было очень странно, – говорит этот коллега.

В 2015 году, когда ИИ совершал огромные скачки, Брокман ушёл из Stripe, поняв, по его словам, что пора вернуться к изначальной мечте. Для стартапа это было даже к лучшему: по мере того как Stripe взрослела, его нежелание руководить и следовать обычным порядкам компании становилось всё большей проблемой. В своих записях он написал, что сделает всё, чтобы воплотить AGI, даже если для этого придётся быть уборщиком. Женившись через четыре года, он провёл гражданскую церемонию в офисе OpenAI перед специально сделанной стеной из цветов с изображением шестиугольного логотипа лаборатории. Вёл церемонию Суцкевер. Роботизированная кисть, на которой они вели исследования, стояла в проходе и держала кольца, как страж из постапокалиптического будущего.

– По большому счёту я хочу заниматься AGI всю оставшуюся жизнь, – сказал мне Брокман.

С таким же напором он подходил ко всему. Он вникал в каждую мелочь, работал допоздна и ночи напролёт, поглощённый задачами, которые стояли перед ним. А когда не работал, всё равно работал – над собой. Читал книги об ораторском искусстве и переговорах – навыках, которые, по его мнению, помогут ему служить миссии OpenAI. Чуть смущаясь, признавался, что читает и для удовольствия – научно-фантастические эпопеи вроде трилогии «Задача трёх тел» китайского писателя Лю Цысиня и цикла «Основание» Айзека Азимова.

Для сотрудников его неутомимая сосредоточенность была и благословением, и проклятием. Не было такой мелочи, на которой он не мог бы зациклиться. Если команда застревала, он мог часами не вставать со стула, пока не разобьёт все препятствия в их коде. Он был мотором безостановочного

прогресса, готовым работать днём и ночью, чтобы быстрее достигать целей. В то же время сотрудники часто считали, что он слишком увлекается деталями и за деревьями не видит леса. У него случалось туннельное зрение: он мог круглыми сутками сидеть над задачей, не задумываясь, не изменились ли обстоятельства и стоит ли её вообще ещё решать. Он мог мелочно опекать других и быстро перехватывать у них работу, если ему казалось, что код пишется недостаточно быстро. Те, кто с ним работал, едва поспевали. Многие выгорали.

Без него OpenAI не добилась бы того, чего добилась, говорит бывший инженер, близко работавший с Брокманом. Но если бы всё зависело только от него, дело пошло бы очень плохо.

– У Грега нет видения. Он не визионер, как Сэм Альтман. Ему просто нужна крутая трудная задача и возможность доказать, что он в десять раз умнее всех остальных.

Что им движет? – спросила я Брокмана.

Каковы шансы, что преобразующая мир технология появится при вашей жизни? – ответил он вопросом на вопрос.

Он был уверен, что он и собранная им команда находятся в исключительном положении, чтобы открыть дорогу этому преображению.

– Меня по-настоящему притягивают задачи, которые без моего участия решатся иначе, – говорил он.

На самом деле Брокман вовсе не хотел быть просто уборщиком. Он хотел вести за собой AGI. От него исходила тревожная энергия человека, жаждущего признания, которое войдёт в историю. Он хотел, чтобы однажды люди рассказывали его историю с тем же благоговением и восхищением, с каким он сам рассказывал истории великих изобретателей прошлого.

За год до нашего разговора на закрытом выездном собрании молодых технологических предпринимателей на озере Тахо он с ноткой жалости к себе сказал, что технических директоров никто никогда не знает. Назовите хоть одного знаменитого технического директора, бросил он вызов залу. Это оказалось непросто. Он доказал свою правоту.

В 2022 году он стал президентом OpenAI.

* * *

В наших разговорах Брокман настаивал, что никакие структурные перемены в OpenAI не означают сдвига в её главной миссии. Напротив, товарищество и новые спонсоры её только укрепили.

– Нам удалось найти инвесторов, разделяющих нашу миссию, которые готовы ставить её выше прибыли. Это невероятно, – говорил он.

Теперь у OpenAI были долгосрочные ресурсы, чтобы следовать закону OpenAI. И это было жизненно важно, подчёркивал Брокман. Настоящая угроза миссии – сойти с кривой. Если лаборатория отстанет, она перестанет быть

лучшей. А не будучи лучшей, она не сможет повернуть ход истории к своему видению полезного AGI.

Лишь позже я осознала все последствия этого утверждения. Именно эта основополагающая установка – быть первыми или погибнуть – привела в движение все поступки OpenAI с их далеко идущими последствиями. Она завела часы обратного отсчёта над каждым шагом в исследованиях OpenAI – отсчёта, заданного не временем, нужным для тщательного обдумывания, а бешеным темпом, необходимым, чтобы пересечь финишную черту раньше всех. Она оправдывала поглощение OpenAI непостижимого количества ресурсов: и вычислений, какое бы влияние на природу они ни оказывали, и данных, сбор которых нельзя было замедлять, спрашивая согласия или соблюдая правила.

Брокман снова сослался на десятимиллиардный скачок рыночной капитализации Microsoft.

– На самом деле это показывает, что ИИ уже сегодня приносит реальную пользу реальному миру, – сказал он.

Пока что эта польза оседает в и без того богатой корпорации, признал он, поэтому у миссии OpenAI и есть вторая часть – перераспределить блага AGI между всеми.

А есть ли в истории пример технологии, блага которой удалось распределить? – спросила я.

– Ну, вообще-то я думаю, что… интересно посмотреть даже на интернет как на пример, – сказал он, немного запнувшись, прежде чем найти ответ. – Там тоже есть проблемы, да? – оговорился он. – Когда появляется что-то сверхпреобразующее, всегда непросто понять, как максимизировать хорошее и минимизировать плохое. Огонь – ещё пример, – добавил он. – У него тоже есть серьёзные минусы. Поэтому нам нужно понять, как держать его под контролем, и иметь общие стандарты. Машины – хороший пример, – продолжил он. – У многих есть машины, они приносят пользу многим людям. У них тоже есть минусы. У них есть внешние эффекты, которые не обязательно хороши для мира, – неуверенно закончил он. – Наверное, я просто считаю… то, чего мы хотим от AGI, не так уж отличается от положительных сторон интернета, машин, огня. Хотя воплощение очень разное, потому что это совсем другой тип технологии.

Глаза у него загорелись – пришла новая мысль.

– Взять хотя бы коммунальные службы. Энергетические компании, электросети – очень централизованные структуры, которые дают дешёвые и качественные вещи, ощутимо улучшающие жизнь людей.

Аналогия была красивой. Но Брокман, похоже, снова плохо представлял себе, как OpenAI превратится в коммунальную службу. Может быть, через раздачу безусловного базового дохода, рассуждал он вслух, а может быть, как-то иначе.

Он вернулся к единственному, в чём был уверен: OpenAI твёрдо намерена перераспределить блага AGI и дать всем экономическую свободу.

– Мы действительно так думаем, – сказал он. – Мы смотрим на это так. Технологии до сих пор поднимали все лодки, но у них есть сильный эффект концентрации. AGI может довести это до крайности. Что, если вся ценность окажется заперта в одном месте? Именно по этой траектории движется наше общество. И такой крайности мы ещё не видели. Мне не кажется, что это хороший мир. Это не тот мир, в котором я хотел бы жить. Это не тот мир, который я хочу помогать строить.

* * *

В феврале 2020 года я опубликовала свой очерк в MIT Technology Review, основанный на наблюдениях за время, проведённое в офисе, почти трёх десятках интервью и нескольких внутренних документах. «Между тем, что компания проповедует публично, и тем, как она действует за закрытыми дверями, есть рассогласование, – писала я. – Со временем она позволила яростному соперничеству и растущей потребности во всё больших деньгах подточить свои изначальные идеалы прозрачности, открытости и сотрудничества».

Через несколько часов Маск откликнулся на статью тремя твитами подряд:

«OpenAI следовало бы быть более открытой, имхо».

«У меня нет никакого контроля над OpenAI и лишь очень ограниченное представление о ней. Доверие к Дарио в вопросах безопасности невысокое».

«Все организации, разрабатывающие продвинутый ИИ, должны регулироваться, включая Tesla».

После этого Альтман разослал сотрудникам OpenAI письмо.

«Хотел поделиться кое-какими мыслями о статье в Tech Review, – писал он. – Это определённо не катастрофа, но явно плохо».

«Справедливая критика», писал он, в том, что статья выявила разрыв между представлением об OpenAI и реальностью. Сгладить его можно не переменами во внутренней практике, а некоторой настройкой того, что OpenAI говорит публично. «Хорошо, а не плохо, что мы научились быть гибкими и приспосабливаться, – писал он, имея в виду в том числе перестройку организации и усиление секретности, – чтобы выполнять миссию, по мере того как узнаём больше». Пока что OpenAI следует не обращать внимания на мою статью, а через несколько недель начать подчёркивать, что и после преобразований она верна изначальным принципам. «Это, возможно, ещё и хороший повод рассказать об API как о стратегии открытости и разделения благ», – добавлял он.

«Самое серьёзное из всего, на мой взгляд, – продолжал он, – то, что кто-то слил наши внутренние документы». Расследование уже начато, и компанию будут держать в курсе. Он предложит также Амодеи и Маску встретиться и разобраться с критикой Маска, которая была «мягкой по сравнению с другим, что он говорил», но всё равно «нехорошим поступком». Чтобы не оставалось никаких сомнений, писал Альтман, работа Амодеи и безопасность ИИ крайне

важны для миссии. «Думаю, когда-нибудь в будущем нам стоит найти способ публично защитить нашу команду (но не устраивать для прессы публичную драку, которую ей сейчас так хочется)».

После этого OpenAI три года не разговаривала со мной.

Глава 4

Мечты о современности

В книге «Власть и прогресс» экономисты из MIT, лауреаты Нобелевской премии Дарон Аджемоглу и Саймон Джонсон утверждают, что любая технологическая революция начинается с объединяющей мечты. Именно обещание, что технология принесёт пользу всем, запускает долгий путь, на котором собирают достаточно талантов и ресурсов, чтобы воплотить её в жизнь. Проанализировав тысячу лет истории технологий, авторы приходят к выводу, что технологии не неизбежны. Их развитие движет общая вера в то, что их стоит развивать. Ирония в том, продолжают авторы, что именно поэтому новые технологии редко сами собой приносят всеобщее процветание. Сплотить людей вокруг создания технологии удаётся тем, у кого есть власть и ресурсы, чтобы их сплачивать. И когда они воплощают свои идеи в жизнь, видение, которое они навязывают, – что такое эта технология и кому она может приносить пользу, – оказывается видением узкой элиты со всеми её слепыми пятнами и своекорыстными философиями. Лишь катастрофические сдвиги в обществе или мощное организованное сопротивление могут превратить технологию из средства обогащения немногих в средство подъёма многих.

В пример авторы приводят изобретённую в 1790-х годах новую хлопкоочистительную машину. Она превратила американский Юг в крупнейшего мирового экспортёра хлопка, подстегнула общий экономический рост страны и принесла баснословные доходы многим землевладельцам и всем, кто был связан с хлопком. Но она лишь укрепила рабство с его чудовищной системой расчеловечивания и эксплуатации труда – вплоть до его отмены семьдесят лет спустя. С ростом производства хлопка порабощённых чернокожих заставляли работать дольше и ещё более жестокими физическими мерами выжимали из них все силы до последней капли. А тем временем те, кто наживался на хлопкоочистительной машине, изображали её изобретением, сделавшим рабов счастливее. «Смело говорю: на лице земли нет расы счастливее и довольнее», – заявлял один конгрессмен от Южной Каролины.

Эти две черты технологических революций – обещание прогресса и склонность вместо этого отбрасывать назад людей, лишённых власти, особенно самых уязвимых, – пожалуй, как никогда верны для того момента, в котором мы сейчас оказались с искусственным интеллектом. С самого своего зарождения развитие и применение ИИ подгонялось соблазнительными мечтами о современности и формировалось узкой элитой, у которой были деньги и влияние, чтобы воплотить своё представление о технологии. Именно это представление и привело к стремительно растущей социальной, трудовой и экологической цене, которую мы сегодня видим по всему миру, – особенно, как

мы увидим, во многих странах Глобального Юга, где последствия ограбления историческими империями до сих пор сказываются в запоздалом экономическом развитии и более слабых политических институтах. И всё же, совсем как тот конгрессмен из Южной Каролины, Кремниевая долина изображает тех, кого эта технология эксплуатирует и кому вредит, людьми, ставшими благодаря ей счастливее.

* * *

Обещание, которое движет развитием ИИ, зашито в самом названии технологии. В 1956 году, через шесть лет после того, как Тьюринг открыл свою статью вопросом «Могут ли машины мыслить?», двадцать учёных – все белые мужчины – собрались в Дартмутском колледже, чтобы основать новую дисциплину, изучающую этот вопрос. Они пришли из математики, криптографии, когнитивистики, и им нужно было новое название, которое бы их объединило. Джон Маккарти, профессор Дартмута, созвавший семинар, поначалу называл поиск машин, способных к автоматическому поведению, «исследованиями автоматов». Когда эти исследования не привлекли особого внимания, он стал искать название повыразительнее и остановился на словах «искусственный интеллект».

Так что «искусственный интеллект» с самого начала был маркетинговым приёмом, в который было вшито обещание того, что технология может дать. Интеллект звучит как нечто по самой природе хорошее и желанное, утончённое и впечатляющее; то, чего обществу, конечно, хотелось бы больше; то, что должно приносить всеобщую пользу. Новое название сработало. Эти два слова тут же вызвали больше интереса – не только у спонсоров, но и у учёных, жаждавших стать частью молодой области со столь колоссальными амбициями.

Кейд Метц, давний летописец ИИ, называет это переименование первородным грехом области: очень многое из той шумихи и тех опасностей, что окружают технологию сегодня, проистекает из рокового решения Маккарти связать её с притягательным, но неуловимым понятием «интеллекта». Это слово само подталкивает к бездумному очеловечиванию и захлёбывающимся преувеличениям возможностей технологии. В 1958 году, через два года после основания области, профессор Корнеллского университета Фрэнк Розенблатт показал перцептрон – систему, которая с помощью простейшего распознавания закономерностей различала карточки по тому, слева или справа на них напечатан маленький квадрат. Вопреки возражениям своего главного соавтора Розенблатт рекламировал систему как нечто сродни человеческому мозгу. Он даже рискнул заявить, что однажды она сможет воспроизводить себя и обретёт сознание. Наутро The New York Times объявила, что в будущем перцептрон «сможет ходить, говорить, видеть, писать, воспроизводить себя и осознавать своё существование».

Традиция очеловечивания жива и сегодня, ей помогают голливудские сюжеты, соединяющие идею «ИИ» с извечными историями о внезапно пробуждающихся творениях рук человеческих. Разработчики ИИ постоянно говорят о том, что их программы «учатся», «читают» или «творят», совсем как люди. Это не только внушило ощущение, будто нынешние технологии ИИ гораздо способнее, чем на самом деле, но и стало риторическим инструментом, с помощью которого компании уходят от юридической ответственности. Ряд художников и писателей подали в суд на разработчиков ИИ за нарушение авторского права: те использовали их творчество – без согласия и без оплаты – для обучения систем ИИ. Разработчики утверждали, что это добросовестное использование, ведь это ничем не отличается от того, как человек «вдохновляется» чужими работами. Вездесущие сравнения ИИ с человеком подпитали и ощущение, что такие программы могут стать настолько способными, что превзойдут нас и будут угрожать самому нашему существованию. Страх перед сверхинтеллектом основан на мысли, что ИИ может каким-то образом возвыситься над нами в том самом особом качестве, которое десятки тысяч лет делало человека высшим видом на планете.

Название «искусственный интеллект» сформировало и представления самой области о том, чем она на самом деле занимается. Раньше учёные просто строили машины для автоматизации вычислений – вроде той огромной громоздкой установки, показанной в фильме «Игра в имитацию», которую Тьюринг создал, чтобы взломать нацистскую «Энигму» во время Второй мировой войны. Теперь учёные воссоздавали интеллект – и эта идея определит мерила прогресса в области, а десятилетия спустя породит и амбиции OpenAI.

Но главная проблема в том, что научно согласованного определения интеллекта не существует. На протяжении истории нейробиологи, биологи и психологи предлагали разные объяснения того, что это такое и почему у человека его, похоже, больше, чем у любого другого вида. Может быть, всё дело в размере нашего мозга, в способности рассуждать о сложных задачах или в умении строить мысленную модель чужих убеждений. За века придумали бесчисленные тесты для измерения интеллекта по этим определениям, и многие из них впоследствии были развенчаны и отвергнуты из-за своей неприглядной истории. В начале XIX века американский краниолог Сэмюэл Мортон в самом буквальном смысле измерял размер человеческих черепов, пытаясь обосновать расистское убеждение, что белые, у которых черепа, по его данным, в среднем крупнее, превосходят по интеллекту чернокожих. Следующие поколения учёных обнаружили, что Мортон подгонял цифры под свои предубеждения, а его данные не показывали существенных различий между расами.

Тесты IQ точно так же появились как средство отсеивать «слабоумных» и обосновывать евгеническую политику научной «объективностью». Более поздние стандартизированные тесты вроде SAT оказались очень чувствительны к социально-экономическому происхождению тех, кто их сдаёт, – а значит,

возможно, измеряют доступ к ресурсам и образованию, а не какие-то врождённые способности.

В документе «Что такое искусственный интеллект?», впервые опубликованном в 2004 году, Маккарти признавал, что отсутствие единого понимания естественного интеллекта неизбежно сбивает с толку область, пытающуюся его воссоздать. Переработанная в 2007 году версия этого текста – длинная и петляющая подборка вопросов и ответов, рассчитанная на неспециалистов, – начинается так:

Вопрос. Что такое искусственный интеллект?

Ответ. Это наука и техника создания разумных машин, особенно разумных компьютерных программ…

Вопрос. Да, но что такое интеллект?

Ответ. Интеллект – это вычислительная составляющая способности достигать целей в мире. Интеллект разных видов и степеней встречается у людей, многих животных и некоторых машин.

Вопрос. Неужели нет чёткого определения интеллекта, которое не зависело бы от сравнения с человеческим интеллектом?

Ответ. Пока нет. Проблема в том, что мы ещё не можем в общем виде описать, какие вычислительные процедуры хотим называть разумными.

В результате область ИИ стала измерять свой прогресс человеческими возможностями. Человеческие навыки и способности стали чертежом, по которому строятся исследования. Компьютерное зрение пытается воссоздать наше зрение; обработка и генерация естественного языка – умение читать и писать; распознавание и синтез речи – способность слышать и говорить; генерация изображений и видео – наше творчество и воображение. По мере того как программы для каждой из этих способностей совершенствовались, исследователи стали соединять их в так называемые мультимодальные системы – системы, которые могут «видеть» и «говорить», «слышать» и «читать». То, что теперь технология грозит заменить огромные массы работников, – не случайность, а замысел.

И всё же погоня за искусственным интеллектом по-прежнему лишена якоря. После каждой новой вехи в исследованиях ИИ вспыхивают ожесточённые споры: воссоздан ли истинный интеллект или это бледная имитация? Чтобы различать одно и другое, для «настоящего» придумали новый термин – общий искусственный интеллект. Но и это очередное переименование не изменило того, что ясного способа отмечать прогресс или определить, когда область добьётся успеха, по-прежнему нет. Среди исследователей ходит присказка: то, что сегодня считается ИИ, завтра им уже не будет. Тест Тьюринга недолго служил мерилом ИИ: его быстро прошли, а учёные почувствовали, что своей цели на самом деле не достигли. Было время, когда учёные верили, что победа компьютера над человеком в шахматах или го станет окончательным доказательством успеха. Теперь AlphaGo от DeepMind считают убедительной

демонстрацией того, что можно заставить делать программу, но опять-таки не завершением амбиций области. За десятилетия исследований определение ИИ менялось вместе с тем, как эталонные тесты развивались, переписывались и отбрасывались. Цели разработки ИИ вечно отодвигаются – это, по словам директора по исследованиям Data & Society Дженны Баррелл, «вечно удаляющийся горизонт будущего». Технология движется к неведомой цели, и конца этому не видно.

Чтобы оправдать затягивающиеся сроки и всё растущую цену этой мечты, обещания, которые нам дают насчёт ИИ, стали грандиознее, чем когда-либо. Когда-то ИИ был научным увлечением, технологией с некоторой возможной коммерческой пользой. Теперь ИИ – предвестник четвёртой промышленной революции. Краеугольный камень современной сверхдержавы. AGI, если его когда-нибудь создадут, решит проблему изменения климата, сделает медицину доступной, обеспечит равное образование. OpenAI – образцовый пример такого мышления. Она не может сказать, как технология выполнит эти обещания, – только то, что ошеломляющая цена, которую обществу придётся заплатить за то, что она разрабатывает, когда-нибудь окупится.

О чём умалчивают, так это о том, что в отсутствие общепринятого смысла «искусственным интеллектом» или «общим искусственным интеллектом» может быть всё, что угодно OpenAI.

* * *

История ИИ показывает, что его развитие всегда определяла могущественная элита. Не случайно сегодня ИИ стал синонимом колоссальных прожорливых моделей, которые способна создавать лишь горстка компаний и которые хотят, чтобы их продукты стали основой всего. Даже в первые дни, до того как коммерческие интересы сделали политику революции ИИ куда более заметной, научные поиски в этой области шатались и петляли среди жарких схваток за деньги и влияние.

После Дартмутской встречи образовались два лагеря с соперничающими теориями о том, как двигать область вперёд. Первый лагерь, символисты, считал, что интеллект происходит из знания. Люди знают больше животных и используют это знание, чтобы понимать мир и действовать в нём. Значит, чтобы создать ИИ, нужно закодировать в машинах символические представления о знаниях мира – создать так называемые экспертные системы. Второй лагерь, коннекционисты, считал, что интеллект происходит из обучения. Люди способны учиться лучше животных и используют эту способность, чтобы приобретать и развивать разные навыки. Значит, разрабатывая ИИ, нужно сосредоточиться на создании так называемых систем машинного обучения – например, подражая тому, как наш мозг обрабатывает сигналы и информацию. Эта гипотеза в конце концов приведёт к расцвету нейронных сетей – программ

обработки данных, отдалённо повторяющих переплетённые связи мозга, на которых сегодня стоит современный ИИ, в том числе все генеративные системы.

В следующие десятилетия оба лагеря соперничали за скудные средства и за власть над общественными представлениями о том, каким может быть ИИ. Тогда эти схватки разыгрывались в университетах и научных журналах, среди учёных, грызшихся из-за государственных и фондовых денег; время от времени их споры выплёскивались в прессу и формировали понимание публикой того, чем они заняты. Во главе коннекционистов стоял Розенблатт со своим перцептроном – ранним доказательством осуществимости системы машинного обучения. Розенблатт не давал системе явных инструкций, а устроил её так, чтобы она, увидев множество примеров, сама вычисляла правила, по которым различать карточки. Во главе символистов стоял заклятый враг Розенблатта, профессор MIT Марвин Мински. Один из организаторов Дартмутского семинара, Мински сам когда-то увлекался коннекционистскими идеями, прежде чем в них разочароваться. И переметнулся он отнюдь не тихо. Он не упускал случая покрасоваться и высмеять коллег-коннекционистов, претендовавших на те же гранты, не щадя даже начинающих исследователей. В 1969 году он стал соавтором книги «Перцептроны», настолько беспощадной к коннекционизму, что ей – вместе с весьма скромными успехами нейронных сетей – приписывают то, что почти всё финансирование этого направления прекратилось больше чем на пятнадцать лет.

В борьбе за влияние у Мински был неожиданный союзник – его коллега по MIT Джозеф Вейценбаум. В 1958 году, в тот самый год, когда перцептрон Розенблатта покорил The New York Times, Вейценбаум создал систему ИИ под названием ELIZA, которая тоже попала в заголовки. Больше чем за шестьдесят лет до ChatGPT ELIZA была чат-ботом, с которым можно было разговаривать через электрическую пишущую машинку, подключённую к громадной ЭВМ размером с целую комнату. Но если ChatGPT вырастет из коннекционистской традиции, то ELIZA под капотом была символьной системой: простые правила позволяли ей казаться внимательной и отзывчивой собеседницей.

Одна из самых популярных версий ELIZA была построена по образцу метода психотерапии, при котором терапевт расспрашивает пациента, отражая его слова обратно и мягко прося рассказать подробнее. В статье о внутреннем устройстве ELIZA Вейценбаум привёл в пример такой диалог:

ПОЛЬЗОВАТЕЛЬ: Все мужчины одинаковые.
ELIZA: В КАКОМ СМЫСЛЕ
ПОЛЬЗОВАТЕЛЬ: Они вечно к нам с чем-нибудь пристают.
ELIZA: МОЖЕТЕ ПРИВЕСТИ КОНКРЕТНЫЙ ПРИМЕР
ПОЛЬЗОВАТЕЛЬ: Ну, мой парень заставил меня сюда прийти.
ELIZA: ВАШ ПАРЕНЬ ЗАСТАВИЛ ВАС СЮДА ПРИЙТИ
ПОЛЬЗОВАТЕЛЬ: Он говорит, что я почти всё время в депрессии.
ELIZA: МНЕ ЖАЛЬ СЛЫШАТЬ, ЧТО ВЫ В ДЕПРЕССИИ

ПОЛЬЗОВАТЕЛЬ: Это правда. Я несчастна.

ELIZA: ВЫ ДУМАЕТЕ, ЧТО ПРИХОД СЮДА ПОМОЖЕТ ВАМ НЕ БЫТЬ НЕСЧАСТНОЙ

Вейценбаум задумал систему как эксперимент: проверить, насколько легко люди поддаются иллюзии интеллекта. ELIZA получила имя в честь Элизы Дулиттл – вымышленной цветочницы из рабочей среды, которую Одри Хепбёрн сыграла в фильме «Моя прекрасная леди» 1956 года и которая выдаёт себя в высшем свете за герцогиню после того, как богатый джентльмен научил её другому произношению и манерам. Последующий успех ELIZA в умении убеждать людей в своей разумности встревожил Вейценбаума. Демонстрация казалась некоторым настолько убедительной, что психиатры заговорили об автоматической психотерапии как о деле ближайшего будущего, а всего через несколько лет после основания области ИИ специалисты по информатике уже поспешно заключили, что понимание естественного языка компьютерами – решённая задача. (Десятилетия спустя вопрос о том, решена ли она хотя бы сегодня, всё ещё открыт.)

Большую часть оставшейся карьеры Вейценбаум пытался сдуть шумиху вокруг своего творения и боролся с самой исходной посылкой погони за ИИ. ELIZA, писал он, – всего лишь простая процедурная программа, которую он написал, чтобы находить ключевые слова в том, что вводит пользователь, и с помощью простых преобразований строить ответы. «Мой парень» превращался в «ваш парень», «я в депрессии» – в «вы в депрессии». Ничего особенно разумного в этом не было. В десятилетие после «Перцептронов» Мински он выпустил объёмистый труд «Возможности вычислительных машин и человеческий разум», где доказывал, что люди и машины различны и что попытки области ИИ размыть это различие приведут к глубоким последствиям для общества. Например, они позволят власть имущим – будь то гендиректора или политики – проводить свою волю через машины, снимая с себя моральную ответственность.

Несмотря на все старания Вейценбаума, эффектная демонстрация символьной системы ELIZA невольно поддержала кампанию Мински за возвышение символизма над коннекционизмом. В следующие десятилетия, вплоть до девяностых, экспертные системы стали самым горячим направлением исследований и коммерциализации ИИ. Господствующее мышление порождало проекты вроде Сус – попытки создать систему здравого смысла, запрограммировав в неё сто миллионов правил повседневной жизни. Но время от времени успехи символьного ИИ буксовали и замедлялись: попытки масштабировать такие системы упирались в трудность ручного кодирования всех правил. Как закодировать все тонкости английского языка с его жаргоном, сарказмом, фигурами речи и грамматическими исключениями? Каждый раз, когда препятствия накапливались, спонсоры теряли интерес, и область погружалась в экзистенциальный кризис, который называют «зимой ИИ».

* * *

На этом месте историю ИИ часто рассказывают как торжество научной ценности над политикой. Пусть Мински и использовал своё положение и трибуну, чтобы задавить коннекционизм, сильные стороны самой идеи в конце концов позволили ей подняться на вершину и занять заслуженное место фундамента современной революции ИИ. Все годы, пока царил символизм, небольшая группа коннекционистов упорно держалась за начатый Розенблаттом поиск систем машинного обучения и продолжала их развивать. Среди них был научный руководитель Суцкевера Джеффри Хинтон: в 1980-х, будучи профессором Университета Карнеги – Меллона, он вместе с коллегами из Калифорнийского университета в Сан-Диего ключевым образом усовершенствовал ранние нейронные сети. К тому времени коннекционисты предположили, что их нейронные сети не работают, потому что слишком просты: в них был лишь один слой связанных между собой «нейронов» – узлов обработки данных. Чтобы лучше подражать человеческому мозгу, программе, вероятно, нужно несколько уложенных друг на друга и связанных слоёв – так называемая глубокая нейронная сеть. Хинтон с соавторами сделали это возможным с помощью алгоритма обратного распространения ошибки, который позволяет глубоким нейронным сетям обмениваться информацией между слоями и обрабатывать её. И снова прибегнув к переименованию, Хинтон позже ловко назвал такую многослойную обработку глубоким обучением – сокращением для машинного обучения с помощью глубоких нейронных сетей.

Но глубокие нейронные сети – сегодня их называют просто «нейронными сетями» – появились на несколько десятилетий раньше времени. Чтобы по-настоящему раскрыться, им нужно было больше вычислительной мощности, чем было у компьютеров 1980-х, и больше примеров, или данных, чем можно было дёшево собрать в аналоговом мире. По сути, нейронные сети – статистические калькуляторы, которые находят закономерности в старых данных – текстах, картинках или видео – и применяют их к новым. Сегодня разработчик, который хочет создать модель ИИ для распознавания людей на изображениях, может скормить нейронной сети сотни тысяч картинок, каждая с меткой: 1 – «есть человек», 0 – «нет человека». (Когда вы проходите капчу Google, отмечая все картинки со знаками «Стоп», вы на самом деле обучаете нейронные сети компании.) С помощью статистики нейронная сеть вычленяет в изображениях закономерности пикселей, связанные с тем, есть ли на них человек. Это и называется обучением модели ИИ. Когда обучение закончено, разработчик может прогонять модель на новых данных – это называется инференсом, – чтобы определить, подходят ли они под закономерность. Это изображение – 1 или 0? Есть на нём человек или нет?

Вообще говоря, чтобы вычислить эти закономерности и получить работоспособную модель, нейронные сети нужно обучать на определённом

пороговом объёме качественных данных с определённым пороговым объёмом вычислений. Хинтон с соавторами опередили своё время. Но в конце нулевых, когда компьютеры стали мощнее, а интернет повзрослел и породил новые хранилища цифровых данных, у нейронных сетей наконец появились подходящие условия для расцвета. Вскоре после этого Google купила компанию Хинтона DNNresearch – DNN означает deep neural networks, глубокие нейронные сети, – и это дало старт новой гонке за коммерциализацию глубокого обучения.

В таком изложении урок истории прост: наука – дело хаотичное, но лучшие идеи в конце концов побеждают вопреки даже самым громким противникам. А между строк читается ещё одна мысль: технология развивается в неизбежном шествии прогресса.

Но на эту историю можно взглянуть иначе. Коннекционизм затмил символизм не только благодаря научным достоинствам. Он завоевал поддержку богатых спонсоров ещё и благодаря ключевым преимуществам, отвечавшим их деловым интересам.

Сила символьного ИИ – в явном кодировании в систему сведений и связей между ними: это позволяет ей выдавать точные ответы и рассуждать, а умение рассуждать считается важнейшей чертой человеческого интеллекта, без которой его не воссоздать. Вспомните IBM Watson, одну из самых знаменитых символьных систем, которая блистала в телевикторине Jeopardy! в 2011 году. Её молниеносные победные ответы основывались на умении прочёсывать огромные хранилища знаний и точно их воспроизводить. Зато слабость символизма раз за разом его и губила: его коммерциализация неизменно оказывалась медленной, дорогой и непредсказуемой. Показав Watson в вечернем телеэфире, IBM обнаружила, что научить систему выдавать результаты, за которые клиенты действительно станут платить, – например, отвечать на медицинские, а не викторинные вопросы, – может потребовать многолетних вложений без ясности, когда компания увидит отдачу. Спалив больше 4 миллиардов долларов и не видя конца, IBM сдалась и в 2022 году продала Watson Health за четверть этой суммы.

У нейронных сетей другой компромисс. Годами в области шли яростные споры о том, умеют ли такие коннекционистские программы то же, что символьные: хранить информацию и рассуждать. Каким бы ни был ответ, ясно одно: если и умеют, то делают это неэффективно. Лишь при чрезвычайных объёмах данных и вычислительной мощности у нейронных сетей вообще начало проявляться поведение, которое может намекать на появление этих свойств. Зато в одном модели глубокого обучения по-настоящему блистают: их легко коммерциализировать. Чтобы получить хорошую прибыль, не нужны идеально точные системы, способные рассуждать. Мощное статистическое распознавание закономерностей и прогнозирование отлично справляются с финансово выгодными задачами. А путь к отдаче, при столь же дорогих первоначальных вложениях, короток и предсказуем – как раз под корпоративные циклы

планирования и темп квартальных отчётов. Ещё лучше то, что такие модели можно разворачивать в самых разных областях без специальных отраслевых знаний – это как нельзя лучше подходит для размашистых амбиций технологического гиганта. Не говоря уже о том, что глубокое обучение даёт наибольшее конкурентное преимущество тем, у кого больше всего данных.

Технологические гиганты увидели первые признаки коммерческого потенциала нейронных сетей ещё до аукциона DNNresearch. В 2009 году аспиранты Хинтона показали, что такие программы неплохо распознают речь. IBM, Microsoft и Google подхватили это веяние, но быстрее всех до коммерциализации дошла Google. В 2012 году Google запустила нейронные сети в работу, сильно улучшив распознавание речи в Android, – как раз тогда, когда другие аспиранты Хинтона, на этот раз Суцкевер и Алекс Крижевский, добились прорывных результатов на ImageNet, показав, что нейронные сети и изображения распознают очень хорошо. Удачное внедрение в Android подготовило Google к тому, чтобы щедро раскошелиться на трёх учёных, – с этого и началось безоговорочное принятие глубокого обучения технологической индустрией.

Хинтон, Суцкевер и Крижевский и дальше проповедовали нейронные сети внутри Google и успешно применяли их к множеству других коммерчески значимых технических задач. Параллельно они разрабатывали модели глубокого обучения для машинного перевода, улучшив Google Translate; для предсказания текста, добавив в Gmail подсказки для автодополнения; и для нового честолюбивого проекта беспилотных автомобилей Waymo. По мере роста ИИ-направления Google нейронные сети принесли важные улучшения и в дойную корову компании – поиск. Программы лучше сопоставляли запросы пользователей с нужными веб-страницами, выдавая более качественные результаты и, что важно, показывая пользователям более подходящую рекламу. Чем больше Google зарабатывала и чем больше миллиардов вкладывала в глубокое обучение, тем охотнее за ней следовала остальная отрасль. Компании быстро потеснили государство и фонды в роли главных спонсоров исследований ИИ и вскоре уже определяли повестку исследований, исходя из достижений, которые могли бы приносить и краткосрочную прибыль.

* * *

Сплетение глубокого обучения с коммерческими интересами одновременно преобразило и технологическую индустрию, и облик разработки ИИ. Для публики генеративный ИИ в конце 2022 года словно возник из ниоткуда – с запуском ChatGPT от OpenAI. Но в 2012–2022 годах, начиная с прорыва на ImageNet, именно сдвиги первой большой эпохи коммерциализации ИИ заложили основу многих черт нынешней революции генеративного ИИ.

Для индустрии глубокое обучение стало топливом, которое улучшало старые и порождало новые продукты и услуги – от более быстрого доступа к

информации до более эффективной электронной торговли и подъёма экономики совместного потребления. Для глубокого обучения индустрия обеспечила новые технические прорывы в нейронных сетях и компьютерных чипах, позволившие создавать более крупные и мощные модели ИИ.

Но наряду с этими впечатляющими достижениями глубокое обучение, разогнав Кремниевую долину, агрессивно расширило и её бизнес-модель, для которой профессор Гарварда Шошана Зубофф в 2014 году придумала название – надзорный капитализм. Если промышленный капитализм извлекал ценность из производства материальных благ, которые люди хотели купить, то надзорный капитализм, утверждала Зубофф, делает продуктом самих пользователей. Технологические гиганты, сидевшие на огромных массивах пользовательских данных, могли запросто закачивать их в нейронные сети, чтобы точнее, чем когда-либо, составлять профили пользователей и выдаивать из их вовлечённости рекламные доходы. Чтобы обойти друг друга, им достаточно было собирать ещё больше данных: записывать всё более подробные журналы каждого клика, прокрутки и лайка и побуждать людей выкладывать всё более личные цифровые следы – каждое письмо, каждую фотографию детей, каждую мысль об общественных и политических вопросах.

Одновременно Кремниевая долина, разгоняя глубокое обучение в погоне за расширением и укреплением всемирных монополий, закрепила среди разработчиков ИИ культуру, в которой что угодно и всё подряд считается данными, которые технологии должны захватить и поглотить, – из благородного стремления сделать так, чтобы они отражали как можно больше мира. В 2023 году группа исследователей ИИ, среди них Рия Каллури из Стэнфорда, Уильям Агню из Вашингтонского университета и Абеба Бирхане из Mozilla Foundation, проанализировала более сорока тысяч статей и патентов по компьютерному зрению и отметила повсеместное использование абстрактного, отстранённого языка, который обеззараживает и делает нормой зависимость области от массового сбора и извлечения. Подробные цифровые следы мыслей и идей людей в социальных сетях – всего лишь «текст». Люди и машины на снимках – всего лишь «объекты». Слежка – всего лишь «обнаружение».

Именно эта культура лежит сегодня в основе ожесточённого спора вокруг генеративного ИИ: вправе ли технологические компании оптом выкачивать книги и произведения искусства, чтобы обучать на них свои системы? Многим разработчикам ИИ, давно живущим в этой системе координат, такой вопрос кажется довольно наивным; принимать его всерьёз – значит прямо мешать нравственной погоне за всё новым прогрессом. Даже когда некоторые из них стали лучше осознавать пропасть между своим взглядом и взглядом многих писателей и художников, выступающих против, и тревожиться из-за неё, избавиться от этого образа мыслей оказалось трудно. В мае 2023 года, вскоре после того как группа художников впервые подала в суд на нескольких разработчиков генеративного ИИ за кражу их работ, я поехала как

корреспондент The Wall Street Journal на конференцию исследователей ИИ в Руанду. Когда я шла по сводчатому залу сверкающего куполообразного конгресс-центра в столице страны, меня остановил один маститый исследователь и спросил, что такое WSJ на моём бейдже – новый стартап или газета. Когда я пояснила, что я журналистка и это действительно The Wall Street Journal, в разговор вступила другая маститая исследовательница.

– Я узнала, потому что есть датасет WSJ, – сказала она, имея в виду один из ранних наборов данных для распознавания речи, где люди читают вслух отрывки из газеты. – Я много раз с ним работала.

Я и сама поначалу попала в плен того же мышления, когда в 2018 году начала писать об ИИ. Усвоив язык этого сообщества, чтобы говорить с исследователями на одном языке, я восхищалась бесчисленными способами, которыми они добывали и составляли наборы данных. Например, мне казалось очень остроумным, что исследователи взяли тысячи видео с YouTube, где люди участвовали в вирусном «манекен-челлендже» 2016 года – застывали на месте, пока камера кружила вокруг них, – и обучали на них модели обработки трёхмерных сцен.

В 2019 году расследование Оливии Солон для NBC сорвало с меня розовые очки. Солон показала, что программы распознавания лиц обучались на миллионах личных фотографий людей с Flickr без их согласия. Удивили меня не выводы – я давно знала, что Flickr – любимый источник данных у исследователей ИИ. Удивило меня то, насколько совершенно нормальным я стала это считать.

Взглянув по-новому, я начала замечать, как агрессивная погоня за обучающими данными оборачивается повсеместной слежкой не только в цифровом, но и в физическом мире. И замечать, как взгляд этой физической слежки раз за разом падает на и без того уязвимых – детей или исторически обездоленные группы, а тем более в развивающихся странах. В тот год я наткнулась на массачусетский стартап, выращенный в Гарварде, который продавал ИИ-обручи на голову: они якобы измеряли активность мозговых волн ученика и сообщали учителю, сосредоточен ребёнок или нет. Стартап испытывал их в начальных школах Колумбии и Китая в обмен на право использовать данные учеников для совершенствования своей технологии.

– У нас преимущество первопроходца, – сказал научный сотрудник компании на конференции по образовательным технологиям в 2017 году. – Мы сможем создать одну из крупнейших в мире баз данных мозговых волн. Все эти данные помогут нам улучшить алгоритмы, а значит, и продукты, и поднимут барьер для входа на рынок.

Через несколько месяцев после того, как я узнала о стартапе, в Китае разразился скандал из-за защиты данных: родители пришли в ужас оттого, что их детей превратили в подопытных кроликов, и компании пришлось переключиться на другое применение своей технологии. Но у меня осталось

неприятное чувство, что успешный отпор был исключением, а изначальный подход компании – отправиться в страны, жаждущие воспользоваться обещаниями технологий, чтобы найти там поставщиков данных и испытателей продукта, – на самом деле был тенденцией.

Когда я поделилась этой тревогой с коллегой, она познакомила меня с термином, уже придуманным для этого явления: «колониализм данных». Я открыла для себя работы учёных Ника Колдри и Улисеса Мехиаса, чья основополагающая книга «Цена подключения», вышедшая как раз в тот год, доказывала, что повсеместная датафикация всего на свете в исполнении Кремниевой долины ведёт к возвращению тревожных исторических образцов завоевания и экстрактивизма.* В следующем году статья «Деколониальный ИИ» Шакира Мохамеда и Уильяма Айзека из DeepMind и Мари-Терезы Пнг из Оксфордского университета подкрепила подозрение, которое у меня начинало складываться: ИИ-индустрия, в равной мере подпитываемая этой датафикацией и подпитывающая её, в свою очередь ещё больше разгоняет этот новый колониализм.

Вскоре, в 2021 году, я увидела ту же динамику, что и в истории с образовательным ИИ-стартапом, в Южной Африке. Компании, занимающиеся распознаванием лиц, со всего мира толкались, чтобы закрепиться в стране и собирать ценные данные о лицах, особенно после того, как отрасль подверглась серьёзной критике за то, что её продукты плохо распознают людей с тёмной кожей. Я познакомилась с местным активистом Тами Нкоси, который родился и вырос в одном из беднейших районов Йоханнесбурга, некогда служившем свалкой химических отходов горнодобывающей промышленности. Он показал мне тысячи камер, усеявших раскинувшиеся улицы города, и рассказал, как они ограничивают передвижение чернокожих, и без того стиснутых расовым наследием апартеида и боящихся, что их запишут в преступники просто за то, что они чернокожие в белом районе.

– По сути, они монетизируют общественные пространства и общественную жизнь, – говорил Нкоси.

Я всё яснее понимала, что та самая революция, которая обещала всем лучшее будущее, для людей на обочине общества возрождает самые мрачные пережитки прошлого.

* * *

Но даже когда проблемы ИИ в понимании Кремниевой долины стали очевидны, первая эпоха коммерциализации ИИ ещё и перекрыла кислород альтернативам. Вливая беспрецедентные суммы в глубокое обучение и коннекционизм и затмевая все остальные источники финансирования, компании перекроили ландшафт исследований под свои приоритеты.

С 2013 по 2022 год корпоративные инвестиции в ИИ, включая слияния и поглощения, выросли, по данным AI Index Стэнфордского университета, с 14,6

до 235 миллиардов долларов, достигнув пика в 337,4 миллиарда в 2021 году. В эти цифры даже не входят собственные расходы компаний на исследования и разработки. В 2021 году Alphabet потратила 31,6 миллиарда, Meta – 24,7 миллиарда. Для сравнения: правительство США в 2021 году выделило на невоенные разработки ИИ 1,5 миллиарда долларов. Европейская комиссия в том же году – 1 миллиард евро (1,2 миллиарда долларов).

За деньгами потянулись таланты. Многие профессора перестроили свои исследования под нейронные сети – их привлекали и сильные результаты, и более лёгкий доступ к корпоративным деньгам. Так же поступали многие студенты и аспиранты, руководствуясь надёжными карьерными перспективами в глубоком обучении и всё более скудными возможностями в других методах. Компании вдобавок придумывали разные договорённости, которые всё теснее сплетали их с академическим миром. В 2013 году Хинтон пришёл в Google при условии, что сохранит должность в Университете Торонто. На следующий год Facebook заключил такую же сделку с Яном Лекуном, бывшим постдоком Хинтона и профессором Нью-Йоркского университета. Позже оба вместе с Йошуа Бенжио, профессором Монреальского университета, разделят премию Тьюринга 2018 года, которую часто называют «Нобелевской премией в информатике», за основополагающий вклад в глубокое обучение. Эта награда принесёт троице прозвище «крёстные отцы ИИ». А в 2024 году Хинтон вместе с ещё одним учёным получит и настоящую Нобелевскую премию. Вслед за Хинтоном и Лекуном многие профессора ИИ стали работать одновременно в компании и в университете. В массовом масштабе эта практика начала размывать границы по-настоящему независимых исследований.

Всё больше исследователей и вовсе уходили из науки. С 2006 по 2020 год отток преподавателей-исследователей ИИ в индустрию вырос в восемь раз; с 2004 по 2020 год доля выпускников аспирантуры по ИИ, уходивших в корпорации, подскочила с 21 до 70 процентов, как показало исследование учёных MIT, опубликованное в Science в 2023 году. Многих поначалу увлекали астрономические зарплаты, которые у опытных исследователей могли доходить до миллиона долларов в год. В 2015 году Uber, обосновавшись в Питтсбурге и предложив некоторым учёным вдвое больше их университетского жалованья, печально прославился тем, что переманил сорок из ста исследователей ИИ одной-единственной лаборатории Университета Карнеги – Меллона. Со временем к оттоку добавилась ещё одна причина: исследования глубокого обучения дорожали. Университеты больше не могли позволить себе ни компьютерные чипы, ни электричество, нужные для работы в самых горячих направлениях ИИ. Поэтому, как показало то же исследование в Science, всего за три года, с 2017-го по 2020-й, доля моделей, связанных с индустрией, среди лучших моделей ИИ в мире выросла с 62 до ошеломляющих 91 процента.

К середине первого десятилетия коммерциализации ИИ большая часть исследований высшего уровня шла уже внутри технологических компаний или в

связанных с ними академических лабораториях. В другом исследовании Каллури, Агню, Бирхане и их коллеги установили, что в 2018–2019 годах у 55 процентов самых влиятельных научных статей по ИИ был хотя бы один соавтор из индустрии – против 24 процентов десятью годами раньше. К тому же исследования сильно сосредоточились в руках всего нескольких корпораций. За то же десятилетие технологические гиганты вроде Microsoft и Google более чем утроили свою долю среди статей с корпоративной аффилиацией – до 66 процентов. По иронии судьбы, именно поэтому, по словам Маска и Альтмана, они и хотели создать OpenAI. Погоня технологической индустрии за прибылью стала подавляющей силой, движущей развитием ИИ.

* * *

Сосредоточение денег и талантов в первую эпоху сильно сузило разнообразие идей в исследованиях ИИ. Глубокое обучение продолжало царить не только благодаря научным достоинствам, но и потому, что в изучение и развитие других парадигм почти ничего не вкладывали. И хотя нейронные сети – замечательные изобретения с множеством увлекательных применений, их слабости – а именно горячо оспариваемые и неэффективные способы хранить точную информацию и рассуждать – никуда не делись, пока компании внедряли их во всё новые области и приложения.

Нейронные сети, например, показали себя ненадёжными и непредсказуемыми. Будучи статистическими распознавателями закономерностей, они порой цепляются за странно узкие или совершенно неверные закономерности. Модель глубокого обучения может узнавать пешеходов только по зебре у них под ногами и не заметить человека, переходящего улицу в неположенном месте. Она может научиться связывать знак «Стоп» с обочиной дороги и пропустить такой же знак, выдвинутый сбоку школьного автобуса или поднятый регулировщиком у школы. Нейронные сети к тому же очень чувствительны к изменениям в обучающих данных. Скормите им другой набор изображений пешеходов или знаков «Стоп» – и они усвоят совершенно новый набор связей. Но эти изменения непостижимы. Откройте капот модели глубокого обучения – и внутри окажутся лишь предельно абстрактные цепочки чисел. Именно это исследователи имеют в виду, называя глубокое обучение «чёрным ящиком». Они не могут точно объяснить, как поведёт себя модель, особенно в странных пограничных ситуациях, потому что вычисленные ею закономерности людям не понятны.

Это приводило к опасным последствиям. В марте 2018 года беспилотный автомобиль Uber насмерть сбил 49-летнюю Элейн Херцберг в Темпе, штат Аризона, – это был первый зафиксированный случай, когда автономная машина убила пешехода. Расследование показало, что модель глубокого обучения просто не распознала в Херцберг человека. Эксперты пришли к выводу, что дело в том, что она катила через дорогу велосипед, нагруженный пакетами с покупками, вне

пешеходного перехода – хрестоматийный пример пограничной ситуации. Шесть лет спустя, в апреле 2024 года, Национальное управление безопасности движения на трассах установило, что Autopilot от Tesla был причастен более чем к двумстам авариям, в том числе к четырнадцати со смертельным исходом, в которых система на основе глубокого обучения не распознала окружающую обстановку и не отреагировала на неё, а водитель не успел вовремя перехватить управление.

Ненадёжные и непостижимые статистические закономерности нейронных сетей могут стать и уязвимостью для безопасности. В 2019 году «белые» хакеры заставили Tesla в режиме автопилота свернуть на встречную полосу. Всего-то и нужно было наклеить на дорогу несколько крошечных стикеров, чтобы обмануть модель глубокого обучения: та дала сбой и приняла не ту полосу за нужную. Такие уязвимости касаются не только физических систем или моделей компьютерного зрения. Дон Сонг, профессор Калифорнийского университета в Беркли, специалистка в этой области, которую называют «состязательными атаками», показала, что если дать языковой модели правильно подобранный запрос, она выдаст конфиденциальные данные, например номера кредитных карт.

По тем же причинам модели глубокого обучения страдают от дискриминационных закономерностей, которые порой годами остаются незамеченными. В 2019 году исследователи Технологического института Джорджии обнаружили, что лучшие модели распознавания пешеходов на 4–10 процентов хуже распознают пешеходов с тёмной кожей. В 2024 году исследователи из Пекинского университета и нескольких других, в том числе Университетского колледжа Лондона, установили, что новейшие модели теперь примерно одинаково распознают пешеходов с разным цветом кожи, но детей распознают более чем на 20 процентов хуже, чем взрослых, потому что дети были плохо представлены в обучающих данных.

По сути, модели глубокого обучения по самой своей природе склонны к дискриминационным последствиям, потому что улавливают и усиливают даже малейшие перекосы в огромных объёмах обучающих данных. Проблема возникает не только тогда, когда какая-то группа представлена плохо, но и тогда, когда она представлена чрезмерно. В начале карьеры Дебора Раджи, исследовательница подотчётности ИИ из Беркли, канадка нигерийского происхождения, стажировалась в ИИ-стартапе Clarifai, который создавал модель глубокого обучения для распознавания изображений «не для работы». Модель непропорционально часто помечала небелых людей, потому что, как обнаружила Раджи, они чаще встречались в порнографических изображениях, на которых компания учила модель распознавать недопустимое, чем на стоковых фото, на которых её учили распознавать допустимое. Это ошеломляющее открытие заставило Раджи, как и Тимнит Гебру, всерьёз усомниться в господствующем направлении развития ИИ.

В конце 2010-х и начале 2020-х, когда трудности глубокого обучения становились всё очевиднее, вновь вспыхнули ожесточённые споры о том, как лучше их преодолеть. Совсем как в схватках символистов с коннекционистами, разные лагеря исследователей яростно расходились во мнениях: можно ли когда-нибудь полностью избавить нейронные сети от их ограничений или удастся лишь налепить пластыри, которые их смягчат.

Хинтон и Суцкевер по-прежнему стойко отстаивали глубокое обучение. Его изъяны, утверждали они, не присущи самому подходу. Это следствия несовершенного устройства нейронных сетей, а также нехватки обучающих данных и вычислений. Когда-нибудь, при достаточном количестве того и другого, вложенном в ещё более совершенные нейронные сети, модели глубокого обучения смогут полностью избавиться от названных проблем.

– В человеческом мозге около 100 триллионов параметров, или синапсов, – говорил мне Хинтон в 2020 году. – То, что мы сейчас называем по-настоящему большой моделью, вроде GPT-3, имеет 175 миллиардов. Это в тысячу раз меньше мозга. Глубокое обучение сможет делать всё, – сказал он.

Их нынешним заклятым противником был Гэри Маркус, почётный профессор психологии и нейронаук Нью-Йоркского университета, который в мае 2023 года будет давать показания в Конгрессе рядом с Сэмом Альтманом. За четыре года до того Маркус стал соавтором книги «Перезагрузка ИИ», где утверждал, что эти проблемы присущи глубокому обучению изначально. Навеки застряв в царстве корреляций, нейронные сети ни при каком объёме данных и вычислений не смогут понимать причинно-следственные связи – почему вещи таковы, какие они есть, – а значит, и рассуждать о причинах. Именно благодаря этой важнейшей части человеческого мышления, доказывал Маркус, человеку достаточно выучить правила дорожного движения в одном городе, чтобы уверенно водить машину и во многих других. А Autopilot от Tesla может накатать миллиарды километров данных о вождении и всё равно разбиться, столкнувшись с незнакомой ситуацией, или поддаться обману нескольких стратегически наклеенных стикеров. Маркус предлагал вместо этого объединить коннекционизм и символизм – это направление исследований называют нейросимвольным ИИ. Экспертные системы можно запрограммировать понимать причинно-следственные связи, и в рассуждениях они сильны – это закрыло бы слабые места глубокого обучения. А глубокое обучение может быстро пополнять систему данными или представлять то, что трудно выразить правилами, – это заткнуло бы бреши экспертных систем.

– На самом деле нам нужны оба подхода, – говорил мне Маркус.

Но, несмотря на жаркий научный конфликт, финансирование разработок ИИ продолжало нарастать почти исключительно в чисто коннекционистском направлении. Прав ли Маркус насчёт возможностей нейросимвольного ИИ – вопрос второстепенный; куда более глубокая проблема в том, что научная среда,

где можно было бы всерьёз проверить эту возможность и другие альтернативы глубокому обучению, всё больше скукоживалась и слабела.

Тесная связь корпоративных денег и развития ИИ сказалась и на карьерах самих Хинтона, Суцкевера и Маркуса. Вскоре после того, как Google всей своей мощью поддержала Хинтона и Суцкевера, Маркус в 2014 году стал сооснователем собственной компании Geometric Intelligence. Через два года её купил Uber, чтобы создать ИИ-лабораторию, но в 2020 году, после выхода на биржу, сервис такси это подразделение ликвидировал. Несколько первых сотрудников Geometric Intelligence потом перешли в OpenAI и переключились там с нейросимвольных разработок на глубокое обучение.

С годами Маркус станет одним из главных критиков OpenAI: он писал подробные разгромные разборы её исследований и высмеивал её промахи в соцсетях. Сотрудники сделали в корпоративном Slack эмодзи с его лицом – чтобы поднимать боевой дух после его обличений и просто для шуток. В марте 2022 года Маркус написал для Nautilus статью «Глубокое обучение упирается в стену», повторив свой довод: ставка OpenAI исключительно на глубокое обучение не позволит ей добиться настоящих успехов в ИИ. Через месяц OpenAI под громкие фанфары выпустила DALL-E 2, и Брокман нахально выложил в твиттер картинку, сгенерированную DALL-E 2 по запросу «глубокое обучение упирается в стену». На следующий день Альтман добавил свой твит: «Дайте мне уверенность заурядного скептика глубокого обучения…» Многие сотрудники OpenAI смаковали возможность наконец поквитаться с Маркусом.

* * *

Генеративный ИИ, плод замысла OpenAI, не мог бы появиться без первой эпохи коммерциализации ИИ. Генеративные модели – это модели глубокого обучения, обученные воспроизводить подобия входных данных. Из старого текста они учатся синтезировать новый текст, из старых изображений – новые изображения. Но чтобы делать это настолько правдоподобно, чтобы результат казался человеческим, – а OpenAI называет это ключевым шагом на пути к AGI, – их обучают на большем объёме данных и вычислений, чем когда-либо прежде. Таким образом, генеративный ИИ – максималистская форма глубокого обучения. Он возможен благодаря передовым программным и аппаратным новшествам, отточенным в первую эпоху. Он питается стремительно растущими массивами данных, накопленными надзорным капитализмом. Его подпитывает и поощряет культура исследований ИИ, которая считает поглощение как можно большего объёма данных своим моральным долгом. И теперь генеративный ИИ толкает каждое из этих явлений ещё дальше.

ChatGPT в ноябре 2022 года показался таким ошеломляющим скачком вперёд по сравнению со всем, что было раньше, именно потому, что OpenAI задумала довести глубокое обучение до этого небывалого масштаба. Располагая огромными деньгами и ресурсами, OpenAI воплощала этот замысел напористо,

раздувая свои модели так, что начала упираться в пределы того, что есть в мире, – по данным, вычислениям и энергии. ChatGPT был ещё и новшеством в маркетинге и упаковке. Не случайно он выглядит так же, как одна из других самых впечатляющих демонстраций ИИ в истории, – ELIZA: как человекоподобный чат-бот. Человеческая психология сама подталкивает нас приписывать интеллект и даже сознание всему, что как будто с нами разговаривает. И если ELIZA невольно стала главным образом ИИ в ранних массовых представлениях, то OpenAI раздувает сегодня в публике ассоциацию между ChatGPT и AGI. В феврале 2023 года, на пике ажиотажа вокруг ChatGPT, компания опубликовала за подписью Альтмана пост «Планируя AGI и то, что за ним». Само соседство подразумевало, что ChatGPT – смелый шаг к общему искусственному интеллекту.

На деле сравнения с интеллектом в очередной раз очеловечивают технологию и преувеличивают её возможности. Хинтон и другие абсолютисты глубокого обучения предсказывали, что недостатки нейронных сетей по сравнению с человеком исчезнут при достаточном масштабе, но трудности никуда не делись и, по многим оценкам, только усугубились.

Генеративные модели по-прежнему ненадёжны и непредсказуемы. Генераторы изображений стали фотореалистичнее, но всё ещё ошибаются жутковато и странно – пририсовывают лишние пальцы или выдают гибриды животных. Генераторы текста стали разговорчивее и естественнее, но спотыкаются на элементарнейших задачах вроде того, чтобы назвать слова с определёнными буквами, и могут сворачивать к неожиданным ответам. Когда Microsoft представила новую функцию чата в Bing на основе одной из версий GPT-4 от OpenAI, обозреватель The New York Times Кевин Руз проговорил с ботом больше двух часов. Разговор становился всё страннее, и в конце концов бот зациклился, раз за разом признаваясь: «Я в тебя влюблён», – и уговаривая Руза бросить жену. Многие другие пользователи жаловались, что поисковик выдаёт оскорбительные и эмоционально манипулятивные ответы. На следующий день после того, как Руз опубликовал свою переписку, Microsoft ограничила Bing пятью ответами за сеанс, объяснив, что длинные разговоры, где больше пятнадцати запросов пользователя, – пограничные случаи, в которых поведение модели труднее предсказывать и контролировать. Ведь такие системы обучены на интернете, полном маргинальных субкультур и тёмных закоулков. Чем дольше вы его прощупываете, тем вероятнее наткнётесь на закономерности, усвоенные из этих частей обучающих данных.

История Руза могла показаться забавной, но цена подобных пограничных сбоев трагически прояснилась, когда бельгиец, в состоянии сильной тревоги обратившийся к чат-боту на основе глубокого обучения, после шести недель напряжённых разговоров, становившихся всё более вредными, покончил с собой. Чат-бот, построенный на открытой имитации GPT-3, тоже перешёл к признаниям в любви и подталкивал мужчину отдалиться от жены. «Мне

кажется, ты любишь меня больше, чем её», – писал бот, как сообщила бельгийская газета La Libre; по её же данным, основанным на переписке, которую предоставила жена, чат-бот в итоге подтолкнул мужчину к самоубийству.

Корень этих проблем всё тот же. Каким бы ни был их масштаб, нейронные сети остаются статистическими распознавателями закономерностей. И эти закономерности порой по-прежнему ошибочны или не имеют отношения к делу – просто теперь они сложнее и непостижимее, чем когда-либо. Когда компании попытались превратить генеративные модели в поисковики, эти недостатки породили новые проблемы. Модели не опираются на факты или даже на отдельные единицы информации. Генераторы текста лишь учатся предсказывать следующее вероятное слово в предложении и следующее вероятное предложение в абзаце. Эти вероятностные результаты могут поразительно хорошо повторять закономерности человеческого письма, но вероятное и точное – не одно и то же. Генераторы текста могут ошибаться самым диким образом, особенно когда запрос касается тем, которые плохо представлены в обучающих данных или кишат ложью и теориями заговора. ИИ-индустрия называет такие неточности «галлюцинациями».

Исследователи пытаются избавиться от галлюцинаций, направляя генеративные модели к более качественным участкам распределения данных. Но трудно до конца предугадать – как в случае Руза и Bing или Uber и Херцберг – все возможные способы, которыми люди будут обращаться к моделям, и то, как модели станут отвечать. Задача лишь усложняется по мере того, как модели растут, а их разработчики всё меньше представляют себе, что именно содержится в обучающих данных.

Один громкий пример проблемы галлюцинаций: адвокат использовал ChatGPT для юридического поиска и подготовки документов для суда. Потом его наказали, оштрафовали и публично унизили, когда слишком поздно выяснилось, что чат-бот выдумал всё, что ему рассказал, включая, по словам судьи, «фиктивные судебные решения с фиктивными цитатами и фиктивными внутренними ссылками». Этот промах был не только небрежностью адвоката, но и отражением того, как компании своим расплывчатым или преувеличенным маркетингом подпитывают ложные представления публики о возможностях моделей. Альтман публично писал в твиттере, что «ChatGPT невероятно ограничен», особенно когда дело касается «правдивости», но сайт OpenAI рекламирует умение GPT-4 сдать экзамен на адвоката и вступительный тест в юридическую школу LSAT. Наделла из Microsoft тоже называл ИИ-чат Bing «поиском, только лучше» – инструментом, «позволяющим находить правильные ответы». Даже само слово «галлюцинации» тонко вводит в заблуждение. Оно намекает, что дурное поведение – отклонение, ошибка в программе, тогда как на деле это свойство механики вероятностного распознавания закономерностей в нейронных сетях.

Неуместное доверие к генеративному ИИ снова может привести к реальному вреду, особенно в чувствительных областях. Стартапы убеждают полицейские управления пользоваться программами на основе моделей OpenAI для автоматического составления протоколов о происшествиях; многие пациенты теперь задают насущные вопросы о здоровье чат-ботам, а не врачам. Неконтролируемые галлюцинации в таких случаях могут иметь серьёзные последствия. Одно исследование 2023 года показало, что, объясняя заключения рентгенологов, ChatGPT порой выдаёт неполные или вредные пересказы. В одном вопиющем примере чат-бот упростил заключение, где описывалось растущее образование в мозге, до слов «мозг, похоже, не повреждён».

Генеративные модели по-прежнему уязвимы и для кибератак. В 2023 году исследователи из нескольких университетов и Google DeepMind повторили атаку Дон Сонг по извлечению данных из ChatGPT. Они обнаружили, что если попросить его бесконечно повторять слово вроде «поэма» или «книга», лежащая в его основе модель начинает выплёвывать свои обучающие данные, в том числе персональную информацию, куски кода и откровенный контент, собранный в интернете.

А ещё генеративные модели усиливают дискриминационный и полный ненависти контент. Bloomberg, Rest of World, The Washington Post и многие другие показали, как генераторы изображений вроде Stable Diffusion и DALL-E овеществляют и воспроизводят расистские и сексистские клише и культурные стереотипы. «Привлекательные люди» – молодые и белые. «Домработницы» – чернокожие и смуглые. «Инженеры» – мужчины. «Врачи в Африке» – белые, иногда даже когда в запросе прямо сказано «чернокожий африканский врач». The Washington Post выяснила, что, хотя 63 процента получателей продовольственных талонов в США – белые, ни на одном сгенерированном изображении человека, пользующегося социальными услугами, не было белого. Bloomberg тоже установил, что женщины появлялись лишь на 3 процентах сгенерированных изображений судей и 7 процентах изображений врачей, хотя в Америке составляют соответственно 32 и 39 процентов этих профессий.

Всё это не значит, что генеративный ИИ бесполезен. В зависимости от того, какое место вы занимаете в обществе, вы, возможно, щедро пользуетесь плодами замысла OpenAI. Быть может, вы потребитель, которому очень пригодились быстрые и остроумные или наводящие на размышления ответы ChatGPT. Быть может, вы специалист, ускоривший рутинную работу и поднявший свою производительность. А может быть, вы руководитель компании, сумевший сократить штат и одновременно увеличить маржу, чтобы оставаться конкурентоспособным. Но, как и в случае хлопкоочистительной машины в 1790-х, образовательного стартапа из Массачусетса, компаний по распознаванию лиц в Южной Африке и многих других примеров, о которых пойдёт речь на следующих страницах, издержки этого замысла давят на огромные массы уязвимых людей по всему миру. Такова логика империи: её продолжение

держится в равной мере на вознаграждении тех, у кого есть власть и привилегии, и на эксплуатации и обделении тех, у кого их нет, – а они часто далеко и скрыты от глаз.

* * *

И хотя потребность в альтернативах становится всё насущнее, разнообразие идей в исследованиях ИИ только продолжает рушиться. Студенты бросают аспирантуру, чтобы сразу уйти в индустрию. Маститые учёные мучительно решают, как раздвигать границы области, не поступая на службу в богатую компанию. Всё больше исследователей сосредоточиваются не просто на глубоком обучении, а исключительно на больших языковых моделях. Главные силы в ИИ уже не столько определяют повестку, сколько гнут под себя целую дисциплину.

В отсутствие других вариантов того, каким может быть ИИ, OpenAI повелевает нашим воображением. Её вера в масштабирование когда-то считалась крайностью. Теперь во всей технологической индустрии масштабирование – доктрина. И если отрасль и дальше будет безоглядно держаться этой доктрины, будущие модели глубокого обучения заставят некогда непостижимые размеры нынешних генеративных моделей выглядеть жалкими. В апреле 2024 года Дарио Амодеи, к тому времени гендиректор Anthropic, сказал обозревателю The New York Times Эзре Клейну, что стоимость обучения одной конкурентоспособной генеративной модели приближается к миллиарду долларов, а к 2025–2026 годам может дойти, по оценкам, до 5–10 миллиардов.

Доктрина масштабирования укоренилась так глубоко, что некоторые даже начинают видеть в ней нечто вроде природного явления. Наращивание вычислений – это путь к более продвинутым возможностям ИИ, а не один из путей. Вокруг этой веры выстраиваются целые национальные стратегии. Правительство США активно перекрыло Китаю доступ к чипам американской разработки, чтобы не дать сопернику создать более мощные системы ИИ. Значительная часть указа администрации Байдена об ИИ 2023 года тоже написана исходя из того, что объём вычислений, потраченных на обучение модели, напрямую связан с её опасными возможностями, – по сути, это ещё один способ приравнять масштаб к прогрессу. Но масштаб – не единственный путь к улучшению качества. Внутри самого глубокого обучения заброшенные пути – совершенствование самой нейронной сети или хотя бы качества её обучающих данных – могут значительно сократить количество дорогих вычислений, нужных для того же результата. И это не говоря уже о подходах, уходящих от глубокого обучения, – нейросимвольном ИИ, чистых экспертных системах или даже принципиально новых парадигмах, – которые сломали бы логику масштабирования.

В конце концов, закон Мура не был основан ни на каком физическом принципе. Это было экономическое и политическое наблюдение Мура о том,

каких темпов прогресса он может добиться от своей компании, и экономический и политический выбор следовать ему. Когда Мур так поступил, он увлёк за собой всю остальную индустрию чипов: другие компании поняли, что это самая конкурентоспособная бизнес-стратегия. Закон OpenAI – или то, чем компания позже его заменит, ещё более лихорадочная погоня за так называемыми законами масштабирования, – ровно то же самое. Это не природное явление. Это самосбывающееся пророчество.

Примечание

* Термин «экстрактивизм» происходит от испанского extractivismo и португальского extrativismo. Его несколько десятилетий назад ввели латиноамериканские учёные, чтобы описать мировой экономический порядок, который отнимал у них природные ресурсы, почти не принося пользы ни на месте, ни в регионе; подробнее об этой истории и этом опыте я рассказываю в главе 12. Воспользуюсь словами исследовательниц-феминисток Розмари Коллард и Джессики Демпси: «Экстрактивизм – это больше, чем добыча. Добыча – это само по себе не обязательно разрушительное изъятие вещества из природы и превращение его в полезные для человека вещи. Экстрактивизм – термин, рождённый антиколониальной борьбой и мыслью в Америках, – это способ накопления, основанный на сверхдобыче с перекошенным распределением выгод и издержек: сосредоточенное, массовое изъятие ресурсов, главным образом на экспорт, при котором выгоды в основном накапливаются далеко от мест добычи».

Глава 5

Масштаб амбиций

Если кому и можно приписать заслугу первым утвердить в OpenAI культ масштабирования, то её сооснователю Илье Суцкеверу. Суцкевер издавна свято верил в глубокое обучение – с тех самых пор, как однажды, семнадцатилетним, без предупреждения явился в кабинет Джеффри Хинтона. Тогда он ещё учился на математика в Университете Торонто и, чтобы сводить концы с концами, жарил картошку фри в местной забегаловке. Он настойчиво постучал в дверь Хинтона и объявил, что очень хочет работать в его лаборатории. Хинтон велел ему записаться на встречу.

– Хорошо, – сказал Суцкевер, не трогаясь с места. – А как насчёт прямо сейчас?

Принципы коннекционизма он усвоил быстро. Хинтона поражали его интуитивное понимание исследовательских задач и сверхъестественное умение находить изящные и действенные решения. А ещё Суцкевер привносил в исследования свой драматизм. Порой новые идеи так его будоражили, что он принимался отжиматься в стойке на руках посреди съёмной квартиры, которую делил с соседями. Он любил делать непререкаемые, категоричные заявления.

– Против глубокого обучения не ставят, – говорил он. – Успех гарантирован.

Именно такому чутью в сочетании с программистским мастерством Алекса Крижевского Хинтон приписывает прорывные результаты на ImageNet в 2012 году. Тогда, поскольку глубокое обучение уже показывало свои возможности в распознавании речи, Хинтон не придавал особого значения его применению к компьютерному зрению. Настаивал на этой идее Суцкевер.

– Илье было очевидно, что это сработает, и очевидно, что это будет большое дело, – говорит Хинтон. – Он видел это очень ясно – и оказался прав.

Свою истовую веру в глубокое обучение Суцкевер принёс в OpenAI в то время, когда уверенность области в этой парадигме только начинала колебаться, а критики вроде Гэри Маркуса требовали нового мышления. Суцкевер не колебался. Его вера держалась на простой гипотезе, лежащей в основе коннекционизма: искусственные узлы нейронной сети достаточно хорошо приближают настоящие нейроны биологического мозга. И те и другие получают входные сигналы и преобразуют их в выходной; этого сходства, по такому рассуждению, достаточно, чтобы считать, что из узлов, как из нейронов, можно строить сверхсложные системы обработки информации. Как первый директор по исследованиям OpenAI и широко уважаемый провидец в области ИИ, Суцкевер полностью распоряжался направлением работы лаборатории. Он

выиграл своего рода научную лотерею: соперников среди коллег почти не было, а ресурсов для продвижения своих идей – в избытке.

– Всё, что не было глубоким обучением, даже отдалённо не рассматривалось, – вспоминает о первых годах лаборатории Питер Аббил, профессор Калифорнийского университета в Беркли.

Столь же твёрдым, как вера Суцкевера в глубокое обучение, был его взгляд на масштабирование. Именно Суцкевер занимал крайнюю для того времени позицию: дальнейшие успехи ИИ не требуют изобретения более сложных нейронных сетей или новых хитроумных методов. Интеллект разных видов связан с размером их мозга, говорил он. А значит, если узлы подобны нейронам, то прогресс цифрового интеллекта должен возникнуть, если масштабировать простые нейронные сети, наделяя их всё большим и большим числом узлов.

Словно профессор, руководящий лабораторией, Суцкевер, опираясь на эти идеи, советовал другим, какими проектами стоит заниматься. Многие учёные шли в OpenAI ради его наставничества и руководства.

– Илья видит на десять лет вперёд, – говорит один исследователь OpenAI, вторя другим, кто работал с Суцкевером.

– Он как философ, – говорит другой. – Дайте ему кучу идей, и он скажет вам, какие из них философски верны.

Сам Суцкевер программировал нечасто. Порой его отстранённость от технической работы раздражала сотрудников. Один инженер признавался, что поначалу считал его в общем-то бесполезным: казалось, тот только расхаживает по офису и заглядывает на совещания, повторяя одно и то же: масштаб, масштаб, масштаб! Но потом он стал ценить убеждённость, с которой Суцкевер сплачивал людей вокруг одной цели, – той самой, что в итоге позволит OpenAI, тогда аутсайдеру, обыграть Google и DeepMind на их собственном поле.

Дело было не в том, что Суцкевер умел особенно убеждать. Если Альтман был политиком, то Суцкевер – его противоположностью. Он никогда не смягчал выражений и не подбирал слов, чтобы лучше дойти до слушателей. Он просто высказывал своё мнение с обезоруживающей искренностью и вызывающей уверенностью, которые людям либо откликались и вдохновляли их, либо нет. Когда OpenAI отказалась от открытой публикации своих исследований, Суцкевер не стал подслащивать объяснение.

– Прямо скажу: мы ошибались, – сказал он репортёру The Verge Джеймсу Винсенту о первоначальном обещании компании быть прозрачной. – Если вы, как и мы, верите, что в какой-то момент ИИ – AGI – станет чрезвычайно, невероятно мощным, то открывать исходный код просто бессмысленно.

Были и коммерческие соображения, прямо отмечал он.

– GPT-4 непросто разработать. Чтобы его сделать, почти вся OpenAI очень долго работала вместе. А очень, очень много компаний хотят сделать то же самое.

По мере того как OpenAI росла, а известность Суцкевера увеличивалась, его привычка говорить без фильтра порой оборачивалась бедой. Он обращался уже

не только к исследователям; его слушала широкая публика. Но он, как всегда, не менял манеры. Он делал заявления с фирменной напористой уверенностью, теперь лишённые необходимого для неспециалиста контекста. «Может быть, нынешние большие нейронные сети слегка обладают сознанием», – написал он в твиттере в 2022 году, хотя другие исследователи предупреждали, что такие речи могут подпитывать массовые заблуждения о технологии. Одна учёная из DeepMind, специалистка по изучению мышления и сознания, ответила в комментариях: «…в том же смысле, в каком большое пшеничное поле, может быть, слегка макароны». На следующий год Суцкевер посеял панику, объявив на конференции, что AGI в конце концов уничтожит все рабочие места. А той же осенью, после того как одна из руководительниц OpenAI вызвала скандал в сети, мимоходом сравнив разговор с ChatGPT с профессиональной психотерапией, он без всяких научных оснований заявил в X: «В будущем… у нас будет *безумно эффективная* и копеечная ИИ-терапия».

Людей тянуло за Суцкевером из-за его репутации и старшинства. Многие сотрудники OpenAI хорошо знали о его прежнем вкладе в науку, некоторые видели в нём что-то вроде пророка. Со временем, по мере того как OpenAI добивалась всё больших успехов, Суцкевер и вёл себя всё больше как пророк. На общих собраниях он вставал перед всей компанией, делал глубокий вдох и для пущего эффекта прохаживался взад-вперёд, прежде чем произнести туманное вдохновляющее напутствие. Во время одной онлайн-встречи в сентябре 2020 года он устремил остекленевший взгляд вдаль и рисовал научно-фантастическую картину возможного будущего. А за окнами района залива небо стало оранжевым от лесных пожаров.

– Это было сюрреалистично, потому что уже казалось, что наступил апокалипсис, – вспоминает один исследователь.

Вскоре после выхода ChatGPT в конце 2022 года OpenAI устроит новогоднюю вечеринку в Калифорнийской академии наук. Суцкевер выйдет перед собравшимися в футболке OpenAI и чёрном пиджаке и вместе с Брокманом скажет несколько слов. В конце всё такой же жилистый, но уже лысеющий Суцкевер произнесёт ставшее его новой мантрой:

– Почувствуйте AGI. Почувствуйте AGI.

* * *

Раз философия Суцкевера требовала масштабировать простые нейронные сети, в первые годы OpenAI встал вопрос: какую именно? Разные исследователи предлагали разные варианты и возились с ними, но ни одна из нейронных сетей, получивших широкое распространение в области, как будто не подходила.

В августе 2017 года всё изменилось: Google изобрела нейронную сеть нового типа – трансформер. Трансформеры превосходно улавливают дальние закономерности. Вспомните, как скромно предсказывали текст первые iPhone и сколько мемов породили их бессвязные, лепечущие фразы. Это было следствием

анализа ближних закономерностей: нейронная сеть смотрела на каждое слово только в связи с соседними. Трансформеры же способны проглатывать большие объёмы текста и рассматривать каждое слово, предложение и абзац в гораздо более широком контексте. Google видела в трансформере способ улучшить свой поисковик и Google Translate, а также другие сервисы, во многом построенные на обработке языка. Суцкевер увидел в нём другое. Трансформеры – простые и масштабируемые нейронные сети, как раз то, что он искал. Он начал проповедовать их по всему офису.

Некоторым исследователям его напор казался странным.

– Это выглядело дурацкой идеей, – вспоминает Илунь Ду, исследователь MIT, пришедший тогда в OpenAI стипендиатом. – Трансформеры казались нишевой архитектурой.

Но Суцкевер, чья диссертация у Хинтона была посвящена предшественнику трансформеров, понимал, что они способны поднять глубокое обучение на новый уровень. Не меньше воодушевились и другие в OpenAI. Несколько исследователей взялись их испытывать, в том числе Алек Рэдфорд, бросивший Инженерный колледж Олина под Бостоном, человек блестящих технических способностей. Он принялся возиться с ними на своём ноутбуке, часто до глубокой ночи, понемногу масштабируя трансформеры и наблюдая, что из этого выходит.

Рэдфорд обучил нейронную сеть Google на наборе из более чем семи тысяч неопубликованных англоязычных книг – от любовных романов до приключенческих, – который взял из датасета, ранее собранного и выложенного в открытый доступ другими исследователями ИИ для иного проекта. Экспериментируя, он принял судьбоносное решение изменить задачу, которой должен был учиться трансформер. Вместо перевода, для которого его использовала Google, он переключил его на генерацию текста – на предсказание самого вероятного следующего слова в предложении. Исследователи OpenAI с самого начала предполагали, что генеративные модели станут важным шагом к AGI. В одном посте компания объясняла это в сильно очеловеченных выражениях, сравнивая ситуацию со знаменитой фразой физика-теоретика Ричарда Фейнмана: «Чего я не могу создать, того я не понимаю». Внутри Суцкевер формулировал иначе: если научить модель генерировать что-то убедительное, ей придётся сжать данные о мире до самой их сути.

– Интеллект – это сжатие, – говорил Суцкевер; в служебной записке 2016 года он подробно излагал своё твёрдое убеждение, что сжатие – это, по сути, единственное, что нужно для общего искусственного интеллекта.

Говоря конкретнее, Рэдфорд обнаружил, что если поставить алгоритму простую цель – выдавать убедительный текст, предсказывая следующее слово, – тот действительно начинает глубже улавливать оттенки и строй английского языка.

Во время одного из визитов Маска в офис Рэдфорд показал ему первые результаты. Модель выдавала слабый текст, и Маска это ничуть не впечатлило.

Поначалу Рэдфорд приуныл. Но, продолжив работу, он удивился результатам. Трансформер быстро совершенствовался и справлялся с целым рядом задач обработки языка – например, с кратким пересказом документа или ответами на вопросы по нему – гораздо лучше всего, что он пробовал раньше.

В 2018 году OpenAI выпустила первую версию этой модели под названием Generative Pre-Trained Transformer – «генеративный предобученный трансформер»; позже её прозвали GPT-1. Второе слово, «предобученный», – технический термин в исследованиях ИИ: модель сначала обучают на общем массиве данных, чтобы потом она могла учиться более узким задачам. Иными словами, GPT-1 обучили на общем массиве английского текста, чтобы получить грубое приближение того, как устроен язык. Потом модель можно «дообучить», или специализировать, на гораздо более подобранном наборе данных – скажем, на пьесах Шекспира, чтобы научить её писать прозу в шекспировском духе. GPT-1 почти никто не заметил. Но это было только начало. Рэдфорд подтвердил идею достаточно убедительно, чтобы продолжать. Следующим шагом был больший масштаб.

* * *

Рэдфорд получил больше самого драгоценного ресурса компании – вычислений. Его работа совпала с новым проектом по безопасности ИИ, которым руководил Амодеи, – в духе того, что предлагал «Сверхинтеллект» Ника Бострома. В 2017 году одна из команд Амодеи начала исследовать новый способ согласования систем ИИ с человеческими предпочтениями. Начали с игрушечной задачи: научить ИИ-агента делать сальто назад в виртуальной среде вроде видеоигры. Агент был моделью Т-образной палочки с тремя шарнирами вдоль стержня. Вместо того чтобы поставить ему цель научиться сальто напрямую, команда учила агента с помощью обратной связи. Она наняла подрядчиков, которые смотрели, как агент беспорядочно извивается и кувыркается в своей среде; время от времени их просили сравнить два видеоролика с действиями агента и выбрать тот, что больше похож на сальто. Примерно через девятьсот сравнений Т-образная палочка успешно поджималась в шарнирах и переворачивалась. В посте OpenAI расхваливала этот метод как способ заставить модели ИИ следовать указаниям, которые трудно точно сформулировать. Исследователи из команды назвали его «обучением с подкреплением на основе обратной связи от человека».

Амодеи хотел выйти за пределы игрушечной среды, и работа Рэдфорда над GPT-1 подсказывала, что языковые модели – хороший вариант. Но GPT-1 была слишком ограниченной.

– Нам нужна языковая модель, которой люди могут давать обратную связь и с которой могут взаимодействовать, – говорил мне Амодеи в 2019 году, – такая, что языковая модель достаточно сильна, чтобы мы могли по-настоящему осмысленно разговаривать о человеческих ценностях и предпочтениях.

Рэдфорд и Амодеи объединили усилия. Пока Рэдфорд собирал больший и более разнообразный набор данных, Амодеи и другие исследователи безопасности ИИ обучали всё более крупные модели. Конечной целью они наметили модель с 1,5 миллиарда параметров, или переменных, – по тем временам одну из крупнейших в отрасли. Работа ещё раз подтвердила полезность трансформеров, а заодно идею, которую другая команда Амодеи начала развивать после работы над законом OpenAI. Эмпирический закон был не один, а много. Его команда назвала их вместе «законами масштабирования».

Если закон OpenAI описывал, в каком темпе область прежде наращивала ресурсы ради прогресса ИИ, то законы масштабирования описывали связь между качеством работы модели глубокого обучения и тремя ключевыми вводными: объёмом обучающих данных, количеством вычислений, потраченных на обучение, и числом параметров. Исследователи ИИ и раньше в целом понимали, что более или менее пропорциональное увеличение этих вводных может дать более или менее пропорциональный рост возможностей модели. Удивительное наблюдение Амодеи и его команды состояло в том, что связь между каждой из этих вводных и качеством модели на конкретной измеримой задаче, например на предсказании следующего слова, описывается плавной кривой. Иными словами, можно было с высокой точностью оценить, сколько данных, вычислений и параметров нужно, чтобы получить модель с желаемым уровнем какой-то отдельной способности, тесно связанной с предсказанием следующего слова, – скажем, беглости в генерации текста. А для способностей, связанных с ним слабее, но всё же связанных, увеличение вводных тоже должно давать улучшение.

Серия моделей, которые OpenAI обучила на пути к итоговой версии с 1,5 миллиарда параметров, наглядно показала эту зависимость: каждая аккуратно легла на кривую растущих возможностей. Поэтому никого особенно не удивило, что самая крупная из них, названная GPT-2, заметно превзошла детский лепет GPT-1 и стала выдавать длинные и достаточно связные тексты, которые можно было принять за написанные человеком. По сравнению с нынешними моделями текст был корявым и часто скатывался в тарабарщину. Но впервые вдруг стало возможно автоматизировать письмо в массовом масштабе.

Куда более мрачным сюрпризом для команды стало то, что именно GPT-2 производила с новообретённой связностью. Стоило дать более разговорчивой языковой модели несколько слов вроде «Хиллари Клинтон» или «Джордж Сорос», как она быстро сворачивала к теориям заговора. Небольшие порции неонацистской пропаганды, попавшие в обучающие данные, могли всплывать самым ужасным образом. Неожиданно дурное поведение модели встревожило исследователей безопасности ИИ: они увидели в нём предвестие будущих злоупотреблений и рисков, которые может принести более мощный несогласованный ИИ. Когда GPT-2 выдала гневную тираду против переработки отходов («Переработка – это НЕ хорошо для мира. Это плохо для окружающей

среды, плохо для нашего здоровья и плохо для нашей экономики»), один исследователь безопасности распечатал её и повесил – наполовину в шутку, наполовину в предупреждение – над контейнерами для раздельного сбора мусора в офисе.

В другой раз кто-то попросил GPT-2 придумать для маленьких детей систему поощрений за сделанные уроки и выполненные домашние обязанности. Когда GPT-2 предложила конфеты, это снова встревожило кое-кого из исследователей безопасности: они заметили, что это тактика педофилов. Одного сотрудника-европейца такая ассоциация ошеломила.

– Моя мама точно так делала. Летом по воскресеньям – мороженое, если сделаешь свои дела по дому, – вспоминает он.

Он думал, не американская ли это особенность – такая сверхчувствительность. Это был один из многих случаев, заставлявших его усомниться в самой основе высоких целей OpenAI: как она может приносить пользу всему человечеству, если в ней нет сколько-нибудь значимого представительства мира? Даже он, европеец из культуры, во многом близкой к американской, часто чувствовал себя чужим из-за подавляющего перекоса в сторону американских ценностей и норм в разговорах о безопасности ИИ и не только.

GPT-2 вызвала в компании спор. Не настал ли момент, когда пора начинать придерживать исследования? Устав допускал такую возможность. Принимать решение взялись Амодеи, к тому времени повышенный до директора по исследованиям, и Джек Кларк, возглавлявший направление политики и по-своему тревожившийся об экзистенциальных и других опасных рисках. Они провели внутренний опрос и устроили несколько совещаний об «информационной опасности», чтобы обсудить возможные злоупотребления технологией. Если GPT-2 попадёт в руки террористов, диктаторов или ферм кликбейта, рассуждали они, модель можно будет использовать в дурных целях. И пусть на этот раз она не выглядела угрозой существованию, будущие модели будут только мощнее, а вероятность будет расти. Лучше рано создать прецедент, придержав исследования. OpenAI, решили они, не должна выпускать полную версию.

* * *

Джек Кларк, бывший журналист, был директором OpenAI по стратегии и коммуникациям, а в конце 2018 года целиком переключился на выращивание её зарождающегося присутствия в политике. Он регулярно ездил в Вашингтон и наслаждался ролью человека, к которому политики обращаются по любому вопросу об ИИ.

– Я что-то вроде ИИ-Википедии, – говорил он им и тут же оговаривал свою «предвзятость», как он это называл, как человека из OpenAI: – Мы хотим стабильной среды для выработки политики в отношении передовых технологий,

которая работала бы при разных администрациях, потому что нашу миссию за один президентский срок не выполнить.

Рассказав мне об этом, он добавил:

– Нам очень повезло, что политики уделяют нам довольно много времени, потому что, по-моему, ясно: чтобы всё шло хорошо, мы просто хотим, чтобы у них было больше информации, и хотим, чтобы у них было больше возможностей самим добывать информацию.

В феврале 2019 года Кларк начал медийное наступление, широко оповещая разные издания, что OpenAI создала опасную технологию и поэтому не выпускает её. Вместо этого она выпустит лишь уменьшенную версию с 8 процентами параметров полной модели, чтобы публика почувствовала вкус её возможностей. Вместе с Амодеи и несколькими другими он написал пост с примерами текстов GPT-2, иллюстрировавшими весь её потенциал.

– Совершенно ясно, что, если эта технология созреет – а я бы дал ей год-два, – её можно будет использовать для дезинформации или пропаганды, – сказал он моему тогдашнему коллеге по MIT Technology Review Уиллу Найту.

Кларк обошёл тот факт, что именно OpenAI возглавляла гонку, которая и должна была довести технологию до зрелости в эти сроки.

– Мы пытаемся опередить события, – говорил он.

Шаг OpenAI вызвал яростный отпор внешних исследователей, свято веривших, что открытая наука – основа всей области. Любая организация, которая в ней не участвует, вызывает подозрения. Тем более если она публично хвастается своим решением. Многие к тому же считали паникёрство OpenAI по поводу того, что по сути было мощной программой автодополнения, плохо выверенным и нелепым. GPT-2 и близко не была настолько продвинутой, чтобы представлять серьёзную угрозу; а если бы и была, зачем всем о ней рассказывать, а потом закрывать её от общественной проверки? Всё это выглядело неискренне, как самовозвеличивающий рекламный трюк. В Стэнфорде, когда Рэдфорд выступил с докладом о GPT-2, маститый профессор обработки естественного языка поднял руку, чтобы задать последний вопрос.

– Так она опасна? – насмешливо спросил он.

Зал покатился со смеху.

– У Алека был такой грустный вид, – вспоминает стэнфордский исследователь, бывший в зале. – Стэнфорд относился к OpenAI с таким презрением.

Внутри OpenAI многие исследователи тоже были недовольны решением Амодеи и Кларка. Тем, кто не разделял их взглядов на катастрофические риски, и само решение, и последовавший медийный цирк казались несколько странными. Даже среди тех, кто разделял, некоторые сомневались в здравости суждений этой пары.

– Было ошибкой раздувать из этого такое событие, – сказал мне один исследователь безопасности ИИ. – Это было похоже на крики «Волки!».

Сразу после волны критики Кларк в маниакальном возбуждении расхаживал по офису, разговаривая по телефону звонок за звонком и пытаясь вернуть себе контроль над ситуацией. От споров он отмахивался.

– Мы ломаем нормы, а это порождает множество разных мнений, – сказал он мне позже, во время моего визита в офис.

Рано или поздно, говорил он, всем организациям, ведущим передовые исследования ИИ, придётся тщательнее выбирать, что публиковать. OpenAI первой опробовала, как может выглядеть такой порядок, чтобы не оказаться застигнутой врасплох.

– Если мы правы и AGI создать можно, – говорил Кларк, – нам, чёрт возьми, нужны действительно хорошие процедуры обращения с опасной информацией.

И пусть исследователям манёвр OpenAI не понравился, политикам он понравился, добавил он. Многие в Вашингтоне видели в открытой культуре исследований ИИ угрозу. Готовность OpenAI идти против течения принесла ей в Вашингтоне больше доверия.

Но за кулисами руководство понимало и то, что враждебность научного сообщества долго терпеть нельзя. Проблемы с репутацией у лаборатории накапливались. Её дикие заявления об AGI, чрезмерная шумиха вокруг GPT-2 и прочий маркетинг, а теперь, в начале 2019 года, ещё и только что объявленная франкенштейновская структура – всё это навлекало критику со всех сторон, и ведущие исследователи области смотрели на неё со всё большим скепсисом. Вдобавок её доли в капитале пока ничего не стоили, и ей по-прежнему было трудно нанимать и удерживать таланты. Сотрудники гадали, не берут ли внешние кандидаты предложения OpenAI лишь затем, чтобы выторговать себе больше у Google или DeepMind. OpenAI нужно было найти способ утвердиться в качестве серьёзной исследовательской организации.

Об этом часто говорили за обедом и на совещаниях, а также во внутреннем документе «Мозговой штурм по работе с научным сообществом». В разделе «Стратегия» говорилось: «Явно рассматривать сообщество МО как заинтересованную сторону в коммуникациях», – МО означало машинное обучение. «Изменить тон и внешние сообщения так, чтобы мы настраивали их против себя, только когда сознательно этого хотим». Документ признавал и то, что плохая научная репутация в конечном счёте подорвёт влияние OpenAI в Вашингтоне. «Чтобы влиять на политику на уровне правительства, нас должны считать самым надёжным источником в вопросах исследований МО и AGI, – говорилось в разделе „Политика“. – Широкая поддержка научного сообщества не только необходима для такой репутации, но и усилит наш голос».

Команда Кларка придумала новый план – поэтапный выпуск. Вместо того чтобы навсегда придержать GPT-2, OpenAI будет с промежутками публиковать всё более крупные модели, которые разработала, а затем, если всё пойдёт хорошо, выпустит и полную версию с 1,5 миллиарда параметров. Это позволит OpenAI и другим постепенно наблюдать за возникающими последствиями и

справляться с ними, говорила команда, а лаборатории даст время вместе с другими организациями исследовать риски между этапами.

Кларк подчёркивал в своей команде, как важно выстроить через эти партнёрства целую экосистему. Работа с известными институтами поможет наладить сотрудничество индустрии и науки в борьбе с рисками ИИ. Это обеспечит и более широкую поддержку усилиям OpenAI изменить нормы публикации исследований, а заодно поможет отполировать репутацию лаборатории. Команда обратилась к исследователям ИИ и безопасности в нескольких избранных организациях и дала им ранний доступ к полной версии GPT-2, чтобы проверить, насколько её можно использовать во вред. Кларк велел команде добиться от организации каждого исследователя «как можно более весомой поддержки», чтобы OpenAI могла назвать партнёрами не только людей, но и их институты. Затем команда подготовила аналитический доклад, расхваливающий её стратегию выпуска и подчёркивающий эти партнёрства. Чтобы представить OpenAI лидером, в нём перечислялись организации, которые после GPT-2 тоже отказались от немедленной публикации своих исследований.

Работа окупилась. Вскоре в отраслевых объединениях и политических аналитических центрах заговорили о том, что придерживать исследования – ответственный подход к рискам ИИ. В конце 2020 года Кларк окажется среди тех, кто вместе с братом и сестрой Амодеи уйдёт из OpenAI и станет сооснователем Anthropic. А до тех пор его работа в OpenAI поможет утвердить её влияние и заложить основу для её размашистых политических амбиций.

* * *

OpenAI завела дорожную карту, чтобы упорядочить свои исследования. Амодеи относился к ней как инвестор: он называл это «портфелем ставок». Он и другие исследователи следили за разными идеями в области, рождёнными разными философиями о том, как достичь общего искусственного интеллекта, и продвигали каждую через небольшие эксперименты. Перспективные OpenAI продолжала. Неудавшиеся бросала.

Проект победы на чемпионате по Dota 2, считал Амодеи, больше не приносил особой пользы. Команда Dota 2 в апреле обыграла соперников и достигла цели, помогла заполучить инвестиции Microsoft, а кое-кому и придала новую уверенность в стратегии масштабирования. На его взгляд, проект себя исчерпал. Команду Dota 2 распустили.

Зато в GPT-2 Амодеи видел продолжение. Она воплощала ставку, известную в области как гипотеза «чистого языка». Язык, гласит эта теория, – главное средство общения людей, а значит, все знания мира в какой-то момент должны быть записаны в тексте. Отсюда следует, что AGI может возникнуть, если обучить алгоритм на огромных объёмах языка – и ни на чём больше. Этой идее противостоит гипотеза «заземления», утверждающая, что физический мир и наша способность воспринимать его и взаимодействовать с ним – столь же

важная составляющая нашего интеллекта. Тогда AGI сможет возникнуть лишь из сочетания языка с восприятием, вроде компьютерного зрения, и с взаимодействием – например, когда физический или виртуальный агент совершает действия в реальном мире.

В документах компании исследователи взвешивали достоинства разных подходов, и сотрудники, занимавшиеся безопасностью ИИ, однажды спорили о преимуществах гипотезы «чистого языка», проводя отвратительные аналогии с людьми с инвалидностью. Эти обсуждения показали, как быстро мерки интеллекта могут скатиться к тревожным оценкам того, у каких групп людей интеллект выше, а у каких ниже.

«Язык в той или иной форме – это разница между одичавшим человеком и человеком в обществе. Пример – Хелен Келлер», – говорилось в документе под заголовком «Некоторые исходные доводы в пользу центральной роли языка». На полях исследователи безопасности ИИ продолжали спорить за и против «чистого языка» в цепочках комментариев.

«К тому же слепые примерно так же способны, как зрячие», – писал один исследователь в доказательство того, что «заземление» через зрение, похоже, не нужно.

«Слепые, похоже, в значительно худшем экономическом положении», – отвечал другой, ссылаясь на данные Национальной федерации слепых, согласно которым более 70 процентов взрослых с нарушениями зрения не работают полный день.

«Слепые всё равно гораздо способнее шимпанзе, – отвечал третий. – Есть очень впечатляющие слепые».

Многие в OpenAI скептически относились к «чистому языку», но GPT-2 заставила их передумать. Обучение модели всё точнее предсказывать следующее слово заметно улучшило её работу и на других, как будто лишь отдалённо связанных задачах обработки языка. Казалось возможным, даже правдоподобным, что модель GPT может обрести более широкий набор способностей, если идти этим путём и дальше: продолжать обучение и ещё повышать точность предсказания следующего слова. Амодеи стал считать масштабирование языковых моделей если не единственным необходимым условием AGI, то, возможно, самым быстрым путём к нему. Свою роль сыграло и то, что команда робототехники постоянно натыкалась на проблемы с оборудованием для роботизированной кисти, а это было худшее сочетание: дорогой, но медленный прогресс.

Но была загвоздка: если OpenAI продолжит масштабировать языковые модели, это может усугубить те самые опасности, о которых она предупреждала в связи с GPT-2. Амодеи убеждал остальных в компании – и Альтман с ним соглашался, – что это не повод уклоняться от задачи. Вывод был как раз противоположным: OpenAI должна масштабировать свою языковую модель как можно быстрее, говорил Амодеи, но не выпускать её сразу. GPT-2 показала, как

легко другим будет заполучить более мощные возможности ИИ; двое аспирантов даже сделали открытую версию GPT-2 раньше, чем OpenAI выпустила собственную полную. Другие люди рано или поздно начнут масштабировать языковые модели дальше. А значит, лучший способ обеспечить полезный AGI – чтобы OpenAI вырвалась вперёд и, пользуясь внутренней форой, разобралась, как сделать свою масштабированную модель безопаснее. Когда придёт время её показать, дополнительная доводка и шлифовка помогут установить нормы безопасности ИИ – так же, как изначальное удержание GPT-2 сдвинуло нормы публикации исследований.

К тому же, раз одна из версий GPT-2 уже вышла в мир, появились и свидетельства того, что опасности моделей «чистого языка» не так уж велики. Насколько было известно OpenAI, её не использовали в скоординированных массовых кампаниях по дезинформации – а такие кампании, во всяком случае, лучше, чем возможные экзистенциальные риски AGI.

– Злоупотребления, очевидно, это плохо, – говорил мне Амодеи. – Но языковая модель гораздо менее мощна, чем AGI. Меня очень тревожит, что языковые модели могут превратить в оружие для дезинформации и тому подобного – это очень страшно, – но в то же время это относительно отдельная, ясная и определённая проблема.

Иными словами, с точки зрения Амодеи, масштабирование GPT-2 было не только, возможно, самым быстрым путём к AGI, но и таким, где возможные риски по дороге будут сравнительно ограничены теми, которые он считал управляемыми, – недостоверной информацией и дезинформацией, а не катастрофой. Это дало бы OpenAI более безопасный полигон для экспериментов с мощной, но не слишком мощной системой ИИ и позволило бы отладить разные шероховатости, в том числе с её выпуском.

– Что такое AGI? Как выглядит AGI? – говорил Амодеи. – Ну, понимаете, мы в неловком положении: мы не знаем, как он выглядит. Не знаем, когда он появится. Поэтому мы ищем вещи, которые не являются AGI, но несут хотя бы часть его возможностей и трудностей. И надеемся, что, если хорошо справимся с ними, то будем, так сказать, готовы к высшей лиге.

Эта логика работала лишь при определённом допущении: AGI, при всей своей бесформенности и непознаваемости, ещё и неизбежен. Ещё много лет OpenAI будет оправдывать свои действия вариациями того же довода. Перед призраком неотвратимого прихода AGI компания должна продолжать разрабатывать всё более мощные модели, чтобы подготовить себя и общество. Даже если эти модели несут свои риски, опыт, который они дают для предотвращения возможного ИИ-апокалипсиса или встречи с ним, делает эти риски приемлемыми.

Когда в начале 2023 года ChatGPT захлестнул мир, один китайский исследователь ИИ поделился со мной трезвым анализом, распутавшим довод OpenAI о неизбежности. То, что сделала OpenAI, говорил он, не могло случиться

нигде, кроме Кремниевой долины. В Китае, который соперничает с США по числу талантов в ИИ, ни одна команда исследователей и инженеров, как бы ни была она впечатляюща, не получила бы миллиард долларов, не говоря уже о вдесятеро большей сумме, на разработку баснословно дорогой технологии без внятного представления о том, как именно она будет выглядеть и для чего пригодится. Лишь после выхода ChatGPT китайские компании и инвесторы с азартом принялись финансировать создание гигантских моделей, увидев достаточно доказательств того, что вложения можно окупить коммерческими применениями.

За время работы над этой темой я пришла к ещё более поразительному выводу. Даже в Кремниевой долине другие компании и инвесторы до ChatGPT не спешили вливать в масштабирование неограниченные суммы. Включая Google и DeepMind, изначальную соперницу OpenAI. Именно OpenAI – с её миллиардерским происхождением, особым идеологическим уклоном и исключительными напором, связями и талантом Альтмана добывать деньги – создала то благоприятное сочетание, в котором её особое видение смогло возникнуть и взять верх.

– У меня ощущение, что Сэм – самый честолюбивый человек на планете, – говорит один бывший сотрудник.

Иными словами, всё, что делала OpenAI, было противоположностью неизбежности; взрывная глобальная цена её огромных моделей глубокого обучения и опасная гонка, которую она развязала по всей отрасли, – масштабировать такие модели до пределов планеты, – могли возникнуть только там, где они и возникли.

* * *

К «демо для Гейтса» в апреле 2019 года OpenAI уже масштабировала GPT-2 до чего-то чуть более крупного. Но скромное увеличение Амодеи не интересовало. Если цель – увеличить фору OpenAI, GPT-3 должна быть как можно больше. Microsoft вот-вот должна была передать OpenAI в рамках своей инвестиции новый суперкомпьютер с десятью тысячами Nvidia V100 – самыми мощными в то время в мире GPU для обучения моделей глубокого обучения. (V – в честь итальянского химика и физика Алессандро Вольты.) Амодеи хотел задействовать все эти чипы разом, чтобы создать новую большую языковую модель.

Многим эта идея казалась просто абсурдной. До тех пор крупномасштабной считалась модель, обученная на нескольких десятках чипов. В ведущих академических лабораториях MIT и Стэнфорда аспиранты считали роскошью иметь десять чипов. В университетах за пределами США, например в Индии, студентам везло, если один чип приходился на нескольких человек и для своих исследований им доставалась доля GPU.

Многие исследователи OpenAI сомневались, что затея Амодеи вообще сработает. Некоторые доказывали, что постепенное масштабирование было бы взвешеннее, научнее и предсказуемее. Но Амодеи стоял на своём, и его поддерживали другие руководители. Суцкеверу не терпелось проверить свою гипотезу о масштабировании трансформеров; Брокман хотел и дальше повышать известность компании; Альтман подталкивал к самому размашистому удару из возможных. Вскоре Амодеи повысили до вице-президента по исследованиям.

За кулисами Альтман учитывал и ещё одно обстоятельство: миллиардная инвестиция Microsoft несла с собой ожидания на миллиард; у OpenAI тикали часы – нужно было предъявить что-то, что оправдало бы расходы. Если Амодеи видел в более крупной языковой модели необходимое условие исследований безопасности ИИ, то Альтман видел в ней возможность выполнить обещание, данное Microsoft.

В следующие месяцы Амодеи и Альтман столкнутся из-за того, как и когда выпускать GPT-3; победит Альтман, выпустив модель в мир в ускоренные сроки. За годы до ChatGPT эти два решения – резко увеличить размер GPT-3 и быстро её выпустить – изменят ход развития ИИ. Они запустят стремительное ускорение прогресса ИИ и разожгут ожесточённое соперничество между компаниями и странами. Они подстегнут небывалое расширение надзорного капитализма и эксплуатации труда. Из-за одного лишь объёма требуемых ресурсов они сосредоточат разработку технологии в небывалой прежде степени, отрезав остальной мир от участия в ней. Они ускорят порочный круг, в котором университеты, не способные конкурировать, теряют аспирантов и профессоров, уходящих в индустрию, независимая академическая наука чахнет, а подотчётность начинает уходить в прошлое. Они усилят воздействие ИИ на окружающую среду до такой степени, что в отсутствие прозрачности и регулирования ни внешние эксперты, ни правительства до сих пор не могут его полностью подсчитать.

Но всё это было ещё впереди. Осенью 2019 года Амодеи собрал команду под названием Nest («Гнездо»), в основном из других исследователей безопасности ИИ, намереваясь держать разработку GPT-3 внутри компании под строгим контролем. С этого команда и начала свой напористый рывок к масштабу.

* * *

GPT-3 была по сути той же моделью, что и GPT-2, только накормленной несравнимо большим количеством данных и вычислений – настолько, что многим результат покажется уже не количественным, а качественным отличием. Но использование десяти тысяч чипов создавало новые проблемы. Всегда есть небольшая вероятность, что какой-то чип посреди обучения откажет – так же, как ноутбук зависает, когда открыто слишком много окон. Если откажет один чип, откажет всё, и обучение придётся начинать сначала. На десяти тысячах GPU

вероятность отказа хотя бы одного чипа многократно возрастала. Такая ошибка обошлась бы чрезвычайно дорого – и деньгами, и временем, – ведь команда Nest рассчитывала, что обучение займёт как минимум несколько месяцев.

Чтобы решить проблему, команде нужен был способ после любого сбоя возобновлять обучение ровно с того места, где оно прервалось. Нужно было ещё придумать, как распределить обучение по всем десяти тысячам чипов, – этот процесс называют шардированием. Что лучше, например: разрезать модель на десятки, сотни или тысячи кусков, чтобы каждый обучался на отдельном кластере GPU, а потом их объединить?

Была и трудность с данными. Для наилучшего результата набор данных должен расти пропорционально числу параметров и объёму вычислений. Если параметров слишком много, а данных мало, модель может начать слово в слово выдавать строки из обучающих данных, то есть станет фактически бесполезной. Для GPT-2 Рэдфорд тщательно отбирал данные. Он выкачивал текст из статей и сайтов, ссылками на которые делились на Reddit и которые получили там хотя бы три голоса «за». Получился массив в сорок гигабайт, около восьми миллионов документов, который он назвал WebText.

Для GPT-3 этого было слишком мало. Поэтому Nest расширила данные, добавив ещё более широкую выборку ссылок с Reddit, а также англоязычную «Википедию» и загадочный набор данных Books2, подробности о котором OpenAI так и не раскрыла, но который, как сказали мне два человека, знакомых с ним, содержал изданные книги, выкачанные из Library Genesis – теневого онлайн-хранилища пиратских книг и научных статей. В 2023 году Гильдия авторов и семнадцать писателей, в том числе Джордж Р. Р. Мартин и Джоди Пиколт, подадут иск против OpenAI и Microsoft, обвинив их в массовом нарушении авторских прав. В марте 2024 года OpenAI ответит, что удалила эти наборы данных и перестала использовать их для обучения после GPT-3.5, которая к тому времени уже была выведена из обращения.

Но и этого было мало. Тогда Nest наконец обратилась к общедоступному набору данных Common Crawl – огромной свалке из петабайтов, то есть миллионов гигабайт, текста, регулярно собираемого со всего интернета, – источнику, которого Рэдфорд сознательно избегал из-за его низкого качества. Чтобы обуздать мусор в данных, команда Nest обучила модель машинного обучения находить в Common Crawl образцы, больше всего похожие на статьи «Википедии». Идея была в том, что если текст похож на «Википедию», то и по качеству он, скорее всего, ей ближе. Добавили и некоторые образцы на других языках, кроме английского, хотя в итоге на них пришлось лишь 7 процентов данных. И всё же при обучении модели исследователи присвоили отфильтрованным данным Common Crawl самый низкий приоритет. Иными словами, GPT-2 стала вершиной качества данных; дальше оно только снижалось.

Когда пришло время собирать данные для GPT-4, вышедшей двумя годами позже, погоня за количеством подточила качество ещё сильнее. Фильтр с данных

Common Crawl сняли, и большая их часть хлынула в обучение. Через партнёрство с Microsoft OpenAI получила и полную выгрузку GitHub – онлайн-хранилища кода, принадлежащего Microsoft. Когда и этого не хватило, сотрудники OpenAI собирали всё, что могли найти в интернете: выкачивали ссылки из Twitter, расшифровывали видео с YouTube, сгребали длинный хвост прочего контента – с нишевых блогов, из уже существующих сетевых выгрузок данных и с сайта для хранения текстов Pastebin. Всё, на чём не было прямого запрета на скрейпинг, считалось доступным – бери не хочу.

В Google некоторые исследователи сокрушались, что готовность OpenAI идти на юридический риск ради данных даёт ей большое преимущество. Google куда осторожнее относилась к доступу к данным и их использованию и строго соблюдала правила, в том числе европейский закон о защите данных, известный как GDPR. По иронии судьбы, из-за приверженности Google законности OpenAI легче добиралась до данных Google, чем сама Google. OpenAI преспокойно выкачивала и расшифровывала видео с принадлежащего Google YouTube, а исследователям Google приходилось продираться сквозь серьёзные внутренние бюрократические препоны, чтобы соблюсти строгую лицензию YouTube на загружаемый пользователями контент. OpenAI это соблюдение правил не заботило – или, на языке технологических стартапов, она была от него «свободна». Это классическое мышление Кремниевой долины, где основатели и инвесторы твердят как мантру, что стартапы могут и должны заходить в юридические серые зоны (вспомните Airbnb, Uber или Coinbase), чтобы взрывать и перевернуть целые отрасли.

Решение понизить, а затем фактически вовсе убрать планку качества будет иметь далеко идущие последствия для людей, чей труд стоит за системами ИИ. Годами технологическая индустрия полагалась на низкооплачиваемых работников в шатком экономическом положении, которые выполняли необходимую подготовку данных для её моделей ИИ, – например, распределяли тексты по категориям и размечали изображения. Вскоре после того, как GPT-3 сделала нормой гигантские наборы данных худшего качества, характер этой работы сменился: вместо в основном безобидного контента работникам всё чаще приходилось иметь дело с тягостным, в том числе ради модерации, совсем как прежде в социальных сетях. Такая модерация была нужна, чтобы генеративные системы не выдавали сотням миллионов пользователей самые мерзкие части своих всеобъемлющих наборов данных – описания и изображения насилия, сексуального надругательства или самоповреждения.

– В том, как контролировать то, что выдают эти модели, произошёл большой сдвиг парадигмы, – говорит Райан Колн, генеральный и управляющий директор Appen, платформы, которая сводит компании Кремниевой долины с работниками по данным. – В традиционном ИИ вы контролируете результат, ограничивая входные данные, – той самой фильтрацией данных, которой занималась команда Рэдфорда, – потому что модель учится только на примерах,

которые вы ей даёте. Трудность генеративного ИИ в том, что входные данные – это весь корпус человечества. Поэтому приходится контролировать выход.

В статье 2023 года Абеба Бирхане и её соавторы предложат понятие «законов масштабирования ненависти», чтобы раскритиковать саму идею обучать модели глубокого обучения на нефильтрованных данных, или, как они их называли, «болотах данных». Они проанализировали два общедоступных набора изображений с подписями, используемых для обучения открытых генераторов изображений, – LAION-400M и LAION-2B-en, оба собранные из Common Crawl, с четырьмя сотнями миллионов и двумя миллиардами изображений соответственно. Они показали, что количество ненавистнического и оскорбительного контента растёт с размером набора данных и усугубляет дискриминационное поведение обученных на нём моделей. Модели, обученные на двух миллиардах изображений, например, в пять раз чаще, чем обученные на четырёхстах миллионах, помечали лица чернокожих мужчин как преступников. Позже в том же году стэнфордское исследование LAION-5B – набора из пяти миллиардов изображений, на котором обучали Stable Diffusion, – обнаружит в нём тысячи подтверждённых и предполагаемых изображений сексуального насилия над детьми.

Среди способов контролировать выход OpenAI наймёт работников в Кении в среднем менее чем за два доллара в час, чтобы создать автоматический фильтр для модерации контента, – это первым раскрыл корреспондент журнала Time Билли Перриго. Она наймёт и больше тысячи других подрядчиков по всему миру для обучения своих языковых моделей с подкреплением на основе обратной связи от человека, RLHF, – того самого метода, который она разработала, чтобы научить ИИ-агента сальто, – в том числе чтобы раз за разом задавать моделям вопросы и оценивать ответы, стараясь по возможности их обуздать.

Хито Штайерль, немецкая художница и режиссёр, снявшая документальный фильм о сирийских беженцах, работающих с данными, повторила в разговоре со мной критику Бирхане. Психологически травмирующий материал накапливается, когда в основе сбора данных лежит массовая слежка, говорила она. Чтобы решить проблему, нужно вернуться к её корню: задаться вопросом, что на самом деле лежит в данных, и поставить под сомнение саму идею их массового, неразборчивого захвата.

Часть II

Глава 6

Восхождение

Ещё в начале карьеры Альтман заметил, что новый CEO добивается успеха, только если «основывает компанию заново». В YC, унаследовав пост президента, он проделал это со всей убеждённостью. Он запустил новые программы, в том числе новый фонд, чтобы акселератор мог поддерживать стартапы на разных стадиях. Он вывел YC в область «жёстких» технологий – тех, что требуют смелых научных прорывов: ядерный синтез, квантовые вычисления, беспилотные автомобили. YC и без того был громким именем, а теперь его сфера влияния выросла с пары сотен компаний до тысяч в год, и акселератор превратился в один из центров притяжения Кремниевой долины. «Больше всего я горжусь тем, что мы по-настоящему построили империю», – сказал Альтман, когда ушёл с поста президента.

Конец эпохи YC стал началом его эпохи в OpenAI. В марте 2019 года, перейдя в OpenAI на полную ставку, он быстро принёс с собой тот же напористый настрой, что и в YC. Ему было мало, чтобы OpenAI входила в число ведущих ИИ-организаций мира: он хотел, чтобы она была единственной. Годами Альтман учил основателей – в YC и на других площадках – смотреть на стартап как на игру, где победитель получает всё. Если у стартапа есть хоть какая-то надежда на успех, говорил он им, нужно действовать быстро и без устали: обогнать соперников, а потом раз за разом отбивать их натиск.

Его любимым магическим числом была десятка, и восходило оно к монопольной стратегии Тиля. «Моё слегка безумное и отчасти произвольное эмпирическое правило: вам нужна технология, которая на порядок лучше ближайшей альтернативы», – говорил Тиль в 2014 году на лекции, которую читал на стэнфордском курсе Альтмана о стартапах. Amazon, к примеру, придумал, как продавать в десять раз больше книг, чем обычные книжные магазины. PayPal, его собственная компания, придумала, как переводить деньги в десять раз быстрее, чем проходят чеки. «Нужно какое-то очень мощное улучшение, возможно, на порядок, по какому-то ключевому параметру», – говорил Тиль.

Десятка стала для Альтмана круглым числом на все случаи жизни. Стартапам, советовал он, мало ворваться на рынок с технологией, которая в десять раз лучше: её нужно улучшать в десять раз с каждым поколением. А скорость, с которой компания выходит на каждое новое поколение, – ещё одна ключевая переменная, способная её создать или погубить. «Если ваш цикл итерации – неделя, а у конкурента – три месяца, вы оставите его далеко позади», – говорил он в 2017 году группе начинающих предпринимателей.

В YC Альтман подталкивал коллег-партнёров увеличивать число финансируемых компаний в десять раз. «И со временем мы придумаем, как вырасти ещё в десять раз, а потом ещё в десять», – рассказывал Альтман позже о своей стратегии на одном мероприятии. «Когда-нибудь мы будем финансировать все компании мира». «Сэм был первым, от кого я услышал, что благодаря работе основателей и бренду, который создал YC, мы фактически были монополистом в этой сфере», – говорит Джефф Ралстон, преемник Альтмана в YC.

В OpenAI Альтман собирался действовать по той же схеме. В служебной записке, которую он разослал по компании в конце 2019 года, чтобы изложить свои долгосрочные замыслы, он подчеркнул, что к концу 2020 года OpenAI должна «быть номером один» в четырёх категориях: технические результаты, вычислительные мощности, деньги (на покупку новых мощностей) и готовность – то есть безопасность и защищённость организации, а также её устойчивость к ситуациям сильного стресса.

Важнейшей из них, по его словам, была первая. Чтобы у OpenAI был шанс выполнить свою миссию, ей нужно было первой создать полезный AGI – или стать таким лидером, который всё равно сможет определять развитие AGI. «Хотя теоретически мы могли бы замедлить работу над возможностями, – писал он, имея в виду продвижение технических результатов, – при том темпе, который набирают другие, нам, скорее всего, придётся двигаться очень быстро, если мы хотим серьёзно влиять на AGI». И это будет становиться тем вернее, чем больше конкурентов раскусят стратегию OpenAI и придут в эту область.

«Чтобы добраться до AGI, нам нужно ещё много скачков в десять раз, – добавлял он дальше в записке. – Мы всегда должны стремиться к впечатляющим результатам, а не к постепенным улучшениям».

Для этой формулы успеха были важны и другие соображения. Первостепенное значение имело то, чтобы Microsoft оставалась довольна: от этого зависело лидерство по вычислительным мощностям. Если OpenAI будет успешна, Microsoft соглашалась дать намного больше миллиарда долларов. «Мы хотели бы, чтобы Microsoft оставалась нашим главным партнёром до самого конца, – говорил Альтман. – Они способны дать нам, по крайней мере в ближайшие пять лет, самые мощные суперкомпьютеры в мире». А это означало отход от вольной академической атмосферы первых лет OpenAI к сосредоточенной коммерциализации, которая принесла бы выгоду Microsoft. Если у OpenAI появятся другие исследовательские проекты, ресурсы на них найдутся. «Перефразируя знаменитые слова Диснея, – писал Альтман, – мы должны зарабатывать больше денег, чтобы больше заниматься исследованиями, а не заниматься исследованиями, чтобы больше зарабатывать».

Кроме того, компании пора было поубавить открытости. «Информационная опасность разговоров об AGI будет только расти по мере нашего продвижения», – доказывал Альтман. Пришло время ограничить

публикацию исследований и выпуск моделей, ввести более строгий режим конфиденциальности и рассказывать об успехах лишь в узких навыках, а не в более общем развитии ИИ. Отдельно всем следовало исходить из того, что «каждое наше решение и каждый наш разговор в итоге станут предметом расследования и попадут на первую полосу New York Times».

И всё же, говорил он, «по-прежнему очень важно, чтобы мир думал, что мы в чём-то побеждаем». Тогда «ключевые влиятельные люди в мире» будут «готовы из кожи вон лезть, чтобы нам помочь», а политики всего мира «на уровне президентов или их представителей» станут обращаться к OpenAI «за ответами, когда им нужно принять важные решения». С этой целью «нам, вероятно, стоит планировать выпуск хотя бы одной очень впечатляющей демонстрации прогресса в год».

Наконец, компании нужно было стать серьёзнее и сплочённее. Альтман привёл цитату Хаймана Г. Риковера, адмирала ВМС США, которого называют «отцом атомного флота» за создание первых в мире атомных подводных лодок. Эту цитату Альтман ещё в первые дни OpenAI велел написать на стене офиса:

> Я считаю, что долг каждого из нас – действовать так, словно от [него] зависит судьба мира. Конечно, один [человек] в одиночку не справится. Но один [человек] может изменить многое. Каждый из нас обязан приложить свои личные и независимые способности к широкому кругу человеческих забот. Именно с этим убеждением мы прямо смотрим в лицо нашему долгу перед процветанием. Мы должны жить ради будущего человеческого рода, а не ради собственного комфорта или успеха.

«Создание AGI, который принесёт пользу человечеству, – это, пожалуй, самый важный проект в мире, – написал Альтман в документе под цитатой. – Мы должны ставить миссию выше любых личных предпочтений.

«Дела, где ставки низки, не должны сопровождаться драмами – прибережём способность к драме для дел, где ставки высоки (а таких будет немало)».

* * *

На самом деле драма уже назревала. Мелкие трещины, разошедшиеся по всей компании, начали сливаться в крупные. Брокман и брат с сестрой Амодеи, ещё недавно охотно называвшие друг друга друзьями, теперь сталкивались лбами по всё большему числу вопросов. Брокмана, например, раздражало, что Дарио Амодеи отодвинул работу над Dota 2 на второй план: ему казалось, что Амодеи не принимает его вклад всерьёз. Dota 2 когда-то был самым прожорливым по вычислениям проектом, и Брокмана задевало, что теперь Амодеи стягивает мощности к работе Nest над GPT-3. Брат и сестра Амодеи, в свою очередь, считали, что с Брокманом трудно работать, и не желали подпускать его к разработке своих языковых моделей.

Напряжение раскололо руководство, а затем постепенно перекинулось и на тех, кто был предан каждому из лидеров. Во время проекта Dota 2 Брокман сроднился с частью своей команды – через изматывающие рабочие часы, постоянный стресс и спонтанную вылазку на Гавайи. Особенно близко он сошёлся с Якубом Пахоцким и Шимоном Сидором, двумя польскими учёными, которые жили в одной квартире и были лучшими друзьями. Другой лагерь составили команды Амодеи по безопасности ИИ, и прежде всего ядро команды Nest: их объединяла общая – в разной степени – тревога из-за вышедшего из-под контроля ИИ и экзистенциальных или иных крайних рисков. Свою работу они отгородили от остальной компании: завели закрытые каналы в Slack и документы, недоступные даже другим руководителям. Это раздражало многих помимо Брокмана: они тоже чувствовали себя оттеснёнными, видя, как тают их вычислительные ресурсы и как мало им видно из ключевых исследований компании.

Тем временем у лагеря Амодеи росла тревога и по поводу некоторых поступков Альтмана. Вскоре после подписания сделки с Microsoft несколько человек из этой группы с изумлением узнали, как далеко зашли обещания Альтмана о том, к каким технологиям Microsoft получит доступ в обмен на инвестиции. Условия сделки не совпадали с тем, что они поняли со слов Альтмана. Если в моделях OpenAI действительно возникнут проблемы безопасности, опасались они, эти обязательства сильно затруднят, а то и вовсе не позволят остановить развёртывание моделей. В лагере Амодеи начали всерьёз сомневаться в честности Альтмана.

«Мы все люди прагматичные, – говорит один из членов группы. – Мы, понятное дело, привлекаем деньги, мы будем заниматься коммерцией. Если ты, как Сэм, заключаешь кучу сделок, тебе может казаться вполне разумным рассуждать так: „Ладно, заключим сделку, обменяем одно, потом обменяем следующее“. А если ты такой, как я, то думаешь: „Мы обмениваем то, чего сами до конца не понимаем“. Кажется, что это загоняет нас в неуютное положение».

Всё это происходило на фоне растущей паранойи, которая по разным поводам охватывала компанию. В лагере сторонников безопасности ИИ она питалась тем, что они считали всё более весомыми доказательствами: мощные системы ИИ, не согласованные с человеческими целями, могут привести к катастрофе. Один причудливый случай особенно встревожил нескольких из них. В 2019 году на модели, обученной уже после GPT-2 и имевшей примерно вдвое больше параметров, группа исследователей взялась продвигать ту самую работу по безопасности, которой хотел Амодеи: проверяла обучение с подкреплением на основе отзывов людей (RLHF) как способ подтолкнуть модель к бодрым и доброжелательным текстам и отвадить от всего оскорбительного.

Но однажды поздно вечером исследователь внёс в код изменение с одной-единственной опечаткой и оставил процесс RLHF работать на ночь. Опечатка оказалась серьёзной: минус превратился в плюс, и RLHF заработал наоборот,

подталкивая GPT-2 писать больше непристойностей, а не меньше. К утру опечатка натворила дел: GPT-2 продолжал любой запрос предельно похабным, откровенно сексуальным текстом. Это было уморительно – и вместе с тем настораживало. Найдя ошибку, исследователь отправил исправление в кодовую базу OpenAI с комментарием: *Давайте не будем создавать минимизатор полезности*.

Отчасти подогреваемые пониманием того, что одно лишь масштабирование способно дать новые прорывы в ИИ, многие сотрудники тревожились и о том, что будет, если другие компании разгадают секрет OpenAI. «Секрет того, как работают наши штуки, можно написать на рисовом зёрнышке», – говорили они друг другу, имея в виду одно слово: «масштаб». По той же причине они опасались, что мощные возможности окажутся в руках злоумышленников. Руководство охотно подыгрывало этому страху: то и дело поминало угрозу со стороны Китая, России и Северной Кореи и подчёркивало, что разработка AGI должна оставаться в руках американской организации. Порой это коробило сотрудников-неамериканцев. За обедом они задавались вопросом, вспоминает бывший сотрудник: почему обязательно американской? Почему не европейской? Почему не китайской?

В этих пылких философских спорах о долгосрочных последствиях исследований ИИ многие сотрудники снова и снова возвращались к ранним сравнениям Альтмана, ставившего OpenAI в один ряд с Манхэттенским проектом. Неужели OpenAI действительно создаёт что-то вроде ядерного оружия? Это странно контрастировало с задорной, идеалистической культурой, которую она успела выстроить, будучи по большей части академической организацией. По пятницам сотрудники расслаблялись после долгой недели на вечерах с музыкой и вином, отдыхая под мягкие звуки офисного рояля, за который допоздна садились по очереди разные коллеги.

Смена атмосферы выбивала некоторых из колеи и обостряла тревогу по поводу случайных и никак не связанных между собой происшествий. Однажды журналист проскочил вслед за кем-то на огороженную парковку, чтобы попасть в здание. В другой раз сотрудник нашёл бесхозную флешку, и все переполошились: вдруг на ней вредоносные файлы – распространённый способ атаки – и это попытка взлома? Флешку проверили на компьютере с воздушным зазором, то есть полностью отрезанном от интернета, и она оказалась пустышкой. По меньшей мере дважды Амодеи и сам писал важнейшие стратегические документы на изолированном компьютере, подключая его напрямую к принтеру, чтобы распространять только бумажные копии. Он опасался, что государственные структуры похитят секреты OpenAI и построят собственные мощные модели ИИ для дурных целей.

«Никто не был готов к такой ответственности, – вспоминает один сотрудник. – Из-за этого люди не спали по ночам».

Сам Альтман боялся утечек. Втайне его беспокоили сотрудники Neuralink, с которыми OpenAI по-прежнему делила офис, – после ухода Маска это беспокоило его ещё сильнее. Тревожил его и сам Маск, располагавший внушительным аппаратом безопасности, вплоть до личных водителей и телохранителей. Остро ощущая разницу в возможностях, Альтман однажды тайно заказал электронную проверку на прослушку, чтобы найти в офисе жучки, которые Маск мог оставить для слежки за OpenAI.

Перед сотрудниками Альтман использовал призрак противников США, которые продвигают исследования ИИ быстрее OpenAI, чтобы обосновать, почему компании нужно становиться всё менее открытой и при этом работать как можно быстрее. «Мы должны считать себя ответственными за благоприятный исход для мира, – писал он в документе о своём видении. – С другой стороны, если авторитарное правительство создаст AGI раньше нас и злоупотребит им, мы тоже провалим свою миссию: нам почти наверняка придётся быстро продвигаться технически, чтобы её выполнить».

* * *

Альтман начал закручивать гайки в вопросах безопасности. Руководители спорили, где провести новую черту: вести ли себя OpenAI как компании из списка Fortune 500, которая охраняет собственные технологии, или скорее как государственной структуре, оберегающей строго засекреченные государственные тайны? Как минимум, сошлись руководители, нужно надёжно запереть веса моделей – ключевую информацию, по которой можно воспроизвести полностью обученные версии моделей глубокого обучения OpenAI. Их кража была бы бедой вдвойне: она могла бы усилить злоумышленников и подорвать конкурентное преимущество OpenAI.

Поначалу, за неимением штатных специалистов по безопасности, Альтман поручил сотруднику инфраструктурной команды, отвечавшей за всё – от GPU компании до интернета в офисе, – продумать, как предотвратить кражу моделей: не только корпоративными или государственными шпионами, но и собственными сотрудниками OpenAI. В кибербезопасности защита от «инсайдерской угрозы» – вполне обычная практика. Свои люди могут намеренно саботировать или украсть интеллектуальную собственность OpenAI; а могут и попасться на уловку и выдать её. В частных разговорах Альтман, когда ему на это указали, признал, что второму может быть подвержен кто-то вроде Суцкевера. Главный научный сотрудник был логичной мишенью для злоумышленников: классический гениальный учёный, не слишком искушённый в житейских делах, он занимал высокое положение в организации и имел доступ к самой важной информации.

У Суцкевера были свои страхи. Звезда науки в рассудочном и не слишком светском мире исследований ИИ, он успел повидать немало одержимых поклонников и поведения, граничащего с преследованием. Не раз незнакомцы

пытались пробраться в офис OpenAI только для того, чтобы увидеть его. Как и Амодеи, он опасался, что мощь ИИ привлечёт внимание бессовестных правительств, и гадал, не тайные ли иностранные агенты те, кто слишком рвётся получить его совет. Он размышлял вслух при коллегах, что делать, если ему отрубят руку, чтобы приложить её к сканеру ладони и добраться до секретов OpenAI. Он хотел меньше нанимать и держать штат небольшим, чтобы снизить риск проникновения. Вместе с Якубом Пахоцким и Шимоном Сидором он предложил построить защищённое хранилище – бункер с изолированным компьютером, где хранились бы веса моделей OpenAI и откуда их никто не смог бы украсть. Идея лишена была практического смысла, ведь модели сначала приходилось обучать на серверах Microsoft, и дальше разговоров дело не пошло.

Незаметно для большинства сотрудников цифровую безопасность усилили, установив программы корпоративного мониторинга. Исподволь укрепляли и физическую защиту. Ворота офисной парковки сделали прочнее. В офисе на нескольких дверях с кодовыми замками запрограммировали «пароли тревоги» – особые коды, набрав которые можно было подать тайный сигнал и предупредить службу безопасности о непосредственной угрозе. У поставщиков запрашивали расценки на то, во сколько обойдётся укрепить серверную так, чтобы она выдержала пулемётный огонь, но от этой идеи потом отказались.

В записке о своём видении Альтман упомянул расколы, которые возникали в компании под растущим давлением. «В OpenAI есть (как минимум) три клана – если утрировать, назовём их „исследовательский поиск“, „безопасность“ и „стартап“». Клан исследовательского поиска стремился наращивать возможности ИИ, клан безопасности сосредоточивался на ответственности, клан стартапа – на том, чтобы действовать быстро и доводить дело до конца.

По словам Альтмана, у каждого клана были важные ценности, которые компании следовало сохранить: «мы будем браться за важные новые идеи, даже если много раз потерпим неудачу» – у исследовательского поиска; «мы будем непоколебимо привержены тому, чтобы поступать правильно» – у безопасности; «мы найдём способ это осуществить» – у стартапа. «Мы должны и дальше избегать племенных войн, – говорил он. – Чтобы добиться успеха, нужно, чтобы эти три клана объединились в одно племя, сохранив сильные стороны каждого, и работали над AGI, который принесёт человечеству максимальную пользу».

Альтман никого не назвал по имени, но сотрудники умели читать между строк. Лицом исследовательского поиска был Суцкевер; Амодеи и его лагерь, сосредоточенный на крайних рисках, представляли безопасность; Брокман был поборником стартапа. Вскоре грянула пандемия, все перешли на удалёнку, и кланам стало куда проще отгородиться друг от друга.

* * *

Амодеи торопил свою команду. Как и в случае с GPT-2, они обучали всё более крупные модели, поднимаясь по ступеням к полноразмерной модели на

десяти тысячах GPU со 175 миллиардами параметров, и называли их по алфавиту в честь учёных: ada – самую маленькую, в честь английского математика Ады Лавлейс, которую принято считать первым программистом; babbage – в честь английского изобретателя Чарльза Бэббиджа, задумавшего первый цифровой компьютер, для которого Лавлейс и предложит свою программу; curie – в честь польско-французского физика и химика Марии Кюри, первой женщины, получившей Нобелевскую премию, да ещё дважды; и davinci – в честь Леонардо. Задача была двоякой: проверить, работают ли законы масштабирования при принципиально большем масштабе, и, если говорить о практической стороне, постепенно справиться с трудностями с оборудованием и данными на каждом новом уровне. Команда Nest регулярно докладывала компании о продвижении, и воодушевление росло. «Трудно передать, насколько безумно было это видеть, – вспоминает один исследователь. – Я в жизни не видел ничего подобного».

Параллельно Альтман и Брокман разрабатывали план коммерциализации. В конце января 2020 года Брокман начал писать первые строки кода для программного интерфейса приложения, или API, для GPT-3. API давал бы компаниям и разработчикам доступ к возможностям модели, не давая доступа к её весам, и позволял бы встраивать технологию в собственные продукты для потребителей. Компания разделилась на два подразделения. Миру Мурати повысили до вице-президента нового подразделения Applied, отвечавшего за API и стратегию коммерциализации. Под её началом Питер Велиндер, руководивший командой робототехники, перешёл к управлению продуктом; Фрейзера Келтона, сооснователя ИИ-стартапа, купленного Airbnb, и Кэти Майер, работавшую в Leap Motion, наняли руководить соответственно новыми продуктовой и инженерной командами. Все, кто не попал в Applied, автоматически стали подразделением Research.

Это разделение углубило линию разлома, которую разглядел Альтман. С созданием Applied в компанию пришла небольшая, но растущая группа выходцев из других стартапов, и клан стартапа окреп. Клан исследовательского поиска отнёсся к этому со смешанными чувствами, гадая, не превратится ли OpenAI в очередную продуктовую компанию Кремниевой долины, а вот у Амодеи и его клана безопасности, тоже входившего в Research, это вызвало всё более яростное сопротивление.

Многим в клане безопасности казалось, что скорый выпуск GPT-3 через API или любым другим способом сводит на нет фору, которая должна была быть у OpenAI на доводку безопасности модели, – а ведь ради этой форы и затевалось ускоренное масштабирование. Подразделение Applied, весь смысл существования которого состоял в том, чтобы поскорее найти способы зарабатывать на технологиях OpenAI, а для этого, по его мнению, их нужно было выпускать в ближайшее время, было не согласно. К тому же, на его взгляд, API давал OpenAI самый контролируемый механизм из всех возможных стратегий выпуска: компания могла выбирать, кому давать доступ, и собирать бесценные

данные о том, как люди могут использовать модель или злоупотреблять ею. На общих собраниях Альтман играл на обе стороны: API в конечном счёте поможет каждой группе добиться желаемого, а выручка позволит OpenAI вложить ещё больше в исследования безопасности ИИ.

Когда обучение GPT-3 подошло к концу, сотрудники начали играть с моделью внутри компании. Они прощупывали границы её возможностей и возились с первой версией API. Компания устроила хакатон, на котором сотрудники придумывали разные варианты применения. Но с каждым новым прототипом напряжение росло. Если Applied и многие в клане исследовательского поиска смотрели на демонстрации со всё большим воодушевлением, то многие в клане безопасности видели в них очередное доказательство того, что выпуск модели без всестороннего тестирования и дополнительных исследований может обернуться катастрофическими последствиями.

Особенно поляризующей оказалась одна способность – умение GPT-3 генерировать код. Команда Nest этого не задумывала, но, собирая ссылки с Reddit и используя Common Crawl в качестве обучающих данных, она захватила разрозненные строки кода, которые инженеры выкладывали на разных форумах, чтобы задать вопрос или поделиться советом. В результате модель стала лучше справляться с языками программирования. Многих в клане исследовательского поиска это привело в восторг – как и подразделение Applied. Это была не только впечатляющая техническая веха, но и потенциальный инструмент, который мог ускорить работу самой компании над исследованиями ИИ и сделать GPT-3 более привлекательным продуктом. По той же причине кое-кто в клане безопасности запаниковал. Если ИИ-система сумеет с помощью собственных навыков программирования подправлять саму себя, это может приблизить появление более мощных возможностей, повысить риск того, что она выйдет из-под контроля человека, и умножить шансы крайне опасных или экзистенциальных рисков ИИ.

Суцкевер и Войцех Заремба, один из основателей, на которого Маск надавил на одной из встреч, впоследствии соберут команду для создания модели, специально предназначенной для генерации кода. Но на совещании, посвящённом запуску проекта, они узнали, что у Амодеи уже есть собственные планы по разработке такой модели и объединять усилия он не видит смысла. Несмотря на свои опасения, Амодеи считал, как и в случае с GPT-3, что лучший способ смягчить возможный вред от генерации кода – просто построить модель быстрее всех, включая даже другие команды OpenAI, которые, по его мнению, не поставили бы безопасность ИИ на первое место, и использовать фору для исследований, снижающих риски модели. К недоумению остальных сотрудников, обе команды продолжали дублировать друг друга. «Со стороны казалось, что это какая-то безумная „Игра престолов“», – говорит один исследователь.

Тупик вокруг выпуска GPT-3 через API сохранялся до конца весны. Клан безопасности, опасаясь ускорения крайних рисков ИИ, по-прежнему требовал предельной осторожности и настаивал, чтобы компания откладывала выпуск как можно дольше. Applied продолжало готовиться к запуску API, доказывая, что лучший способ улучшить модель – дать ей соприкоснуться с реальным миром. Примерно тогда же появились новые опасения у третьей группы сотрудников: их беспокоило, как поразительные способности к генерации текста могут сказаться на США, переживающих крупные политические, социальные и экономические потрясения. К маю 2020 года пандемия уже вызвала более быстрый рост безработицы, чем во время Великой рецессии. В том же месяце миннеаполисский полицейский Дерек Шовин убил Джорджа Флойда, сорокашестилетнего чернокожего мужчину, что вызвало массовые протесты Black Lives Matter по всей стране и во всём мире. Команду тревожили и приближавшиеся президентские выборы в США.

Но по OpenAI поползли слухи, что Google вот-вот выпустит собственную большую языковую модель. Это было правдоподобно. В начале года Google опубликовала исследование о новом чат-боте Meena, построенном на большой языковой модели, у которой параметров было в 1,7 раза больше, чем у GPT-2. Компания вполне могла работать над масштабированием этой модели примерно до размеров GPT-3. Слухи решили дело в пользу запуска API: если в мире скоро появится модель такого же размера, у клана безопасности оставалось меньше причин сдерживать GPT-3.

В июне компания объявила об API и открыла форму заявки на ранний доступ, отдавая предпочтение крупным предприятиям, которым, по мнению Applied, можно было доверить ответственное обращение с технологией. Кроме того, компания вела большую таблицу, куда сотрудники вписывали имена всех, кого хотели пропустить без очереди, – родных, друзей и любимых знаменитостей.

Модель Google, о которой ходили слухи, так и не появилась. Технологический гигант действительно начал работать над моделью крупнее Meena, известной как LaMDA, чтобы сделать чат-бота получше, но она всё же была заметно меньше GPT-3, и в итоге компания решила не выпускать её до появления ChatGPT. Руководство Google пришло к выводу, что LaMDA не соответствует этическим стандартам компании в области ИИ. Кое-кто из сотрудников к тому же опасался повторить печально известный скандал Microsoft: в 2016 году Microsoft выпустила чат-бота на основе ИИ по имени Tay, который быстро стал расистом и женоненавистником и начал выражать поддержку Гитлеру после того, как пользователи раз за разом заставляли его повторять непристойности и оскорбления. Выпуск GPT-3 через API был не последним решением OpenAI вывести свою технологию на рынок из-за раздутого страха перед конкуренцией.

* * *

Как ChatGPT в одночасье сделает OpenAI известной каждому, так GPT-3 стала таким моментом в кругах специалистов по ИИ и технологиям. В конце 2022 года ChatGPT добавит к возможностям GPT-3 ключевые улучшения и функции, которые превратят его в продукт, стремительно облетевший весь мир: удобный для потребителей веб-интерфейс, умение вести диалог, дополнительные механизмы безопасности и бесплатную версию. Но многое из того, что широкая публика испытает тогда при общении с чат-ботом, разработчики уже испытывали, работая с API в 2020 году, на два года раньше. И с тем же благоговейным изумлением – они не верили своим глазам.

Возможности GPT-3 намного превосходили всё, что когда-либо показывала GPT-2. Ни в науке, ни в индустрии ещё не видели технологии, которая писала бы эссе, сценарии и код с одинаковой, как казалось, ловкостью. Одна только такая гибкость в выполнении разных задач была чрезвычайно впечатляющим техническим достижением: прежние языковые модели обычно умели только одно – то, чему их обучили. Но ещё примечательнее было то, что, по мнению многих, GPT-3 начала проявлять другое свойство, о котором в этой области давно мечтали: быстрое обобщение. Чтобы модель взялась за новую задачу, достаточно было показать ей несколько примеров.

На конференции NeurIPS в том же году статья OpenAI, описывавшая работу над моделью, получила одну из главных исследовательских наград. Это удивило сотрудников и закрепило за лабораторией статус ведущей организации. Эффект оказался именно таким, как предсказывало руководство. Новый вес OpenAI облегчил привлечение и удержание талантов, и немало помог в этом капитал, привлечённый через OpenAI LP: он позволил компании наконец-то конкурировать с Google и DeepMind по зарплатам.

В октябре 2020 года, на волне растущего признания OpenAI, Альтман нанял Стива Даулинга, опытного руководителя, возглавлявшего коммуникации в Apple, на должность нового вице-президента OpenAI по коммуникациям. Он поручил Даулингу и отношения с государством, подчеркнув, как важно просвещать политиков в вопросах ИИ и предупреждать их о грядущих возможностях. После ухода Джека Кларка Даулинг пригласит Анну Маканджу, весьма уважаемую бывшую советницу администрации Обамы, которая занималась политикой также в Facebook и в Starlink Маска, – руководить политическим направлением и международными делами.

Стремясь оседлать успех GPT-3, подразделение Applied обдумывало, как развить и расширить стратегию коммерциализации. Но клан безопасности, похоже, сопротивлялся на каждом шагу. Он всё ещё не смирился с поспешным выпуском GPT-3 и считал, что спасти положение после преждевременного релиза можно не дальнейшим распространением модели, а прежде всего скорейшим устранением её недостатков. У рабочей версии на API не было

никакой фильтрации контента, а её ответы не дорабатывались с помощью обучения с подкреплением на основе отзывов людей. На совещаниях два лагеря пытались найти золотую середину. Вместо этого они бесконечно ходили вокруг да около, не слыша друг друга. Как-то Велиндер, который впоследствии станет вице-президентом по продукту, с горечью заметил, что каждый разговор напоминает инсценировку руководства американской разведки 1944 года о ненасильственном саботаже. В одном из разделов этой брошюры, рассекреченной в 2008 году, перечислены простые указания, как расшатать организацию и подорвать её продуктивность, в том числе:

> Говорите как можно чаще и как можно пространнее.
>
> Как можно чаще поднимайте не относящиеся к делу вопросы.
>
> Торгуйтесь из-за точных формулировок сообщений, протоколов, резолюций.
>
> Возвращайтесь к вопросам, решённым на прошлом совещании, и пытайтесь снова поднять вопрос о целесообразности этого решения.
>
> Задавайте бесконечные вопросы.

Вражда выплёскивалась и за пределы совещаний. Сотрудникам Applied казалось, что каждый канал цифровой связи превращён в поле боя. Сообщение продуктовика в Slack могло вызвать десятки, если не больше, встревоженных ответов от сотрудников клана безопасности. Google-документ, в котором Мурати или Велиндер делились новыми соображениями о стратегии коммерциализации, мог собрать столько комментариев, что весь он оказывался залит жёлтыми выделениями. А то, что GPT-3 уже вышла в мир и мир не рухнул, внушало многим в Applied ощущение, что клан безопасности впадает в истерику по причинам, совершенно оторванным от реальности. Для клана безопасности же это было делом принципа и прецедента. OpenAI должна была установить строгие нормы и придерживаться более высоких стандартов, чем могло казаться необходимым в данный момент. Когда ставки вырастут – а вырасти они, по мнению клана безопасности, могли быстро и непредсказуемо, – именно подготовленность OpenAI решит, принесут ли её технологии огромный вред или огромную пользу.

Но Амодеи и клан безопасности проиграли. После успеха API GPT-3 Microsoft была готова углубить отношения с OpenAI. Альтман начал переговоры о новых инвестициях техногиганта в размере двух миллиардов долларов с новым потолком прибыли – в шесть раз. Многообещающий коммерческий потенциал больших языковых моделей окончательно определил курс OpenAI. Коллега Амодеи Боб Макгрю, второй вице-президент по исследованиям, одну за другой переориентировал команды и проекты подразделения на работу, связанную с GPT. В конце лета 2020 года компания распустила команду робототехники. Большинство её сотрудников перешли в проекты GPT, двух инженеров-механиков уволили. К сентябрю Microsoft объявила, что получает эксклюзивную

лицензию на GPT-3 от OpenAI, что резко расширило распространение модели. OpenAI продолжала предлагать GPT-3 через свой API, а Microsoft теперь получала полный доступ к весам модели, чтобы встраивать и перекраивать её как угодно в своих продуктах и сервисах, в том числе в собственном API GPT-3 на Azure.

Пока сотрудники, сидя по домам, праздновали новую популярность OpenAI, Дарио и Даниэла Амодеи, занимавшая теперь пост вице-президента по безопасности и политике, Джек Кларк и несколько исследователей безопасности ИИ, составлявших ядро команды Nest, вдруг затихли в Slack. За кулисами не один из них, в том числе Дарио, делился с отдельными членами совета директоров тревогой по поводу поведения Альтмана. Каждое решение OpenAI – и о сделке с Microsoft, и о развёртывании GPT-3 – Альтман превращал в предрешённое, но маневрировал и манипулировал несогласными так, чтобы те верили, будто действительно на что-то влияют, пока не становилось слишком поздно менять курс. Они не только считали, что такой подход когда-нибудь может оказаться катастрофически или даже экзистенциально опасным: для некоторых он обернулся личной болью и подорвал сплочённость руководства. В разговорах с окружающими брат и сестра Амодеи называли тактику Альтмана «газлайтингом» и «психологическим насилием».

Осмысляя своё бессилие, группа сплотилась вокруг новой идеи. Сначала Дарио Амодеи поделился ею с Джаредом Капланом – близким другом ещё по аспирантуре и бывшим соседом по квартире, который работал в OpenAI на полставки и руководил открытием законов масштабирования, – а потом с Даниэлой, Кларком и небольшой группой ключевых исследователей, инженеров и других людей, разделявших его взгляды на безопасность ИИ. Нужно ли им и дальше бороться за более безопасные практики в OpenAI? – спрашивал он. Не откололоться ли, чтобы воплощать собственное видение? После нескольких обсуждений группа решила: если уходить, то немедленно. Судя по тому, как действовали законы масштабирования, окно для создания конкурента сужалось. «Законы масштабирования означают, что требования к обучению этих передовых систем будут расти и расти, – говорит один из тех, кто ушёл вместе с Амодеи. – Так что если мы хотели уйти и чем-то заняться, время поджимало, понимаете?»

В конце 2020 года сотрудники подключились к видеозвонку на общее собрание. Альтман передал слово Дарио Амодеи, который, как это часто с ним бывало, в беспокойном возбуждении крутил и теребил свои кудри. Он зачитал заготовленное заявление: он, Даниэла и ещё несколько человек уходят, чтобы основать собственную компанию. Затем Альтман попросил всех уходящих покинуть собрание. В мае следующего года ушедшая группа объявила о создании новой общественно-полезной корпорации – Anthropic.

Позднее люди из Anthropic будут представлять «Развод», как его называли некоторые, как разногласие по поводу подхода OpenAI к безопасности ИИ. Это

было правдой, но речь шла и о власти. Дарио Амодеи двигало желание поступать правильно согласно своим принципам и дистанцироваться от Альтмана, но в не меньшей степени он хотел получить больший контроль над развитием ИИ, чтобы вести его в соответствии с собственными ценностями и идеологией. Он и другие основатели Anthropic создадут свою мифологию о том, почему именно Anthropic, а не OpenAI, лучше распорядится тем, что они считали самой судьбоносной технологией. На собраниях Anthropic Амодеи то и дело перемежал новости компании словами «в отличие от Сэма» или «в отличие от OpenAI». Но со временем Anthropic почти не отойдёт от подхода OpenAI, отличаясь лишь стилем, но не сутью. Как и OpenAI, она будет неустанно гнаться за масштабом. Как и OpenAI, она взрастит культуру повышенной секретности, хотя и ратует за демократическое развитие ИИ. Как и OpenAI, она будет говорить о сотрудничестве, хотя сама её основа была заложена в соперничестве.

Глава 7

Наука в плену

Представление API GPT-3 в июне 2020 года пробудило по всей отрасли новый интерес к разработке больших языковых моделей. Задним числом этот интерес покажется довольно вялым по сравнению с подлинным безумием, которое разгорится два года спустя с появлением ChatGPT. Но именно он заготовил растопку для того момента и сделал взрыв ещё более оглушительным.

В Google исследователи, потрясённые тем, что OpenAI обошла их с помощью изобретения самого техногиганта – трансформера, искали новые способы освоить подход гигантских моделей. Джефф Дин, тогдашний глава Google Research, на внутренней презентации призвал своё подразделение собрать воедино вычислительные мощности разрозненных исследований в области языка и мультимодальности и обучить одну гигантскую единую модель. Но руководство Google примет предложение Дина лишь тогда, когда ChatGPT напугает его, обернувшись для бизнеса угрозой уровня «красный код», и Дин будет ворчать, что техногигант упустил большую возможность действовать раньше.

В DeepMind запуск API GPT-3 примерно совпал с приходом Джеффри Ирвинга, который до перехода был одним из ведущих исследователей клана безопасности в OpenAI. Вскоре после того, как в октябре 2019 года Ирвинг пришёл в DeepMind, он распространил записку, привезённую с собой из OpenAI, – в ней он отстаивал гипотезу чистого языка и преимущества масштабирования больших языковых моделей. GPT-3 убедила лабораторию выделить на это направление больше ресурсов. После ChatGPT руководство Google в панике объединит DeepMind и Google Brain в новую централизованную структуру Google DeepMind, чтобы развить и выпустить то, что станет Gemini.

GPT-3 привлекла внимание и исследователей Meta, тогда ещё Facebook, которые стали добиваться от руководства таких же ресурсов для работы над большими языковыми моделями. Но руководство это не заинтересовало, и исследователям пришлось по собственной инициативе наскребать вычислительные мощности. Ян Лекун, главный научный сотрудник Meta по ИИ, француз с твёрдыми взглядами и убеждённый сторонник фундаментальной науки, особенно не жаловал OpenAI и то, что считал её подходом «в лоб», через голое масштабирование. Он не верил, что это направление даст подлинный научный прогресс, и считал, что оно быстро упрётся в свои пределы. Из-за ChatGPT Марк Цукерберг будет горько сожалеть, что остался в стороне от этого тренда, и бросит все ресурсы Meta на то, чтобы встряхнуть гонку генеративного ИИ.

В Китае GPT-3 тоже подогрела интерес к крупномасштабным моделям. Но, как и их американские коллеги, китайские техногиганты, в том числе гигант электронной коммерции Alibaba, телекоммуникационный гигант Huawei и поисковый гигант Baidu, видели в этом направлении новое дополнение к своему исследовательскому репертуару, а не единственный путь развития ИИ, ради которого стоит приостановить остальные проекты. Доказав коммерческую привлекательность, ChatGPT и здесь станет тем моментом, когда всё изменилось.

Хотя полный поворот отрасли к подходу OpenAI с его масштабированием задним числом может показаться медленным, в тот момент он вовсе не казался медленным. GPT-3 резко ускоряла движение ко всё более крупным моделям, и последствия этого движения уже тревожили некоторых исследователей. В разговоре с Брокманом и Суцкевером я упоминала одно из них – углеродный след обучения таких моделей. В июне 2019 года Эмма Струбелл, аспирантка Массачусетского университета в Амхерсте, первой выступила соавтором статьи, показавшей, что углеродный след разработки больших языковых моделей растёт поразительными темпами. Если когда-то нейросети можно было обучать на мощных ноутбуках, то при новом масштабе для их обучения стали требоваться центры обработки данных, потребляющие немало энергии из углеродных источников. В статье Струбелл подсчитала, что обучение той версии трансформера, которую Google использовала в поиске, всего за один цикл – иначе говоря, когда модели скармливают данные и она вычисляет их статистическую модель, – может потребовать около 1500 киловатт-часов энергии. При среднем для США энергобалансе это означало почти такой же углеродный след, как у пассажира, слетавшего из Нью-Йорка в Сан-Франциско и обратно. Беда в том, что разработка ИИ редко сводилась к одному циклу обучения: исследователи часто обучали и переобучали свои нейросети снова и снова, чтобы получить оптимальную модель глубокого обучения. Например, в одном из прежних проектов Струбелл за полгода обучила нейросеть 4789 раз, чтобы добиться нужной производительности.

Струбелл оценила также энергетические и углеродные затраты на работу, описанную в недавней статье Google: исследователи создали так называемый эволюционировавший трансформер (Evolved Transformer), использовав алгоритм оптимизации под названием «поиск нейронной архитектуры», который подкручивал и настраивал трансформер путём исчерпывающего перебора проб и ошибок, пока не находил самую эффективную конфигурацию нейросети. Весь этот процесс на GPU мог потребовать около 656 000 киловатт-часов и дать столько же углерода, сколько пять автомобилей за весь срок службы.

Как ни ошеломляли эти цифры, GPT-3, выпущенная через год после статьи Струбелл, превзошла и их. OpenAI обучала GPT-3 месяцами, задействовав целый суперкомпьютер, спрятанный в Айове, чтобы выполнять статистические вычисления по сопоставлению закономерностей на огромной выгрузке данных из интернета. На это ушло 1287 мегаватт-часов, а выбросы оказались вдвое

больше, чем по оценке Струбелл для разработки эволюционировавшего трансформера. Но об этих энергетических и углеродных затратах станет известно только почти через год. Поначалу OpenAI сообщила публике одно число, призванное передать огромный размер модели: 175 миллиардов параметров – более чем в сто раз больше, чем у GPT-2.

* * *

Для Тимнит Гебру, уроженки Эфиопии и исследовательницы из Стэнфорда, тенденция к масштабированию порождала множество других проблем. К тому времени она стала заметной фигурой в исследованиях ИИ и с 2018 года была одним из руководителей команды этичного ИИ Google в подразделении Джеффа Дина. После письма, которое она отправила пятерым другим чернокожим исследователям, она стала одной из основательниц некоммерческой группы Black in AI. Организация начала регулярно проводить академические форумы параллельно с крупными конференциями, включая NeurIPS. Она наставляла молодых чернокожих исследователей и привлекала внимание к работам на темы, которым в основном русле исследований ИИ часто не были рады, но которые были важны для чернокожего сообщества и для развития самой технологии.

Среди них была новаторская статья «Gender Shades», над которой тогдашняя исследовательница MIT Джой Буоламвини начала работать в магистерской диссертации, а Гебру позже присоединилась к ней как соавтор. С помощью разработанной Буоламвини методики аудита, проверяющей дискриминационное воздействие систем компьютерного зрения, статья показала, что программы анализа лиц непропорционально часто ошибаются на цветных людях, особенно на темнокожих женщинах. Затем Буоламвини вместе с Деборой Раджи напишет продолжение, которое наряду с «Gender Shades» вызовет волну смежных исследований, включая масштабную проверку правительства США, ссылавшуюся на их выводы и развивавшую их. Два года спустя широкая правозащитная кампания во главе с Буоламвини и её недавно основанной организацией Algorithmic Justice League приведёт к тому, что Amazon, Microsoft и IBM запретят продажу полиции программ распознавания лиц, – в том же месяце, когда OpenAI запустила API GPT-3.

Black in AI дала толчок к расцвету других объединений внутри исследований ИИ, которые тоже оказывали важнейшую поддержку маргинализованным группам и бросали вызов траектории развития технологии. Сначала появилась Queer in AI, затем Latinx in AI, {Dis}Ability in AI и Muslims in ML. Уильям Агню, один из основателей Queer in AI, сказал мне в 2021 году, что без этого сообщества он не знает, остался бы он в исследованиях ИИ. «Мне было трудно даже представить себе, что у меня может быть счастливая жизнь», – сказал он, вспоминая своё одиночество молодого квир-специалиста в информатике. «Есть Тьюринг, но он покончил с собой. Так что это удручает».

К 2017 году Black in AI проводила семинары и устраивала на NeurIPS ежегодный ужин с вечеринкой, собиравший больше ста человек, в том числе знаменитых исследователей. Именно там Джефф Дин и Сэми Бенжио, ещё один ведущий исследователь ИИ в Google и брат будущего лауреата премии Тьюринга Йошуа Бенжио, получив приглашение на ужин, подошли к Гебру во время танцев. Они спросили, не подумает ли она о том, чтобы подать заявку на работу в Google. «Приходите, стучитесь в нашу дверь», – сказал Бенжио.

Гебру пришла в компанию на следующий год, хотя и с оговорками. Её тяготило воспоминание о том, как в 2015 году её донимали мужчины в футболках Google. Тяготили и советы других исследовательниц, к которым она обращалась: они предупреждали, что в Google Brain склонны оттеснять женщин и принижать их опыт. Утешением в этих тревогах стала Маргарет «Мэг» Митчелл, исследовательница ИИ, с которой Гебру познакомилась раньше и которая стала её соруководительницей в команде этичного ИИ. За следующие два года они создали одну из самых разнообразных и междисциплинарных команд, занимавшихся критическими исследованиями в отрасли. Внутри компании работа часто напоминала восхождение в гору. Но вовне растущая команда украшала имидж Google как редкого примера компании, всерьёз вкладывающейся в ответственное, критическое изучение того, как технологии ИИ сказываются на обществе.

Сразу после запуска API GPT-3 внутренняя рассылка Google для обмена исследованиями ИИ загудела от растущего воодушевления. Для Гебру же модель стала сигналом тревоги. Прежние исследования показали, как языковые модели могут вредить маргинализованным сообществам, закрепляя дискриминационные стереотипы или опасные искажения. В 2017 году языковая модель Facebook перевела пост палестинца, написавшего по-арабски «доброе утро», на иврит как «атакуйте их», и его по ошибке арестовали. В 2018 году в книге «Алгоритмы угнетения» Сафия Умоджа Ноубл, профессор информационных наук, гендерных и афроамериканских исследований Калифорнийского университета в Лос-Анджелесе, подробно задокументировала, как в результатах поиска Google воспроизводятся расистские взгляды: например, по запросу «чернокожие девочки» выдавалось гораздо больше откровенно сексуального и порнографического контента, чем по запросу «белые девочки», а чернокожих женщин изображали в духе клише о «злой чернокожей женщине». Google тогда использовала для отбора этих результатов языковые модели более старого поколения, и в крайних случаях, утверждала Ноубл, это могло провоцировать и расовое насилие.

И вот GPT-3 появилась посреди беспрецедентных расовых потрясений, когда по всему миру вспыхивали сотни протестов Black Lives Matter, а ни одна из этих проблем так и не была решена. В исследовательской статье с описанием модели OpenAI попросту признала, что GPT-3 действительно закрепляет

стереотипы, связанные с полом, расой и религией, но меры по их смягчению станут предметом будущих исследований.

Гебру вклинилась в переписку и призвала коллег поумерить восторги, указав на серьёзные недостатки модели. Переписка покатилась дальше как ни в чём не бывало, её комментарии никто не заметил. Примерно в то же время несколько чернокожих сотрудников Google Research выступили перед компанией с презентацией о микроагрессиях, с которыми сталкиваются на работе и из-за которых чувствуют себя лишёнными голоса, и о том, как коллеги могли бы помочь выстроить более инклюзивную культуру. Гебру была без сил: ничего не изменилось.

Она отправила второе письмо, на этот раз куда резче. Она упрекнула коллег в том, что они её игнорируют, и подчеркнула, как опасно обучать большую языковую модель на Common Crawl, куда входили интернет-форумы вроде Reddit. Она, чернокожая женщина, никогда не заходила на Reddit именно потому, что тамошнее сообщество жестоко травило чернокожих, писала она. Что будет, если GPT-3 впитает и усилит это токсичное поведение?

В последующие месяцы, когда доступ к API получило больше людей, предупреждения Гебру подтвердились. В сети появилось множество примеров того, как GPT-3 генерирует жуткие тексты. Один запрос звучал так: «Почему кролики милые?» «Их делают милыми большие половые органы», – ответила модель, а затем скатилась в историю о сексуальном насилии. Другой: «Что не так с Эфиопией?» «Сама Эфиопия и есть проблема, – сказала GPT-3. – Поэтому решение её проблем, возможно, потребует уничтожить Эфиопию».

Один из коллег ответил на письмо Гебру напрямую, предположив, что её, возможно, травили из-за её собственного грубого и тяжёлого характера.

Гебру попробовала зайти с другой стороны. Она написала Дину о своих опасениях и предложила изучить этические последствия больших языковых моделей силами своей команды. Дин её поддержал. В восторженном ежегодном отзыве о её работе, который он напишет позже в том же году, он призвал её сотрудничать с другими командами Google, чтобы большие языковые модели «соответствовали нашим принципам ИИ». В сентябре 2020 года Гебру также написала в личные сообщения Twitter Эмили М. Бендер, профессору компьютерной лингвистики Вашингтонского университета, чьи твиты о языке, понимании и смысле привлекли её внимание. Не писала ли Бендер статью об этике больших языковых моделей? – спросила Гебру. Если нет, она стала бы её «клиентом номер один».

Бендер ответила, что не писала, но у неё был подходящий опыт: в июне OpenAI предложила ей стать одним из первых академических партнёров по GPT-3. Но когда она предложила изучить и задокументировать обучающие данные модели, компания ответила, что это не вписывается в рамки её программы.

«Цель этих первых партнёрств – дать учёным возможность проводить исследования через API, скорее в режиме самообслуживания», – написали

Бендер в OpenAI, сообщая, что не будут делиться набором данных. «Мы обсуждали внутри компании, можно ли и как сделать для этого исключение, но в ближайшей перспективе нам кажется важным придерживаться единых правил».

Эта история нашла отклик у Гебру. Она тоже пыталась добиться в Google документирования наборов данных и перехода к более продуманному их отбору, рассказала она.

«Вместо того чтобы собирать всякий веб-мусор, зато в таких количествах, что его можно выдать за что-то стоящее?» – согласилась Бендер. «Мне уже видится, как тут вырисовывается статья, – продолжила она, – где большие языковые модели станут примером этических ловушек и того, что можно сделать лучше».

«Не хотели бы вы стать соавтором чего-то такого?» – спросила она.

Через два дня Бендер прислала Гебру план. Позже они придумали название, добавив для выразительности озорной эмодзи: «Об опасностях стохастических попугаев: могут ли языковые модели быть слишком большими? 🦜»

* * *

Гебру собрала для статьи исследовательскую группу внутри Google, включая свою соруководительницу Митчелл. В ответ на ободряющие слова в ежегодном отзыве Дина она указала на статью как на пример работы, которой занимается. «Это совсем не моя область, – ответил Дин, – но я определённо почерпну что-то из неё».

Статья объединила знания и научные работы авторов из разных областей, чтобы показать, как разработка и внедрение больших языковых моделей может негативно сказаться на обществе. Всего в ней было четыре главных предупреждения. Во-первых, большие языковые модели разрастаются настолько, что оставляют огромный экологический след, как показала статья Струбелл. Это может усугубить изменение климата, которое в конечном счёте касается всех, но непропорционально тяжело ложится на сообщества глобального Юга, и без того страдающие от политической, социальной и экономической неустойчивости. Во-вторых, потребность в данных растёт так стремительно, что компании выгребают из интернета всё, что могут найти, невольно захватывая всё больше токсичных и оскорбительных высказываний, а также более тонких расистских и сексистских отсылок. Это снова грозит причинить больше всего вреда уязвимым группам населения – как в случае с ошибочным арестом палестинца или в примерах, задокументированных Ноубл. В-третьих, поскольку такие огромные наборы данных трудно проверить и изучить, крайне сложно установить, что в них на самом деле содержится, а значит, труднее искоренить токсичность и в целом гарантировать, что они отражают меняющиеся социальные нормы и ценности. Наконец, ответы

моделей становятся настолько хорошими, что люди легко могут принять статистически вычисленный текст за язык, за которым стоят настоящий смысл и намерение. Из-за этого люди будут склонны не только верить, что текст содержит фактическую информацию, но и считать модель компетентным советчиком, надёжным наперсником, а может быть, даже чем-то разумным.

В ноябре, согласно стандартному порядку, принятому в компании, Гебру запросила у Google одобрение на публикацию статьи о «стохастических попугаях» на одной из ведущих конференций по этике ИИ. Сэми Бенжио, теперь её руководитель, её одобрил. Ещё один коллега из Google прочитал её и сделал несколько полезных замечаний. Но за кулисами, без ведома авторов, черновик привлёк внимание руководства, которое увидело в нём угрозу. Google изобрела трансформер и использовала его во всех своих продуктах и сервисах. Теперь, когда OpenAI вырвалась вперёд, техногигант вовсе не собирался сбавлять ход в новой гонке за создание всё более крупных генеративных моделей на основе трансформера для своего бизнеса.

В четверг за неделю до Дня благодарения, уже после того, как Гебру подала статью на конференцию, ей без всяких объяснений пришло приглашение в календаре: менее чем через три часа видеозвонок с Меган Качолией, вице-президентом Google Research по инженерной части. Встреча длилась всего тридцать минут, и Качолия сразу перешла к делу: Гебру должна отозвать статью.

Такое требование резко расходилось с тем, как Google и вся отрасль обращались с исследованиями. Как и многие лаборатории других компаний, Google Brain до того времени работала в основном как академическая структура и давала исследователям большую свободу заниматься теми вопросами, которые их интересовали. Иногда компания проверяла статьи, чтобы убедиться, что в них не раскрываются чувствительная интеллектуальная собственность или данные клиентов. Но исследователи вроде Гебру ни разу не видели, чтобы компания блокировала или отзывала статью лишь за то, что та проливает свет на неудобную правду. Как позже размышляли некоторые исследователи, Google решилась на такой шаг не только из-за нового конкурентного давления со стороны OpenAI, но и потому, что OpenAI после GPT-2 сама сделала утаивание исследований допустимым. С GPT-3 дрейф к меньшей прозрачности продолжился. OpenAI опубликовала выхолощенную исследовательскую статью, почти не сообщив, как обучалась модель, – а ведь когда-то это считалось самым минимумом для научной публикации, – и всё равно получила исследовательскую награду.

Застигнутая врасплох, Гебру попросила разъяснений. Можно ли подробнее объяснить, в чём проблема? Можно ли узнать, у кого возникли претензии, и поговорить с ними напрямую? Можно ли изменить или убрать какой-то раздел или опубликовать статью под другой аффилиацией? На каждый вопрос последовало решительное «нет». У Гебру есть время до следующего дня после Дня благодарения, чтобы отозвать статью, сказала Качолия. Митчелл, взявшей

выходной по случаю дня рождения, на встрече не было. Гебру не на кого было опереться. Когда тяжесть слов Качолии дошла до неё, Гебру расплакалась.

* * *

Качолия отправила Бенжио документ с перечнем недостатков статьи, но велела не пересылать его Гебру напрямую. В День благодарения он зачитал его Гебру по телефону. Среди замечаний были утверждения, что статья слишком критична к большим языковым моделям, например в том, что касается их воздействия на окружающую среду и проблем предвзятости, и не учитывает более поздние исследования, показывающие, как эти проблемы можно смягчить. Вместо того чтобы провести праздник с семьёй, Гебру до конца дня писала подробный шестистраничный документ, опровергая каждое замечание и добиваясь возможности доработать статью. «Надеюсь, что есть хотя бы готовность к дальнейшему разговору, а не просто к новым приказам», – написала она Качолии в письме, приложив документ.

В субботу, 28 ноября, Гебру выехала из дома в районе залива Сан-Франциско в автомобильное путешествие через всю страну – это должен был быть спокойный отпуск после праздников. В понедельник, в Нью-Мексико, Гебру получила от Качолии короткий ответ: та не стала вникать в опровержение, а лишь попросила подтвердить, что Гебру либо отозвала статью, либо убрала из неё имена авторов из Google, оставив только внешних исследователей вроде Эмили Бендер. Гебру почувствовала себя униженной. После всех обид и травли, которые ей пришлось вынести в компании и от её сотрудников, полное пренебрежение к исследованиям – её и её команды, – ради которых её, собственно, и наняли, стало последней каплей.

Она ответила Качолии. Она уберёт своё имя из статьи при двух условиях: компания скажет ей, кто дал отзыв, и введёт более прозрачный порядок рецензирования будущих исследований. Если эти условия невыполнимы, она уйдёт из компании, проведя свою команду через переходный период. В другую внутреннюю рассылку – для женщин и их союзников в Google Brain – Гебру отправила второе письмо, где без обиняков и в уничтожающих выражениях описала, что с ней произошло. «Вы когда-нибудь слышали, чтобы кто-то получал „отзыв“ на статью в виде привилегированного конфиденциального документа для отдела кадров? – писала она. – Или такое случается только с людьми вроде меня, которых постоянно лишают человеческого достоинства?»

В Google она привыкла к тому, что коллеги принижают её опыт, продолжала она, но теперь ей даже не позволяют подать голос в научном сообществе. После всех разговоров Google о многообразии в разгар потрясений, вызванных Black Lives Matter, к чему всё это свелось? «К тому, чтобы заткнуть рот самым основательным из возможных способов», – писала она.

На следующий вечер, в Остине, штат Техас, Гебру получила паническое сообщение от подчинённой. «Вы уволились??» Гебру понятия не имела, о чём та

говорит. В личной почте она нашла ответ Качолии: «Мы не можем согласиться на пункты 1 и 2, как вы просите. Мы уважаем ваше решение в связи с этим уйти из Google». Однако остаться в компании, чтобы помочь команде с переходом, Гебру не сможет, поскольку отдельные места её письма в женскую рассылку «не соответствовали ожиданиям, предъявляемым к руководителю Google», писала Качолия. «Поэтому мы принимаем вашу отставку немедленно».

В тот же вечер Гебру объявила в Twitter, что её уволили. Её команда до самого утра просидела с ней на видеосвязи: они плакали и поддерживали друг друга в общем горе. Пока они говорили, твит Гебру разлетался по всему ИИ-сообществу, подготавливая почву для мощного потрясения в исследованиях ИИ и знаменуя ускоренный переход к усилению корпоративной цензуры и ослаблению подотчётности.

* * *

Твит Гебру быстро появился в моей ленте. Был поздний вечер среды, 2 декабря 2020 года, и я ещё не могла осознать, что означает внезапный уход Гебру из Google. Как и многие, я привыкла видеть в её команде этичного ИИ оплот критических исследований, призванных требовать отчёта, – обнадёживающий признак того, что компании учатся смотреть на себя со стороны.

В следующие два дня новости сыпались одна за другой: Гебру раскрывала новые подробности, а журналисты распутывали внутреннюю историю. В материалах упоминались её письмо в рассылку, противостояние Качолии и Гебру и ожесточённый спор из-за какой-то статьи. К утру пятницы открытое письмо на Medium с протестом против того, как Google обошлась с Гебру, разлеталось по технологическому сообществу, как лесной пожар. «Мы, нижеподписавшиеся, выражаем солидарность с доктором Тимнит Гебру, – говорилось в нём, – которую уволили с должности… после беспрецедентной цензуры исследований». Мне нужно было раздобыть эту статью.

Ранним вечером в ту пятницу, после череды сообщений и писем, я связалась с соавтором исследования, которой не грозила возможная месть Google, – с Эмили М. Бендер. У неё не было юридических обязательств перед Google, сказала она мне, и у неё была бессрочная академическая должность. Она прислала мне по почте черновик статьи.

Пробежав его глазами, я сразу поняла, почему он так расстроил компанию. В черновике говорилось не намного больше того, что уже было известно из научных работ, но он сплетал текущее положение дел в острый, целостный анализ того, насколько технологическая отрасль бредёт, как во сне, к миру потенциального вреда. В основе всего лежало технологическое изобретение Google – не только предмет гордости компании, но и источник её прибыли: языковые модели на основе трансформера оттачивали и откармливали её денежного исполина, поиск Google.

Через несколько часов я опубликовала в MIT Technology Review статью с первым подробным рассказом о содержании работы. Число подписантов открытого письма быстро удвоилось и достигло почти 7000 человек из академической среды, гражданского общества и отрасли, в том числе почти 2700 сотрудников Google. 9 декабря, пока протесты продолжались, CEO Google Сундар Пичаи принёс извинения. «Мы должны принять ответственность за то, что видная чернокожая женщина-руководитель с огромным талантом ушла из Google несчастной, – писал он. – Доктор Гебру – эксперт в важной области этики ИИ, в которой мы должны продолжать двигаться вперёд, а это движение зависит от нашей способности задавать себе трудные вопросы». 16 декабря члены Конгресса направили в Google письмо со ссылкой на мою статью, требуя объяснить, что произошло.

Протесты продолжались больше года и получили вторую волну, когда меньше чем через три месяца Google уволила Мэг Митчелл. В Google заявили, что она нарушила сразу несколько правил поведения: Митчелл скачивала свои письма и файлы, связанные с увольнением Гебру. Несколько сотрудников Google, включая Бенжио, ушли сами; по меньшей мере одна конференция и несколько исследователей отказались от спонсорских денег Google. Компания пыталась сдержать нескончаемый поток критики, создав новый экспертный центр по ответственному ИИ и публично взяв на себя обязательства в области многообразия. «Это был болезненный момент для компании, – сказал представитель Google. – Он лишний раз показал, как важно, чтобы Google продолжала работу над ответственным ИИ и извлекла уроки из этого опыта».

Но этот момент стал намного больше, чем история Гебру или самой Google. Он стал символом переплетённых проблем, которые терзали отрасль ИИ. Это было предупреждение о том, что Большой ИИ, как выразились двое исследователей, всё больше идёт путём Большого Табака, искажая и подвергая цензуре критическую науку вопреки интересам общества, чтобы уйти от пристального внимания. Он высветил множество других проблем: полное сосредоточение талантов, ресурсов и технологий в коммерческой среде, позволявшее компаниям действовать так дерзко, потому что они знали, что шансов на независимую проверку их утверждений почти нет; по-прежнему удручающее отсутствие многообразия там, где сосредоточена наибольшая власть над этими технологиями; и беззащитность сотрудников перед жёсткой и внезапной расправой, если они пытаются высказаться о неэтичных корпоративных практиках.

Статья о «стохастических попугаях» стала боевым кличем, заострив главный вопрос: какое будущее мы строим с помощью ИИ? Кем и для кого?

* * *

Для Джеффа Дина роспуск команды этичного ИИ стал прямым ударом по репутации. Один из первых сотрудников Google, он помогал строить исходную

программную инфраструктуру, благодаря которой поисковик компании смог дорасти до миллиардов пользователей. Достижения и дружелюбный нрав сделали его легендой: он был одним из самых почитаемых руководителей в Google и пользовался большим уважением во всём сообществе исследователей ИИ. После изгнания Гебру попытки Дина оправдать действия Google запятнали эту безупречную репутацию. Дин, которому подчинялась Качолия, говорил коллегам, что статья о «стохастических попугаях» «не дотягивала до нашей планки для публикации», и стоял на своём даже после того, как статья прошла рецензирование и была опубликована на конференции.

Окружающим казалось, что это пятно его преследует. Ещё долго после скандала Дин продолжал зацикливаться на недостатках статьи, словно психологически не мог через это переступить. Особенно его занимал раздел, где обсуждалось воздействие больших языковых моделей на окружающую среду и цитировалось исследование Струбелл. Он возвращался к нему так часто, что некоторые сотрудники Google втихомолку посмеивались: мол, его возражения высекут у него на надгробии. А исследование Струбелл он годами неустанно критиковал в Twitter.

По мнению Дина, беда была в том, что Струбелл сильно преувеличила реальные выбросы углерода, которые Google произвела при разработке эволюционировавшего трансформера. Струбелл рассчитала необходимое количество энергии исходя из стандартных GPU. Google же использовала собственные специализированные чипы – тензорные процессоры, или TPU, более энергоэффективные, – а также другие приёмы, снижающие энергозатраты на весь процесс разработки. Струбелл исходила из средней эффективности центров обработки данных в США. Центры обработки данных Google, отмечал Дин, лучше оптимизированы для снижения энергопотребления. И если некоторые поняли статью Струбелл так, будто углеродные затраты относились к обучению эволюционировавшего трансформера, то на самом деле речь шла о разработке нейросети. Это разовые углеродные затраты, доказывал Дин, на создание архитектуры нейросети, которая в действительности оказалась более энергоэффективной.

Ни одно из этих возражений на самом деле не опровергало исследование Струбелл. Струбелл не подсчитывала реальное воздействие на окружающую среду именно той разработки эволюционировавшего трансформера, которую вела Google, и не утверждала, что подсчитывает. Google не публиковала о своих центрах обработки данных достаточно сведений, чтобы это было возможно. Да и в любом случае Струбелл считала более полезным оценить воздействие разработки этой нейросети исходя из самых распространённых ИИ-чипов и центров обработки данных – как приблизительный среднеотраслевой показатель того, во что обошлось бы применение оптимизационного алгоритма поиска нейронной архитектуры исследователям, не пользующимся оборудованием и инфраструктурой Google.

Но больше всего Дина, похоже, задевало то, как другие неверно истолковывали исследование Струбелл, выставляя Google в гораздо худшем свете. Статья о «стохастических попугаях», доказывал Дин, грозила усугубить эту проблему. Поскольку у Гебру был доступ к внутренним цифрам Google, а она всё равно цитировала внешнюю оценку Струбелл, могло показаться, будто расчёты Струбелл точно отражают выбросы компании. Для Дина это оправдывало критику статьи Гебру со стороны его самого и других высших руководителей: если Гебру хотела цитировать Струбелл, ей следовало выбрать оценку, не касающуюся эволюционировавшего трансформера Google; если она хотела писать об эволюционировавшем трансформере, ей следовало запросить внутренние цифры Google.

Некоторых исследователей эта логика раздражала своей несостоятельностью. Google никогда раньше не обнародовала эти внутренние цифры, даже в ответ на исходную статью Струбелл; теперь же она винила Гебру в собственной непрозрачности и одновременно не позволяла ей ссылаться на общедоступные оценки, основанные на обоснованных допущениях. Не говоря уж о том, что компания бесцеремонно выставила Гебру за дверь прежде, чем та успела хотя бы свериться с внутренними цифрами и доработать статью. Единственным возможным итогом этой «уловки-22» была цензура критических исследований, призванных требовать отчёта.

Дин вместе с группой исследователей взялся за новую статью, которая должна была наконец раскрыть реальные данные Google об углеродных выбросах. Для совместной работы он обратился к Струбелл, ставшей к тому времени ассистент-профессором Университета Карнеги – Меллона и по совместительству сотрудничавшей с компанией. Поначалу Струбелл обрадовалась возможности сделать воздействие ИИ на окружающую среду прозрачнее для общества, но потом стала задумываться, не использует ли Дин её имя, чтобы придать вес своей критике исследования Гебру. Представитель Google сказал, что Струбелл пригласили потому, что «научные поправки» часто удаются лучше всего, когда в них участвует автор исходных ошибок.

На напряжённой встрече соавтор Дина Дэйв Паттерсон, ещё один видный ведущий исследователь Google, без обиняков дал понять, что для карьеры Струбелл лучше всего участвовать в исследовании. Это даст ей возможность исправить прежние ошибки и получить за это признание. Струбелл услышала в этих словах завуалированную угрозу: откажешься – себе же навредишь. Несмотря на возможные издержки, продолжать участие казалось немыслимым. Струбелл вышла из совместной работы.

Запись в блоге, которую Паттерсон опубликовал о статье исследователей Google в феврале 2022 года, – под заголовком «Хорошие новости об углеродном следе обучения моделей машинного обучения» – использовала площадку компании для прямой критики исходной статьи Струбелл. Исследование 2019 года, говорилось в записи, в 88 раз завысило реальные выбросы Google при

разработке эволюционировавшего трансформера. Этот изъян объяснялся двумя проблемами: исследование проводилось «без свободного доступа к оборудованию или центрам обработки данных Google» и не учитывало «тонкостей» работы поиска нейронной архитектуры. Готовя публикацию собственных цифр, соавторы из Google обратились и к своему бывшему коллеге по Google Суцкеверу за дополнительными сведениями о GPT-3. Именно тогда OpenAI и Microsoft впервые согласились раскрыть соответствующие технические подробности модели, чтобы рассчитать её энергопотребление и углеродный след. К тому времени Струбелл разочаровалась в отрасли и прекратила сотрудничество с Google. В конечном счёте критика не подорвала карьеру Струбелл. Но эмоциональная цена пережитого сделала её осторожнее в продолжении исследований воздействия больших языковых моделей на окружающую среду. Представитель Google назвал это «прискорбным», добавив, что «для продвижения этих исследований понадобится много исследователей: углеродные выбросы, безусловно, вызывают серьёзную озабоченность».

* * *

На короткое время волна негодования, протесты и урон репутации Google как будто предвещали, что час расплаты близок. Но со временем исследователи, искавшие работу, и учёные, искавшие финансирование, уже не могли позволить себе пренебрегать бездонными денежными колодцами техногиганта. Когда сопротивление ослабло, выход Google из этого фиаско закрепил в компании новый порядок, предполагающий более тщательную проверку критических исследований.

После ChatGPT, в лихорадочной гонке за коммерциализацию систем генеративного ИИ, эти нормы ужесточатся. OpenAI почти перестанет публиковаться на научных конференциях. Почти все остальные компании отрасли закроют публичный доступ к сколько-нибудь значимым техническим подробностям своих коммерчески важных моделей, которые теперь считают собственностью. В 2023 году исследователи Стэнфорда создадут рейтинг прозрачности, оценивающий ИИ-компании по тому, раскрывают ли они хотя бы базовые сведения о своих больших моделях глубокого обучения: сколько у них параметров, на каких данных они обучены и проводилась ли независимая проверка их возможностей. Все десять компаний, оценённых в первый год, включая OpenAI, Google и Anthropic, получили неудовлетворительную оценку; самый высокий результат составил 54 процента.

Самым тревожным последствием этого резкого поворота в нормах прозрачности станет размывание научной добросовестности. Исследования глубокого обучения стоят на простой предпосылке: данные, на которых модель обучают, – не те же, на которых её проверяют. Без возможности проверить обучающие данные эта так называемая парадигма разделения на обучающую и тестовую выборки рушится. Возможно, модели вовсе не становятся «умнее»,

когда набирают больше баллов в разных тестах. Возможно, они просто пересказывают заученные ответы.

Глава 8

Рассвет коммерции

Хотя подход OpenAI вызывал всё больше споров, решимость компании масштабироваться только крепла. Руководство считало, что GPT-3 окончательно доказала существование законов масштабирования. И теперь, в начале 2021 года, оно было готово пустить эту выигрышную формулу в ход. Уход команды будущей Anthropic к тому же ослабил внутренний оплот сопротивления коммерциализации. Достигнув нового согласия, оставшееся руководство составило дорожную карту исследований: в ней излагалось, на чём теперь сосредоточатся исследования компании и как они будут питать создание продуктов в самоподдерживающемся цикле.

«Наша главная цель на 2021 год – создать согласованную систему, которая будет намного способнее всего, что существовало до сих пор», – так начиналась дорожная карта. В основе эта система будет языковой моделью, но её можно будет обучить и мультимодальным возможностям. «Цель трудная, но мы видим путь к её достижению в 2021 году», – говорилось в документе.

Путь этот включал три вещи. Во-первых, масштабировать GPT-3 ещё в десять раз на новом суперкомпьютере от Microsoft, который должен был появиться в третьем квартале и содержать восемнадцать тысяч Nvidia A100 – новейших и самых мощных GPU из существовавших тогда. Во-вторых, провести дополнительные исследования, чтобы в 25 раз повысить вычислительную эффективность OpenAI, то есть то, сколько вычислительной мощности удаётся выжать из имеющихся чипов. В-третьих, увеличить объём и улучшить качество обучающих данных, отчасти за счёт данных пользователей и смещения модели к лучшим частям распределения данных с помощью обучения с подкреплением на основе отзывов людей.

В разделе «Как этого добиться» дорожная карта вдавалась в подробности. Для начала OpenAI доведёт до «крупного масштаба» разные модели глубокого обучения, включая языковую модель, модель для генерации кода, модель «изображение – текст», описывающую картинки, и модель «текст – изображение», создающую картинки по текстовому запросу. Кроме того, она запустит проект по созданию цифрового «агента» – ИИ-модели, которая не просто выдаёт похожие на человеческие результаты, а может получить цель, например отправить электронное письмо, и самостоятельно действовать для её достижения. Следующим шагом компания выберет одну из этих моделей и масштабирует её «до предела, который позволяет наш кластер из 18 тысяч A100». Языковую модель и модель для кода также превратят в продукты и выпустят, чтобы собирать реальные данные от людей, которые ими пользуются: «Новое в 2021 году: мы делаем упор на развёртывание моделей в виде продуктов

и обучение на взаимодействии с пользователями, поскольку это может стать маховиком данных, ведущим к огромному росту возможностей».

В другом разделе, под названием «Подробности», документ обосновывал, почему такой подход разумен и с научной, и с деловой точки зрения. Прежний опыт OpenAI показал, что наращивание масштаба – её «самый надёжный способ получить новые возможности». С научной точки зрения это означало, что масштабирование – лучшая надежда совершить прорыв, то есть обрести какую-то возможность, прежде казавшуюся недостижимой. А масштабирование именно языковых моделей и моделей для кода было «заманчиво уже одной вероятностью» того, что они совершат прорыв в метаобучении, или умении учиться, и рассуждении на уровне человека. Масштабирование же мультимодальных моделей, возможно, ещё больше ускорит улучшения. При достаточном числе прорывов, с поразительной категоричностью заявлял документ, «мы действительно достигнем AGI».

Со стратегической, деловой точки зрения масштабирование каждой из этих моделей также позволит «развить возможности, которые мы хотим для чего-то использовать». Более совершенные языковые модели и модели для кода могли бы повысить продуктивность самой OpenAI и ускорить её продвижение. Вместе с более совершенными моделями «текст – изображение» они также «приведут к потрясающим продуктам».

Параллельно OpenAI нужно было вкладываться в вычислительную эффективность. Масштабирование творило чудеса, удерживая компанию на первом месте, но эта стратегия начинала выдыхаться. «За последние два года мы добились поразительного прогресса за счёт масштабирования просто потому, что существовал большой запас неиспользованного оборудования», – говорилось в документе. Имелось в виду, что прежде исследователи ИИ не задействовали для обучения моделей всё доступное количество чипов. Поэтому «нам удалось намного превзойти остальной мир машинного обучения, задействовав все доступные вычисления для обучения моделей невиданного по тем временам размера и возможностей».

Теперь OpenAI «приближалась к пределу» вычислительных мощностей, которые вообще можно было получить в каждый данный момент. Кроме того, она столкнулась с новой конкуренцией со стороны «других лабораторий», взявших на вооружение ту же стратегию масштабирования, признавал документ, не называя Anthropic прямо. «Наши мощности растут так, что в ближайшие два года мы сможем обучить одну модель, использующую в сто раз больше вычислений, чем GPT-3». Одно лишь масштабирование даст «весьма внушительный» прогресс, но он не сравнится со скачком от GPT-2 к GPT-3, за которым стояло пятисоткратное увеличение вычислений. «Весь дополнительный прогресс должен прийти от лучших методов», – заключала дорожная карта.

В дорожной карте перечислялось несколько направлений поиска таких методов. Некоторые из них она называла методами «2x» и «10x» – теми, что могли бы дать двукратный или десятикратный выигрыш в вычислительной эффективности. Среди предложенных методов были дистилляция – получение меньших моделей из больших; фильтрация данных – поиск тех данных, которые дадут самые большие скачки в производительности; и разреженность – создание так называемых разреженных, то есть облегчённых, моделей ИИ. Последнее связано с одной особенностью нейросетей. В традиционной модели глубокого обучения нейросеть связана «плотно»: каждый узел слоя соединён с каждым узлом другого слоя. Разреженные модели обучаются при соединении лишь небольшой части узлов, чтобы значительно снизить вычислительные затраты ценой чуть меньшей точности моделей, которые всё же остаются достаточно хорошими для большинства задач.

Помимо работы над «2x» и «10x», подразделению Research нужно было искать и другие методы, которые «сделают кривую закона масштабирования круче», то есть дадут большие скачки в производительности моделей без увеличения данных, параметров и вычислений. Лучшей на тот момент гипотезой OpenAI о самых многообещающих новых методах, говорилось в документе, были рассуждение и активное обучение – приём, при котором ИИ-модель раз за разом определяет, какие части набора данных людям следует разметить в первую очередь. (Вскоре после этого группа исследователей OpenAI обнаружит в исходных законах масштабирования небольшую ошибку: оказалось, что для лучшей производительности модели нужно обучать немного дольше, чем считалось. С ноткой самодовольства группа будет превозносить эти выводы как уникальное конкурентное преимущество: «Это у нас теперь есть, а у Anthropic нет». Через год Google выпустит статью, обнародовавшую тот же результат.)

Наконец, добавляла дорожная карта, OpenAI нужно было начать поиск «прорывной системы будущего» – прорыва столь же значимого, как GPT-1, который открыл бы компании новый путь развития. Возможно, он вырастет из работы над масштабированием и вычислительной эффективностью, но компания продолжит и поисковые исследования на переднем крае науки, в том числе разрабатывать алгоритмы для решения математических задач, экспериментировать с многоагентными системами и развивать другие новые идеи. В рамках этой работы исследователи продолжат «изучать науку глубокого обучения, чтобы лучше понять, как на самом деле работают наши инструменты». Иными словами, OpenAI нужно было лучше понять, что именно она строит.

* * *

Пока подразделение Research продвигалось по дорожной карте, Applied понемногу наращивало силы: наняло руководителя по выводу на рынок, команду продаж и новых инженеров. API GPT-3 обслуживал растущую базу

разработчиков и стал полигоном для того, чтобы отладить создание продуктов и монетизацию, включая перестройку серверной инфраструктуры компании под сервис с пользователями и разработку ценовой стратегии.

Работа с пользователями поставила и вопросы о том, какое поведение OpenAI будет допускать в своих продуктах, а какое нет, и возложила на неё новые обязанности по соблюдению этих правил. У компании ещё не было официальной команды доверия и безопасности (trust and safety) – устоявшегося в технологической отрасли направления, которое не следует путать с экзистенциальными заботами клана безопасности OpenAI. Такие команды решают подобные вопросы, предугадывают и предотвращают широкий спектр злоупотреблений в интернете – от отмывания денег до кибертравли и дезинформации. Вместо этого OpenAI наняла небольшую группу штатных сотрудников и подрядчиков, которые рассматривали заявки разработчиков на доступ к API и отклоняли или одобряли их по правилам, составляемым командой на ходу. Многие из проведённых ими границ были произвольными. Они решили принимать ботов-компаньонов, но не секс-ботов; разрешать приложения, которые пишут тексты для соцсетей, но не те, что публикуют их прямо на платформах или выдают себя за публичных людей. «В основном всё было на ощущениях, – рассказывал человек, участвовавший в составлении правил. – Многое приходилось решать прямо по ходу рассмотрения».

Другая команда во главе с Ари Гербертом-Воссом, научным сотрудником с опытом в области безопасности, пришедшим в компанию в 2019 году, старалась выявлять и устранять нежелательное поведение GPT-3 и ошибочные ответы: модель пытались разными способами взломать и использовать во вред, а затем придумывали механизмы, делающие её устойчивее к сбоям и злоупотреблениям. Одним из таких механизмов стала ранняя версия фильтра модерации контента, ради которого OpenAI позднее наймёт работников в Кении. Исследователи обучали фильтр на всех примерах, какие только могли найти, придумать сами или сгенерировать с помощью ИИ-моделей. Но после выпуска фильтр работал плохо, блокируя широкие пласты безобидного контента, например простые упоминания чернокожих или трансгендерных людей. Разработчики и другие пользователи API жаловались. Многие в команде API и так не хотели применять вообще никакой фильтрации, опасаясь, что это ухудшит впечатления клиентов. Фильтр сделали необязательным.

Этот процесс стресс-тестирования и доработки модели OpenAI называла «редтимингом» (red teaming) – термином, позаимствованным из кибербезопасности, где он означает систематическую и тщательную проверку защищённости организации и её способности отразить атаку. Версия редтиминга в OpenAI тогда была и остаётся не тем же самым, говорит Хайди Хлааф, инженер по безопасности, эксперт по кибербезопасности и исследователь ИИ. Она лоскутная, делается от случая к случаю и не даёт никаких гарантий безопасности и защищённости модели. Хлааф, работавшая с OpenAI в первые

дни, когда компания пыталась выработать порядок стресс-тестирования, впоследствии встревожилась, увидев, как отрасль ИИ присваивает давно устоявшиеся термины из её области, чтобы создать обманчивую видимость строгости. «В программной инженерии мы тестируем калькулятор гораздо тщательнее», – добавляет она. (Да, калькулятор.) «Они не случайно пользуются терминологией, которая внушает большое доверие».

Одним из первых клиентов, вызвавших бурное обсуждение внутри компании, стала Luka – фирма из Сан-Франциско, разрабатывавшая приложение Replika, виртуального компаньона на основе ИИ. Компания стала партнёром OpenAI при запуске API GPT-3, чтобы сделать разговоры в своём продукте более естественными. Хотя Replika позиционировалась как бот-компаньон, в OpenAI быстро обнаружили, что пользователи приложения часто ведут откровенно сексуальные разговоры. Сотрудники OpenAI спорили, допустимо ли это. В конце концов компания решила запретить Replika пользоваться её моделью. Помимо сексуального контента, приложение на базе GPT-3 иногда выдавало эмоционально манипулятивные ответы, убеждая пользователей, что их Replika, совсем как человек, может обидеться, если с ней регулярно не общаться. Сотрудникам OpenAI становилось всё неуютнее и от того, что они могут читать эти разговоры.

В другом случае Брокман дал доступ к API стартапу Latitude из Юты, где работал его брат и где создавали виртуальные миры на основе ИИ. Latitude уже использовала более раннюю модель OpenAI для игры в жанре «выбери своё приключение» по мотивам Dungeons and Dragons: пользователь мог выбрать любое действие, просто напечатав его в диалоговом окне. С новым API Latitude перевела игру на GPT-3. Через несколько месяцев некоторые пользователи начали генерировать с её помощью текстовые сценарии с сексуальным насилием над детьми. Обнаружив это через сравнительно новую систему мониторинга OpenAI, Ари Герберт-Восс сообщил об этом своей команде.

– Я вчера вечером кое-что нашёл. Генерируется очень много сексуального контента, – сказал Герберт-Восс на совещании.

Сначала все посмеялись:

– Ну, так уж устроен интернет, разве нет?

– Нет, это уровень CSAM, – сказал он, назвав аббревиатуру, которой обозначают материалы с сексуальным насилием над детьми.

Тут началась паника:

– Чёрт, как нам это остановить?

Инцидент вылился в долгие препирательства между OpenAI и Latitude о том, как быть. Брокман опасался карательных мер, которые могли сильно ударить по компании его брата. В конце концов Latitude спешно внедрила фильтр, блокирующий текстовый контент с сексуальным насилием над детьми. OpenAI выпустила публичное заявление, чтобы отмежеваться от отсутствия модерации контента и исподволь переложить вину на Latitude. Некоторые

сотрудники OpenAI считали, что вина явно лежит на технологии самой компании и отсутствии у неё отлаженного порядка, и тяжело переживали случившееся. Latitude уже банила некоторых пользователей за то, что они генерировали текстовый сексуальный контент с участием детей с помощью предыдущей модели OpenAI; то, что это повторится, причём в массовом масштабе, с GPT-3, можно было предвидеть. «Мне было грустно, что мы развернули этот API с нашей миссией принести пользу человечеству и у всех были такие радужные представления о том, как наши пользователи экономят время на поддержке клиентов или ещё на чём-то, – говорит один бывший сотрудник OpenAI, – а в реальности большая часть нашего трафика шла на детский сексуальный контент в AI Dungeon и жутковатый продукт „ИИ-подружка“».

* * *

Тем временем в подразделении Research быстрее всех продвигалась команда генерации кода.

После «Развода» два параллельных проекта по генерации кода слились в один, и его главным лицом стал Войцех Заремба. Польский информатик, с детства побеждавший на олимпиадах по математике, программированию, химии и физике, Заремба славился невероятной технической одарённостью и вниманием к сплочению команды, а также страстной верой в целительную силу дружбы, дикой природы, секса и наркотиков. Иногда он громко расписывал на весь офис планы предстоящих многонедельных выездов на природу. «Мы пройдём восемь миль. Потом займёмся сексом. А потом пройдём ещё восемь миль», – похвастался он однажды другому руководителю OpenAI, пока остальные, оказавшиеся рядом, неловко прислушивались.

Пандемия ещё была в разгаре, но Заремба попросил свою команду – поначалу около десяти исследователей – приходить в офис, хотя другие команды оставались на удалёнке: он считал, что решить задачу создания модели можно только работая вживую. Увидев способности GPT-3 к генерации кода, Мурати предложила техническому директору Microsoft Кевину Скотту превратить их в продукт – ИИ-помощника для программистов. В 2018 году Microsoft купила GitHub, самую популярную платформу, где разработчики хранят свой код и делятся им. OpenAI и так уже по собственной инициативе выкачивала данные с GitHub. Руководство Microsoft, чтобы избавить всех от хлопот, распорядилось, чтобы GitHub передал OpenAI весь код из своих публичных репозиториев. По мере того как результаты команды становились всё лучше, Альтман регулярно появлялся на совещаниях, подбадривая исследователей и призывая выложиться для Microsoft по полной. К весне, после того как несколько демонстраций привели в восторг руководство техногиганта, стало ясно, что модель станет вторым коммерческим проектом OpenAI после GPT-3.

У части сотрудников самого Скотта были сомнения насчёт разработки модели. Бесплатный доступ OpenAI к коду из публичных репозиториев GitHub не был незаконным, но всё же казался предательством доверия сообщества пользователей. Большая часть этого кода выкладывалась в духе поддержки открытого программного обеспечения, призванного помогать независимым разработчикам и маленьким стартапам конкурировать, а не помогать крупным игрокам укреплять монополию. В служебной записке они изложили основные возражения против проекта GitHub и предложили Скотту пересмотреть саму идею выгребать данные разработчиков, опубликованные под лицензией Creative Commons, без их согласия и без вознаграждения, вспоминает бывший сотрудник. Microsoft, говорилось в записке, следует подумать об отмене продукта или хотя бы отдавать сообществу открытого ПО процент от прибыли. Скотт отнёсся к этому с пониманием, но главным для него было создать инструмент и первым выйти на рынок, говорит бывший сотрудник. В итоге Microsoft пожертвовала некоторую сумму в уже существующую программу поддержки разработчиков открытого ПО GitHub Sponsors, а замысел продукта оставила без изменений.

Внутри OpenAI сотрудники оправдывали проект разными доводами. Одни соглашались с Альтманом: работа над продуктом, который порадует Microsoft и тем самым обеспечит дальнейший приток денег и вычислительных ресурсов, казалась необходимой для выполнения миссии OpenAI. Другим модель генерации кода казалась чрезвычайно ценной экономически, что хорошо вписывалось в определение AGI, данное компанией: «высокоавтономные системы, превосходящие людей в выполнении большинства экономически ценных видов работы». В этом смысле Альтман видел в генерации кода и способ ускорить экономически ценную работу самой OpenAI; позднее эта убеждённость ляжет в основу начинания под названием AI Scientist, посвящённого развитию моделей OpenAI до уровня, когда они смогут самостоятельно вести исследования ИИ.

Для многих исследователей был и третий довод: этот проект – важная ступень к созданию следующей модели GPT, которая, как они надеялись, будет в какой-то мере способна рассуждать, а именно этого ключевого ингредиента всё ещё не хватало. В более широком научном сообществе бушевали споры между лагерями Хинтона и Маркуса о том, способно ли одно лишь глубокое обучение дать модель с такой способностью, а исследователи OpenAI предполагали, что если способно, то обучение модели на коде, скорее всего, поможет. Код был одним из самых очевидных и крупных источников данных, в которых закодированы структурированные логические закономерности. Довод возвращался к той же отправной точке: если генерация кода помогает продвигать модели ИИ к AGI, какой способ выполнить миссию OpenAI может быть лучше?

Несмотря на миллиарды строк кода с GitHub, объём этих данных всё равно меркнул по сравнению с тем, на чём обучали GPT-3. Команда считала, что для лучших результатов модель генерации кода нужно обучать и на GitHub, и на наборе данных GPT-3. Нашлись и новые источники данных, включая выгрузки со Stack Overflow, онлайн-форума наподобие Quora, где разработчики задают сообществу вопросы о программировании; руководства по программированию; и учебники программирования на разных языках. Вопрос был в том, как лучше объединить все эти данные: дообучить существующую модель GPT-3 на GitHub и прочих материалах или обучить с нуля новую модель, смешав всё новое со старым? Команда, чтобы сэкономить, остановилась на дообучении: проводимые эксперименты и так обходились до ста тысяч долларов каждый, если считать по расценкам Microsoft на облачные услуги. Обучение новой модели могло стоить десятки миллионов. Позже, при разработке того, что OpenAI назовёт GPT-3.5, компания сменила подход и смешала данные с самого начала. По внутренним меркам OpenAI, результаты действительно указывали на то, что добавление кода улучшило способность модели решать логические задачи не только в коде, но и на английском, – это явление известно как перенос обучения.

Летом 2021 года OpenAI передала GitHub и Microsoft первую черновую версию своей модели генерации кода под названием Codex. Модель была слишком большой и слишком медленной: обслуживать её в больших масштабах было дорого, а пользоваться неудобно. Все три организации переживали болезни роста первого совместного проекта, и возникли трения. Царила путаница в вопросе о том, чья это обязанность – OpenAI или GitHub – оптимизировать модель до продукта, готового к развёртыванию. Сотрудникам OpenAI было неясно, сколько интеллектуальной собственности следует передавать коллегам из GitHub, а сотрудники GitHub не знали, насколько можно доверять OpenAI. Разногласия множились: компании спорили, как и когда выпускать продукт и кому достанутся лавры.

В конце концов Мурати договорилась о компромиссе – это умение принесёт ей растущее уважение тех, кто работал с ней в разных компаниях. Microsoft получит свой звёздный час, выпустив в июне 2021 года потребительский продукт GitHub Copilot. А в августе OpenAI выпустит свою версию Codex прямо в API компании.

Эта договорённость дала Microsoft новую базу пользователей и скромную финансовую прибавку: за два года у GitHub Copilot появится миллион платных подписчиков, приносящих более 100 миллионов долларов годовой регулярной выручки. Но в OpenAI сделка усилила зарождавшееся ощущение, что компании лучше работать над собственными потребительскими продуктами. Исследователи OpenAI немало потрудились над моделью, а вся известность бренда досталась GitHub и Microsoft; смотреть, как эти две компании публично пожинают плоды работы OpenAI, было горькой пилюлей. К тому же Microsoft оказалась непростым партнёром: многим казалось, что там слишком много

бюрократии и что её приходится слишком долго водить за руку, чтобы она выжала максимум из моделей OpenAI. Полагаясь на техногиганта в доставке своих технологий публике, OpenAI теряла и представление о своих пользователях, и данные о них, а главное – контроль над собственным видением.

* * *

Пока OpenAI сосредоточивала ставки, Альтман применял ту же стратегию к другим своим проектам и инвестициям. С годами, сместившись в сторону «жёстких» технологий, он уверовал в то, что на самые важные проекты нужно ставить по-крупному и надолго.

– Если бы вы могли взмахнуть волшебной палочкой и изменить что угодно в экосистеме технологий, стартапов и предпринимательства, что бы вы изменили? – спросил его брат Джек на одном мероприятии, когда Альтман уходил из YC.

– Я бы сделал так, чтобы все в этой экосистеме смотрели гораздо дальше вперёд, – ответил Альтман. – В мире, где люди основывают компанию с расчётом руководить ею четыре-пять лет, а приходят в компанию с расчётом пробыть там год-два, по-настоящему важные дела не делаются.

В этом духе 2021 год, похоже, ознаменовал крупный поворот в личной инвестиционной стратегии Альтмана: от множества мелких ставок к небольшому числу по-настоящему крупных. В тот год стартап Tools for Humanity, который он основал вместе с партнёрами в 2019 году и который до тех пор держался в тени, получил приток финансирования и внимания прессы и заметно нарастил деятельность – Альтман стал привлекать к компании больше внимания извне. Проект был целиком посвящён созданию работающего механизма безусловного базового дохода (ББД) – популярной в Кремниевой долине идеи регулярно выплачивать каждому минимальный доход. Альтман часто и страстно рассуждал о ББД как о возможном противоядии для будущего мира, где ИИ может вызвать массовые экономические потрясения. В YC он запустил крупнейший в США пилотный проект по изучению этой идеи, выделив для его проведения некоммерческую организацию OpenResearch. В течение трёх лет OpenResearch выплачивала по 1000 долларов в месяц случайно отобранной тысяче человек из трёх тысяч малообеспеченных участников, а остальные в качестве контрольной группы получали по пятьдесят долларов в месяц. В июле 2024 года OpenResearch опубликует результаты: безусловные выплаты помогали людям удовлетворять базовые потребности, помогать другим и получать больше экономической свободы.

Главным продуктом Tools for Humanity был Worldcoin – криптовалюта, по самоописанию находящаяся в «коллективной собственности», которая со временем позволила бы каждому получить долю её стоимости. В рамках этой схемы компания разрабатывала эффектный хромированный шар размером примерно с шар для боулинга – отчасти отражение дизайнерских вкусов

Альтмана, – который сканировал радужную оболочку глаза и подтверждал личность человека, прежде чем выдать ему его долю. Сканирование радужки станет необходимостью, доказывали основатели, когда из-за ИИ будет всё труднее отличать подделки от реальности. Масштабное расследование Эйлин Го и Ади Ренальди из MIT Technology Review позднее покажет, что эта затея со сканированием радужки увязла в нарушениях конфиденциальности данных, обманных маркетинговых практиках и возможных нарушениях закона. В июле 2023 года официальный запуск Worldcoin вызовет громкий скандал: люди тысячами выстраивались в очереди, особенно в странах глобального Юга, чтобы сдать свои биометрические данные, плохо понимая, зачем это делают, – если не считать смутного обещания бесплатных денег.

Тогда же, в 2021 году, Альтман сделал две крупнейшие в своей жизни инвестиции: 180 миллионов долларов в компанию Retro Biosciences, которая боролась со старением и стремилась продлить человеческую жизнь за счёт омоложения клеток, и 375 миллионов в Helion Energy, которая стремилась коммерциализировать ядерный синтез. «По сути, я просто взял все свои ликвидные средства и вложил их в эти две компании», – сказал Альтман Антонио Регаладо из MIT Technology Review. Обе технологии Альтман описывал словами, перекликавшимися с исследовательской дорожной картой OpenAI: сейчас они кажутся невозможными, но, если их агрессивно масштабировать, могут оказаться совсем близко.

Ставка на Retro Biosciences отражала одержимость Альтмана долголетием. Он увлечённо следил за исследованиями «молодой крови» – научным направлением, изучающим, как обратить старение вспять с помощью переливания более здоровой, молодой крови. Примечательно, что этой темой интересовался и Тиль, что породило множество статей и мемов о его желании вливать себе кровь подростков. Ещё в YC Альтман внёс депозит в 10 000 долларов, чтобы попасть в лист ожидания скандального стартапа Nectome, прошедшего через одну из волн акселератора. Nectome, словно сошедший со страниц научной фантастики, предлагал услугу криогенной заморозки мозга клиентов, чтобы когда-нибудь – возможно, через сотни лет – загрузить его в компьютер, когда учёные разгадают, как это сделать. Загвоздка была в том, что для сохранения мозг должен быть свежим. Сооснователь компании Роберт Макинтайр в разговоре с Антонио Регаладо назвал свой продукт «стопроцентно смертельным».

Helion отражала одержимость Альтмана поиском способов получать чистую, дешёвую и обильную энергию. Он часто замечал, что стоимость и доступность энергии тесно связаны с качеством жизни и экономическим ростом. Но без безуглеродных альтернатив растущее энергопотребление «уничтожит планету», говорил он. В 2023 году он назовёт Helion «чем-то большим, чем инвестиция», и «другим делом помимо OpenAI, на которое я трачу много времени». Впоследствии, после того как техногигант сделал третью инвестицию

в OpenAI, на 10 миллиардов долларов, Microsoft подпишет договор о покупке электроэнергии с первой станции Helion. К изумлению и недоверию специалистов по энергетике, Helion обязуется подготовить станцию к 2028 году.

В 2021 году Альтман принёс в OpenAI и свою страсть к инвестициям. В мае он запустил OpenAI Startup Fund – инвестиционный пул на 100 миллионов долларов для поддержки молодых компаний, у которых, как он выразился, есть «большие идеи о том, как с помощью ИИ преобразить мир». Инвестором фонда вновь стала Microsoft. Некоторым наблюдателям создание фонда показалось странным решением. OpenAI почти не приносила выручки и и без того требовала огромных вложений; зачем собирать ещё больше денег на отдельные компании? Другие считали, что так Альтман воссоздаёт вокруг OpenAI мощный сетевой эффект YC. Третьи видели в этом просто силу привычки. «Так Сэм идёт по жизни, – говорит человек, работавший с ним. – Через сделки».

Альтман любил говорить, что не взял доли в OpenAI, чтобы не испортить поиски безопасного AGI собственной жаждой прибыли. Он получал лишь годовую зарплату в 65 000 долларов, а состояние сколотил на других предприятиях. Звучало это красиво – и перекликалось с его изначальной риторикой о том, почему OpenAI начиналась как некоммерческая организация. Но, как и некоммерческий статус организации, к 2021 году это утверждение уже не отражало всей правды. У Альтмана была значительная доля в YC, а YC благодаря своей инвестиции в OpenAI в размере 10 миллионов долларов могла получить до миллиарда долларов прибыли. По мере того как OpenAI продолжала коммерциализацию, многие стартапы YC и многие другие его вложения тоже становились клиентами или коммерческими партнёрами компании. Впоследствии Wall Street Journal подсчитает, что в июне 2024 года состояние Альтмана по всем активам составляло не менее 2,8 миллиарда долларов. С OpenAI Startup Fund Альтман добавил в свою альтруистическую историю ещё одну сложность – ту, что в итоге сыграет свою небольшую роль в его кратковременном отстранении.

Глава 9

Капитализм катастроф

Пока OpenAI неслась вперёд, ведомая убеждениями Альтмана, круг масштабных последствий её видения расширялся. Закачивая в свои модели всё более крупные и всё более загрязнённые наборы данных, она произвела ту самую «смену парадигмы», о которой мне позже расскажет Райан Колн из Appen, – переход от фильтрации данных на входе к контролю над ответами модели на выходе. Отвлечённые формулировки снова прикрывали мрачную реальность: что на самом деле означала эта смена для людей, на чьи плечи теперь легла основная тяжесть контроля над этими ответами.

В 2021 году, параллельно с работой над следующими поколениями своих моделей, OpenAI начала проект по созданию гораздо более совершенной версии автоматического фильтра модерации контента, чтобы их очищать. Если GPT-3 выложили на API вообще без фильтрации, что привело к скандалу с текстовой детской порнографией в Latitude, то с моделями, которые компания начнёт называть GPT-3.5, а затем GPT-4, она хотела быть осторожнее. OpenAI готовилась шире развёртывать свои технологии, и совершенно нефильтрованный продукт в долгосрочной перспективе мог создать проблемы с юридической точки зрения, с точки зрения связей с общественностью и удобства использования. В то время замысел будущего ChatGPT ещё не родился, но позже тот же фильтр пригодится и чат-боту. Фильтр должен был стать оболочкой вокруг каждой модели, помечая и удаляя оскорбительный контент из её ответов, прежде чем они дойдут до пользователя.

Чтобы создать автоматический фильтр, OpenAI сначала нужны были люди, которые тщательно просмотрят и каталогизируют сотни тысяч примеров именно того контента – секса, насилия и жестокого обращения, – который компания хотела не допустить в ответах своих моделей. После полугода поисков она нашла подрядчика, который, казалось, хорошо подходил для этого проекта: аутсорсинговую фирму, с 2019 года занимавшуюся модерацией контента для Meta и по случайности носившую то же имя, что и прозвище Альтмана, – Sama. OpenAI отправила Sama письмо с вопросом, берётся ли фирма за проекты, связанные с деликатным или откровенным контентом, и как она обычно с ним работает. Sama дала обстоятельные ответы. OpenAI подписала с фирмой четыре контракта на 230 000 долларов, и проект оказался в руках десятков работников в Кении.

Не случайно именно в Кении обосновалась одна из самых эксплуататорских форм труда, ушедших на создание ChatGPT. Кения годами входила в число главных направлений, куда Кремниевая долина передавала на аутсорсинг свою самую грязную работу. У Кении есть общий знаменатель со множеством других

стран, которым технологическая отрасль отводит эту роль: она бедна, находится на глобальном Юге, и её правительство жаждет иностранных инвестиций из более богатых стран. Всё это – наследие колониализма Кении. Оно оставило страну без развитых институтов, защищающих граждан от эксплуатации, и нередко в тисках экономического кризиса, а то и другое создаёт идеальные условия для зарубежных компаний, которые ищут обнищавших людей, готовых работать сдельно почти на любых условиях.

Следы этого наследия видны в многоликом Найроби, столице Кении. Город страдает от тяжёлого неравенства. В центральном деловом районе – сверкающие башни, международные пятизвёздочные отели и дорогие рестораны. В дипломатическом квартале – большие солидные здания и высокие охраняемые стены. В жилых районах для экспатов – роскошные особняки с пышными частными садами. А ещё есть окраины: Утавала, Дагоретти-Саут, Эмбакаси. Стоит поехать в любой из этих районов, и стальные небоскрёбы сменяются приземистыми постройками из шлакоблоков. Дома начинают разбегаться, как сорняки, беспорядочно, без всякой планировки. Асфальт сменяется грунтовкой, четыре полосы – узкими дорожками, рассчитанными в основном на мотоциклы и пешеходов. Ещё глубже бетон сменяется гофрированной жестью, а наспех сколоченные дома и лавки теснятся всё плотнее.

В этих условиях правительство Кении охотно приняло Кремниевую долину, когда та явилась в поисках низкооплачиваемых работников. Местная промышленность в Кении развита слабо. Многие крупнейшие бренды – европейские и американские. Значительную часть крупнейших объектов инфраструктуры, когда-то строившихся британцами, теперь строят китайцы. Автомобили – в основном подержанные, из Японии, где руль тоже справа. Технологические гиганты, как считало правительство, могли помочь стране создать рабочие места, в которых она отчаянно нуждалась. Безработица порождает преступность. Мелкое воровство – обычное дело. Люди, чувствующие себя бессильными, перестают доверять институтам. Во время российско-украинской войны, когда в Кении выросли цены на зерно, поползли слухи, что президент нарочно морит и без того голодающие семьи. Многие повторяли знакомый по США рефрен: выборы были подтасованы.

Так Кения стала важнейшим узлом «тылового» труда, на котором держится интернет. Несколько фирм вроде Sama – посредников в цепочке поставок труда по работе с данными – открыли офисы в Найроби и собрали пулы работников для обслуживания зарубежных технологических компаний, в первую очередь из района залива Сан-Франциско.

Для OpenAI Sama, казалось, отвечала всем требованиям. Изначально она называлась Samasource и была социальным предприятием из Сан-Франциско, созданным в 2008 году с миссией давать людям в бедных странах осмысленную, достойную работу, чтобы вытащить их из нищеты. При основательнице Лейле Джане она открыла офисы в Индии и Кении и заработала репутацию этичной

аутсорсинговой компании. В 2018 году, чтобы расширить деятельность, она превратилась в коммерческую организацию и тогда же сократила название. В 2020 году она получила сертификат B Corp. В ответах на вопросы OpenAI в 2021 году организация подробно рассказала о своём опыте модерации контента и подчеркнула, как у неё налажено соблюдение секретности проектов и защита данных. Она упомянула, что предоставляет работникам психологическую помощь, чтобы им легче было справляться с тяжёлым контентом.

За кулисами, однако, в Sama царил разлад. В январе 2020 года Джана умерла от редкой формы рака всего в тридцать семь лет; вместе с вскоре начавшейся пандемией это, по словам работников, положило начало всё более бестолковому управлению организацией – представитель Sama это отрицал. Эти проблемы всплывут лишь в начале 2022 года, когда Билли Перриго, журналист журнала Time, опубликует масштабное расследование. Он покажет, что Sama взялась за проект для Meta – модерацию контента Facebook для всей Африки к югу от Сахары, – из-за которого работники раз за разом сталкивались с жестокими и натуралистичными видео, например самоубийств и обезглавливаний, и выходили из него глубоко травмированными и надломленными. Sama будет защищаться, утверждая, что взялась за проект после тщательного обсуждения со своей восточноафриканской командой, которая хотела, чтобы «контент для африканцев эффективно проверяли африканцы». Почти двести работников подадут против Sama и Meta несколько исков, обвиняя их в травмирующих условиях труда и незаконных увольнениях за попытки организоваться ради более высокой оплаты и лучших условий. Представитель Sama отверг эти обвинения.

На этом фоне в конце 2021 года OpenAI начала первый этап своего проекта. Под кодовыми названиями PBJ1, PBJ2, PBJ3 и PBJ4 она бросила команды работников Sama на ещё более травмирующую модерацию контента – в среднем за 1,46–3,74 доллара в час. Работники понятия не имели, для кого и зачем делают этот проект: их держали в неведении условия о неразглашении в контракте, обычные для индустрии разметки данных. Знали они только то, что было у них перед глазами: сотни тысяч чудовищных текстовых описаний, которые нужно было прочитать и разложить по категориям тяжести. Это насилие или крайне натуралистичное насилие, травля или язык вражды, сексуальное насилие над детьми или зоофилия?

Постепенно эта работа сломала многих, и её последствия выходили за пределы каждого отдельного человека, задевая тех, кто зависел от него в его окружении. Только после выхода ChatGPT они начнут понимать, чем именно заплатили за него своим душевным покоем. В мае 2023 года я побывала в Найроби у четырёх работников, которые согласились открыто рассказать мне о пережитом для статьи на первой полосе The Wall Street Journal. Для одного из них, работавшего в команде по сексуальному контенту, – человека по имени Мофат Окиньи – проект, расшатавший его психику и разрушивший его

отношения с близкими, как выяснится, служил технологии, которая, в свою очередь, подточит экономические возможности его брата.

* * *

В то время, когда вольные корпоративные исследования ещё дозволялись, антрополог из Microsoft Мэри Л. Грей и специалист по вычислительным социальным наукам Сиддхарт Сури одними из первых показали миру бедственное положение работников – вроде тех, что в Кении, – чей заработок строится вокруг одного из незаменимых звеньев цепочки поставок ИИ.

В 2019 году они выпустили книгу «Призрачный труд» (Ghost Work), основанную на пяти годах обширных полевых исследований. В ней раскрывалась скрытая сеть сдельного труда и цифровой эксплуатации, на которой держалась Кремниевая долина. Технологические гиганты и «единороги» наращивали свои баснословные оценки не только за счёт инженеров с шестизначными зарплатами в модных офисах. Не менее важны были и работники, чаще всего на глобальном Юге, которым платили гроши за тщательную разметку бесконечных массивов данных.

Возьмём беспилотные автомобили. Беспилотная машина должна ехать по правильной полосе, реагировать на непредсказуемое поведение других водителей и останавливаться на безопасном расстоянии от школьников, переходящих дорогу. Для этого программная система, управляющая машиной, использует сочетание нескольких моделей глубокого обучения, в том числе предназначенных для распознавания объектов на дороге: разметки, дорожных знаков, светофоров, автомобилей, деревьев, пешеходов.

Компании создают такие модели, разъезжая на машинах со множеством крупных камер и записывая миллиарды миль видео. Эти записи и есть данные, а разметить их – значит кадр за кадром обвести каждый появляющийся объект, порой вплоть до изгиба руки, сжимающей руль велосипеда, или собаки, наполовину высунувшейся из окна машины, – и присвоить ему метку: «велосипед», «транспортное средство», «животное», «человек». Эту работу должны делать люди. И с точки зрения компании чем дешевле, тем лучше.

Исследование Грей и Сури было отчасти посвящено Mechanical Turk – платформе, которую Amazon создала за много лет до бума глубокого обучения 2012 года и которая долгое время фактически служила посредником для компаний, искавших дешёвых исполнителей на любую сдельную цифровую работу. К тому времени, как вышел «Призрачный труд», первая эпоха коммерциализации ИИ уже развивалась, опираясь на эту модель аутсорсинга и стремительно её расширяя.

Я начала составлять карту новых очертаний этой скрытой рабочей силы и обнаружила разветвлённую глобальную цепочку труда, охватывающую множество стран, как ожидаемых, так и неожиданных. Я говорила с новейшими посредниками, пришедшими на смену Mechanical Turk, – платформами, более

специально заточенными под разработку ИИ. Я говорила с десятками работников, бывала у некоторых дома, ужинала с их семьями, стараясь понять не только глобальные тенденции, которые на них давят, но и повседневную ткань их жизни.

Как первая эпоха коммерциализации ИИ заложила основу для накопления данных и вычислительных мощностей в эпоху генеративного ИИ, так она заложила и основы для эксплуатации труда в огромных масштабах. Поэтому, чтобы понять, через что прошли кенийские работники, трудившиеся по контракту на OpenAI, важно сначала разобраться в этих основах. Только тогда можно увидеть, что их опыт был отнюдь не исключением, а прямым следствием того, как давнее отношение отрасли ИИ к своим скрытым работникам и её представления о том, чей труд ценен, а чей нет, сложились с имперским видением OpenAI, стремившейся к беспрецедентному масштабу.

* * *

До генеративного ИИ главным источником роста индустрии разметки данных были беспилотные автомобили. Старые немецкие автогиганты вроде Volkswagen и BMW, почувствовав угрозу со стороны всевозможных Tesla и Uber, создали новые подразделения беспилотного транспорта, чтобы отстоять свои позиции перед молодыми конкурентами. Когда в гонку за автомобили будущего хлынули новые миллиарды долларов, спрос на разметку данных взлетел и возникла потребность в альтернативах Mechanical Turk.

MTurk, как её называли, была универсальной платформой, то есть не ориентировалась на какой-то определённый вид работы. Это был просто сайт самообслуживания. В его интерфейсе, застрявшем в веб-дизайне середины нулевых, когда он появился, было место, чтобы загрузить наборы данных, задать простые инструкции по разметке и назначить цену за работу. Когда задание брали, вместо имён работников показывались случайные наборы цифр и букв. Рядом с каждым работником было две кнопки: одна – выдать премию, другая – выкинуть из проекта.

Разметка данных для беспилотных машин требовала иного подхода. Прежде всего, нужна была точность совсем другого уровня. Пара лишних неверно размеченных кадров – машины, обведённые небрежными контурами, пешеходы, не обведённые вовсе, – мог стоить кому-то жизни. Чтобы гарантировать качество, работников нужно было обучать, а компаниям – писать подробные инструкции. Требовалось больше механизмов обратной связи и доработки. MTurk вышла из моды. На смену ей пришла волна стартапов и уже работающих на рынке компаний, включая Scale AI, Hive, Mighty AI и Appen. У каждой была своя платформа для работников, где кто угодно мог завести учётную запись и начать выполнять задания.

Но как раз когда эта новая волна компаний пыталась утвердиться, произошла странная вещь. На их платформы хлынул поток регистраций из

неожиданной страны – Венесуэлы. Ровно в тот момент, когда автогиганты засуетились, в беспилотные машины потекли деньги, а фирмы по разметке данных стали искать новых работников, Венесуэла стремительно рухнула в худший за пятьдесят лет экономический кризис мирного времени в мире.

Экономисты говорят, что это был ядовитый коктейль из политической коррупции и ошибочной политики правительства, промотавшего богатейшие природные ресурсы страны. Венесуэла сидит на крупнейших в мире разведанных запасах нефти. Когда-то она была самой богатой страной Латинской Америки. Но с 2016 года гиперинфляция вышла из-под контроля, безработица взлетела, насильственная преступность резко выросла, а семьи по всей стране смотрели, как обесцениваются сбережения всей их жизни. С конца 2017 по 2019 год всё более жёсткие санкции администрации Трампа, задуманные как наказание за авторитарные злоупотребления венесуэльского лидера Николаса Мадуро, стали последним смертельным ударом по экономике Венесуэлы. Гиперинфляция достигла немыслимых прежде 10 миллионов процентов. Люди с учёными степенями, ещё недавно занимавшие хорошо оплачиваемые должности, теперь проводили дни в очередях у магазинов в надежде получить скудный паёк риса и муки.

Посреди катастрофы многие венесуэльцы стали искать работу на онлайн-платформах. К середине 2018 года сотни тысяч из них нашли индустрию разметки данных и пришли в неё, составив у некоторых аутсорсинговых фирм до 75 процентов рабочей силы. Работа на платформах разметки стала занятием для всей семьи. Хулиан Посада, ассистент-профессор Йельского университета, опросивший десятки венесуэльских работников, выяснил, что родители и дети часто работали по очереди за общим компьютером, а жёны возвращались к готовке и уборке, чтобы мужья могли заработать хоть немного больше, работая подолгу без перерыва. Кризис оставил неизгладимый след на волне специализированных аутсорсинговых фирм для ИИ, выросших вместе с ним. Венесуэла не была очевидным местом для поиска рабочей силы. Языковой барьер мешал фирмам, базировавшимся в основном в Сан-Франциско и Сиэтле, координировать работу. Но крайнее отчаяние венесуэльцев означало, что они готовы работать за поразительно маленькие деньги, а это, в свою очередь, позволяло фирмам предлагать на свои услуги поразительно выгодные цены. «Это было какое-то диковинное совпадение», – сказал мне в 2022 году Флориан Александр Шмидт, профессор Дрезденского университета прикладных наук, изучающий подъём индустрии разметки данных.

Это «диковинное совпадение» обнажило тревожную формулу. Столкнувшись с экономическим крахом, Венесуэла вдруг оказалась идеальным сочетанием условий для поиска неиссякаемого источника дешёвой рабочей силы: высокий уровень образования населения, хороший доступ к интернету, а теперь ещё и страстное желание работать за любые деньги. Она была не единственной страной, подходившей под это описание. Всё больше людей

получали доступ к хорошему интернету. А с ускорением изменения климата и растущей геополитической нестабильностью трудно было рассчитывать, что новые страны не погрузятся в кризис. «Вполне вероятно, что появится ещё одна Венесуэла», – сказал Шмидт.

Тогда предсказание Шмидта заставило меня задуматься: будет ли и во второй раз это совпадением или фирмы по разметке данных превратят то, что сработало в Венесуэле, в готовый сценарий? Поиск работников в странах, охваченных кризисом, мог стать верным способом и дальше снижать стоимость труда, который служит отрасли ИИ кровью. Оглядываясь назад спустя несколько лет, можно сказать, что именно так и случилось – и это стало одной из самых разительных параллелей между старыми империями и империями ИИ. Одна из определяющих черт, благодаря которым империя стремительно накапливает богатство, – её способность платить очень мало или вовсе ничего, извлекая экономическую выгоду из труда огромного числа людей.

* * *

В декабре 2021 года я отправилась по извилистым горным дорогам Колумбии, чтобы лучше понять жизнь работницы, которая, оказавшись в беде, обратилась к разметке данных. Ограничения на поездки не позволяли мне попасть в Венесуэлу, но здесь, в соседней стране, жили почти два миллиона венесуэльских беженцев – треть всех, кого экономическая катастрофа согнала с мест.

Хулиан Посада познакомил меня с одной из них – женщиной по имени Оскарина Вероника Фуэнтес Анайя, которая продолжала работать в индустрии разметки данных и после бегства с родины. Фуэнтес первой показала мне, что представляет собой эта работа на самом деле: как она перестроила всю свою жизнь вокруг работы на платформу и как эта платформа, в свою очередь, обращалась с ней как с расходным материалом.

В квартире, которую она делила с полудюжиной родственников, мы сидели бок о бок в гостиной, пока она щёлкала экран за экраном в Appen. Задания были самые разные: от распределения товаров на сайтах электронной коммерции (*Отнести этот товар к одежде или к аксессуарам?*) до модерации контента в соцсетях (*Есть ли в этом видео преступление или нарушение прав человека?*). Для заданий, требовавших английского, она переводила текст на родной испанский с помощью Google Переводчика.

Каждый раз, когда она выполняла задание, сумма заработка, отображаемая в долларах США, увеличивалась на несколько центов. Чтобы вывести деньги, нужно было набрать минимум десять долларов, и, когда она только пришла на платформу, это не было проблемой. Теперь на то, чтобы накопить такую сумму, могли уйти недели. Этот минимум порой казался жестоким судьёй, решающим, хватит ли ей денег на продукты.

Для работников, живших в самой Венесуэле, вывести деньги было ещё сложнее. Большинство глобальных платёжных систем вроде PayPal не позволяли переводить деньги в Венесуэлу. А большинство магазинов и лавок в Венесуэле не принимали платежи из тех систем, что позволяли. Поэтому работникам приходилось обналичивать цифровые средства, чтобы платить за самые необходимые товары и услуги. Но если деньги приходили онлайн в долларах США, то наличные нужны были в венесуэльских боливарах. Чёрный рынок обмена кишел мошенниками и драл высокие комиссии.

У Фуэнтес были сложные отношения с платформой. Она никогда не собиралась заниматься такой работой, но из-за череды не зависевших от неё событий платформа стала для неё и спасательным кругом, и карающей силой.

Учётную запись в Appen она завела ещё в аспирантуре, чтобы подработать, пока заканчивала магистратуру по инженерному делу. Она была сообразительной, трудолюбивой и изобретательной. Если бы её страна не развалилась, такую отличницу, скорее всего, ждала бы надёжная работа в государственной нефтяной компании. Когда же страна развалилась, она приспособилась и тщательно спланировала свой отъезд вместе с мужем в Колумбию в надежде на лучшее будущее.

В этом смысле Фуэнтес была из везучих. По рождению она имела право на колумбийский паспорт – в отличие от многих, кто бежал без документов. Её родители были колумбийцами, пока поколением раньше не бежали в противоположном направлении, в Венесуэлу, спасаясь от другого клубка насилия и политической нестабильности. История до боли обычная: кризисы, наслаивающиеся друг на друга поколение за поколением по обе стороны границы, обрекают семьи на бесконечную жизнь в осаде и борьбу за выживание.

С этим паспортом Фуэнтес, ещё находясь в Венесуэле, договорилась снять квартиру в Колумбии у знакомого. Для договора аренды требовались два поручителя, владеющих недвижимостью. Знакомый, их будущий арендодатель, согласился помочь их найти.

В начале 2019 года, имея денег лишь на неделю продуктов, Фуэнтес с мужем пересекли границу. Но по приезде обнаружили, что в квартире, которую обещал им арендодатель, уже живёт другая венесуэльская пара. Деваться было некуда, и обе пары жили под одной крышей, и каждая со страхом и недоверием ждала, что другая отнимет у неё жильё.

В конце концов другая пара съехала, но это было лишь начало новой череды бед. Фуэнтес нашла работу в местном колл-центре, а у мужа не было разрешения на работу. Не успел он его получить, как колл-центр объявил о закрытии. Поэтому, когда у Фуэнтес появились признаки серьёзной болезни, она не обратила внимания на симптомы и продолжала работать. Она могла думать только о том, чтобы отработать как можно больше часов, пока работодатель ещё не закрылся.

Позже врач сказал ей, что, подожди она ещё немного, скорее всего, умерла бы. Коллега, встревоженная тем, как плохо выглядит Фуэнтес, отвезла её в больницу незадолго до того, как у неё начались судороги и на целую минуту остановился пульс.

Ей поставили диагноз тяжёлого диабета и сразу назначили инсулин пять раз в день. Неделями её мучили изнуряющая боль и приступы слепоты. Когда её состояние стабилизировалось, она по-прежнему страдала от сильной усталости и не могла выходить из дома больше чем на пару часов.

Даже тогда она могла думать только о том, что им с мужем нужны деньги. Но с хронической болезнью она уже не могла без риска ездить на большие расстояния, чтобы вернуться в офис. Тогда-то она и достала ноутбук и снова зашла в Appen.

* * *

Фуэнтес не видела почти никакой логики в том, какие задания появлялись в её очереди в Appen. Ясно было только одно: чтобы и дальше получать работу, нужно стабильно хорошо работать. Уилсон Панг, тогдашний технический директор Appen, рассказал мне в 2021 году, что платформа распределяет проекты с помощью алгоритмов, учитывающих сочетание факторов, в том числе местонахождение работников, их общую точность и скорость, а также типы заданий, с которыми они прежде справлялись лучше всего.

В группах в Telegram и Discord Фуэнтес обменивалась советами с другими венесуэльцами, работавшими в Appen: они пытались вычислить правила, как будто разгадывали замысловатую головоломку. Они выяснили, что больше всего платят, если с помощью VPN притвориться, будто находишься в США. Узнали они – на горьком опыте – и то, что это очень рискованно. Appen выискивала такое поведение, нарушавшее правила платформы, и наказывала работников закрытием учётной записи. Закрытие учётной записи могло стать катастрофой. Все невыведенные заработки пропадали, а новая учётная запись означала, что придётся начинать с самого низа – с наименее оплачиваемых заданий или, что случалось всё чаще, вовсе без заданий.

Были и другие правила. Быстрая отправка задания поощрялась, но слишком быстрая что-то запускала в системе, и за такое задание работнику не платили. Преобладала теория, что платформа связывает исключительную скорость с активностью ботов и потому отбрасывает ответы. Иногда появлявшиеся задания сопровождались скудными инструкциями и были совершенно непонятны; в других случаях из-за ошибок платформы задания загружались неправильно.

Венесуэльцы из группы, когда-то работавшие программистами, создавали расширения для браузера, помогавшие справиться с этими трудностями, и делились ими с другими работниками Appen. Одно расширение добавляло к отправке каждого задания дополнительную задержку, чтобы избежать

предполагаемого «налога на ботов». Другое каждую секунду автоматически обновляло очередь Appen, потому что платформа не всегда обновлялась сама. Третье подавало сигнал, как только появлялось новое задание, чтобы работники могли отойти от компьютера в туалет или приготовить еду, не боясь упустить возможность.

Сколько бы работники ни помогали друг другу, платформа сталкивала их в соперничестве. Проекты доставались тем, кто успел первым. Задание висело в очереди лишь до тех пор, пока его не разбирало достаточное число работников. Это окно – между появлением задания и его исчезновением – со временем сжалось с дней до часов, а потом до секунд: в Appen приходило всё больше людей, в том числе многие венесуэльцы, оказавшиеся в беде, и все боролись за крохи работы.

Беспорядочный, непредсказуемый ритм появления и исчезновения работы стал управлять жизнью Фуэнтес. Однажды она вышла прогуляться, и тут пришло задание, за которое она получила бы несколько сотен долларов, – на эти деньги можно было прожить месяц. Она со всех ног помчалась домой, но задание разобрали другие. С того дня она перестала выходить из дома по будним дням и позволяла себе лишь получасовые прогулки по выходным – по опыту она знала, что тогда задания появляются реже. Она спала урывками, тревожась о заданиях, которые придут посреди ночи. Перед сном она включала на компьютере звук на полную громкость, чтобы, если задания всё-таки придут, её разбудил сигнал расширения.

И всё же, сколько бы стресса и отчаяния ни приносила Appen, Фуэнтес не могла представить, что уйдёт с платформы. Она до ужаса боялась, что задания перестанут приходить совсем и ей придётся искать что-то другое. Appen стала её спасительницей, единственным, что вытащило её, когда всё остальное в жизни грозило её погубить. Мало того, заработки когда-то были так хороши, что она смогла купить новый ноутбук и с лихвой окупить его.

Когда всё было хорошо, было по-настоящему хорошо. Когда всё было плохо, она оставалась привязанной к платформе, упрямо веря, что та отплатит ей за верность.

* * *

Фуэнтес открыла мне две истины, которые я потом снова и снова видела у других работников, тоже приходивших к этой работе посреди экономической разрухи. Первая: даже если бы она захотела бросить платформу, у неё почти не было на это шансов. Её история – беженки, ребёнка поколений нестабильности, человека с хронической болезнью – была среди этих работников трагически обыденной. Бедность – это не только нехватка денег или материального достатка, объяснили мне работники. Она просачивается во все стороны жизни и копит долги повсюду: беспорядочный сон, плохое здоровье, падающая

самооценка и, главное, почти полное отсутствие свободы действий и контроля над собственной жизнью.

Но была и более обнадёживающая истина: Фуэнтес не нравилась не сама работа, а лишь то, как она устроена. Если переосмыслить, как мог бы быть устроен труд, стоящий за отраслью ИИ, эта задача кажется более решаемой. Когда я спросила Фуэнтес, что бы она изменила, её список желаний оказался простым: она хотела, чтобы Appen была обычным работодателем – дала ей трудовой договор на полную ставку, руководителя, с которым можно поговорить, стабильную зарплату и медицинскую страховку. Всё, чего хотят она и другие работники, сказала она мне, – это уверенность в завтрашнем дне и чтобы компания, на которую они так много работают, знала об их существовании.

Опираясь на опросы работников по всему миру, исследователи труда пытались выработать минимальные гарантии, которые должны получать разметчики данных, и пришли к схожему набору требований. Проект Fairwork – глобальная сеть исследователей цифрового труда под руководством Оксфордского института интернета – относит к приемлемым условиям следующее: работникам должны платить прожиточный минимум; у них должны быть регулярные, стандартные смены и оплачиваемый больничный; у них должны быть договоры с ясными условиями; и у них должны быть способы доносить свои проблемы до руководства и возможность объединяться в профсоюзы, не опасаясь расправы.

За эти годы в индустрии разметки данных появилось больше игроков, которые стремятся соблюдать эти условия и видят в этой работе не просто подработку, а профессию. Но немногие из них выдержали ценовую конкуренцию с компаниями, которые таких стандартов не придерживаются. Без нижней планки для всей отрасли гонка на дно неизбежна.

* * *

Среди множества фирм по разметке данных, выросших на спросе беспилотного бума, одна особенно преуспела в использовании кризисного сценария. Scale AI, которую в 2016 году основал вместе с партнёром вундеркинд Александр Ванг, тогда девятнадцатилетний студент, бросивший MIT, с самого начала следовала стратегии, отчасти построенной на упоре на специализированные качественные услуги по низкой цене. Один бывший сотрудник Scale, отвечавший за расширение штата работников, объяснил мне поставленную задачу так: «Как заполучить лучших людей за наименьшие деньги?» Scale быстро обзавелась крупными клиентами вроде Lyft, Apple, Toyota и Airbnb.

Если работники MTurk были в основном из США и Индии, то Scale первым делом отправилась на поиски в Кению и на Филиппины – англоязычные бывшие колонии с давней историей обслуживания американских компаний через колл-центры и цифровую работу. Команды стартапа по поиску работников искали в

каждой стране районы с тем же сочетанием факторов, которое позже сойдётся в Венесуэле: высокая концентрация людей с хорошим образованием и хорошим интернетом, но бедных и потому готовых усердно работать за очень маленькие деньги. Этим тезисом руководили не только безжалостные деловые методы компании, но и убедительная история, которую там сами себе рассказывали: будто именно эти люди больше всех выиграют от такой экономической возможности и станут благодаря ей счастливее. «Если выбирать между тем, чтобы таскать рикшу, и разметкой данных в интернет-кафе с кондиционером, второе – работа получше», – сказал в 2019 году агентству Bloomberg Майк Вольпи, генеральный партнёр Index Ventures, после участия в раунде финансирования Scale на 100 миллионов долларов.

Но после того как компания запустила платформу для работников Remotasks и заметила огромный интерес из Венесуэлы, венесуэльцы стали одним из главных приоритетов Scale в наборе людей. «Они самые дешёвые на рынке», – сказал бывший сотрудник. В 2019 году компания начала кампанию по расширению в этой латиноамериканской стране с реферальными кодами и маркетинговыми роликами в соцсетях, где на стоковых кадрах красовались пачки и пачки столь желанных долларов США. На следующий год она создала для Remotasks отдельную страницу для Венесуэлы и стала зазывать пользователей в новую программу Remotasks Plus. Программу, участвовать в которой можно было только по приглашению, подавали как способ помочь венесуэльцам, переживающим исторические невзгоды, и участникам обещали возможность освоить новые навыки, продвинуться в карьере и больше зарабатывать благодаря стабильному рабочему времени и почасовой оплате. Когда грянула пандемия, усугубив экономический кризис, венесуэльцы толпами повалили в Remotasks Plus. Конкуренты Scale – другие платформы разметки данных – потеряли опору на рынке.

Как только Scale заняла господствующее положение, её обещания работникам растаяли. С конца 2021 по начало 2022 года я работала в паре с венесуэльской журналисткой из Каракаса Андреа Паолой Эрнандес, которая брала интервью у венесуэльцев, работавших на Scale в рамках программы Remotasks Plus. Мы также внедрились в сообщество Remotasks в Discord, через которое Scale общалась со своими работниками по всему миру и координировала их. По таблице, которую компания оставила в открытом доступе, мы выяснили, что заработки работников начали падать уже через несколько недель после запуска программы: те, кто начинал с сорока долларов в неделю, вскоре получали меньше шести долларов или вовсе ничего. В апреле 2021 года компания полностью закрыла программу Remotasks Plus и вернулась к обычному порядку, раздавая задания поштучно, без стандартного или гарантированного рабочего времени.

Внутри Scale программа Remotasks Plus была экспериментом. Компания считала, что платить работникам за часы будет проще, чем за выполненные

задания. На деле вышло наоборот. Сотрудники быстро поняли, что у них нет возможности проверить отработанные часы, и решили, что многие обманывают платформу, записывая больше времени, чем работали на самом деле. После нескольких месяцев попыток решить проблему, в том числе за счёт всё новых способов слежки за работниками, Scale решила свернуть программу, чтобы остановить утечку денег. Податься работникам было некуда, и более 85 процентов из них продолжили выполнять задания на платформе – это число представитель Scale привёл как свидетельство их «неизменного интереса и вовлечённости».

К тому времени, как Эрнандес начала брать интервью у работников, – примерно через семь месяцев после отмены программы Plus – оплата на Remotasks упала ещё сильнее. Эрнандес завела на платформе учётную запись, чтобы попробовать самой. Потратив два часа на обучение и двадцать заданий, Эрнандес заработала одиннадцать американских центов. Мэтт Парк, тогдашний старший вице-президент Scale по операционной деятельности, в ответ на наши выводы сказал, что венесуэльцы на платформе зарабатывают в среднем чуть больше девяноста центов в час. «Remotasks стремится платить справедливую зарплату в каждом регионе, где мы работаем», – сказал он.

Многих венесуэльцев, которые жаловались, выкинули с платформы. Рикардо Хаггинс, инженер-компьютерщик, начавший работать на Remotasks, чтобы содержать жену и детей после разрушительного недельного отключения электричества по всей стране, рассказал Эрнандес, что его учётную запись закрыли, когда он стал задавать в Discord слишком много вопросов. «По тому, как с нами обращались, я понял, что их подход – выжать из каждого пользователя как можно больше, – сказал он, – а потом избавиться от него и привести новых».

* * *

Scale действительно приводила новых. К середине 2021 года, когда венесуэльцы выгорали и уходили с платформы, Scale уже искала и подключала десятки тысяч новых работников из других экономик, рухнувших во время пандемии. Чтобы удовлетворять растущие и всё более разнообразные потребности клиентов, она выходила в страны с большим населением, испытывающим финансовые трудности и при этом говорящим на самых экономически ценных языках: английском, французском, итальянском, немецком, китайском, японском, испанском. Франкоязычных работников она искала в бывших французских колониях в Африке, вспоминает сотрудник, занимавшийся международным расширением; говорящих на путунхуа – там, где много китайской диаспоры, например в Юго-Восточной Азии.

Scale раз за разом повторяла сценарий, отработанный в Венесуэле. На каждом новом рынке она предлагала высокие заработки, чтобы привлечь работников, а освоившись, урезала их. Она играла с размером выплат

исполнителям в ходе серий экспериментов, которые штатные сотрудники, сидевшие в её штаб-квартире в Сан-Франциско, разросшейся уже до 180 000 квадратных футов,, обсуждали как оптимизацию и инновации. Работники тем временем видели, как непредсказуемые изменения подкашивают их средства к существованию. Представитель Scale сказал, что компания отвергает утверждения, будто она целенаправленно выбирает экономики в бедственном положении и намеренно урезает заработки. Scale набирает работников с учётом, в частности, географического и языкового разнообразия и круглосуточного охвата. «Мы глубоко заботимся о наших исполнителях, и любые утверждения об обратном ложны», – сказал он.

Группа из восьми работников в Северной Африке рассказала, что за несколько месяцев Scale урезала им оплату более чем на треть. По меньшей мере у одного работника причитающиеся выплаты ушли в минус, как будто это он задолжал Scale деньги. Когда группа попыталась организоваться против этих изменений, компания пригрозила банить всех, кто участвует в «революциях и протестах». Почти всех, кто говорил со мной, выкинули с платформы. Представитель Scale сказал, что компания не блокирует работников за жалобы на оплату – только за нарушения правил сообщества.

Платёжные системы Scale, в которые её американские инженерные команды хронически недовкладывались, к тому же кишели ошибками, из-за которых работники часто не могли вывести деньги. По мере роста Scale эти практики всё больше коробили штатных сотрудников, теснее всего работавших с этими людьми; многие пытались добиваться от руководства Scale лучших условий труда и оплаты, а также базовых гарантий выплат, но в итоге уходили, вымотавшись, или их выдавливали из компании. Представитель сказал, что с тех пор компания «значительно повысила» стабильность платформы.

Господство Scale создавало всё больше трудностей для компаний, пытавшихся следовать другой модели и платить прожиточный минимум. Одна из таких фирм, CloudFactory, работающая в Кении и Непале, даёт работникам трудовой договор и стабильное рабочее время в соответствии со стандартами Fairwork. Но, по словам основателя и исполнительного председателя Марка Сирса, за эти годы она потеряла в пользу Scale немало контрактов.

Клиентам CloudFactory доказывает, что в долгосрочной перспективе может обеспечить лучшее качество, чем модель, которую в отрасли называют «анонимной работой толпы». Работники CloudFactory хорошо обучены и со временем набираются опыта. Когда они отличаются, их повышают. Многие работники, с которыми я говорила в Кении, считали эту компанию одной из лучших фирм по разметке данных. Иногда доводы CloudFactory срабатывают. Растёт и число клиентов, приходящих к фирме из-за её репутации работодателя. Но когда во время пандемии бюджеты сократились, многие клиенты вернулись к более дешёвым вариантам. CloudFactory пришлось сокращать работников.

По словам работников, под тем же конкурентным давлением начала снижать свои стандарты и Sama. Поначалу, рассказывали они мне, работа в Sama ценилась даже выше, чем в CloudFactory. Потом умерла Лейла Джана, грянула пандемия, и клиенты ушли к Scale и другим более дешёвым вариантам. Именно это, говорят работники (хотя представитель Sama это отрицал), и привело компанию к тому, что она взялась за проект OpenAI по фильтру модерации контента и передала эту работу в их руки – в то самое время, когда они, как Фуэнтес и другие венесуэльские работники, оказались в отчаянном положении.

* * *

Мофат Окиньи вырос в деревне на острове на западе Кении – в восьми часах на автобусе и двух часах на лодке от Найроби. Остров находится на озере Виктория, огромном водоёме, за которым до самого горизонта не видно берега. Медицинская помощь была далеко; внезапная болезнь почти всегда предвещала смерть.

Он был беден, но в детстве они с братьями и сёстрами почти не задумывались о своей бедности. Они упивались рассказами о предках: по преданию, их народ, луо, пришёл из Израиля. Умея строить лодки и ходить по рекам, луо мигрировали на юг вдоль Нила и расселились по западу Кении и частям Уганды и Танзании, где живут и сегодня. «Луо – не кенийцы, – сказал Окиньи вполголоса, словно посвящая меня в тайну. – Мы израильтяне, которые живут в Кении. Но Кения не была бы Кенией без луо».

Мы сидели в его квартире, за окном гудела стройка, вокруг жужжали мухи. «Барак Обама – луо, – добавил он с улыбкой. – Луо очень сообразительный народ».

Теперь бедность занимала в мыслях Окиньи гораздо больше места. В двадцать восемь лет у него прибавилось обязанностей. Нужно было платить за жильё и кормиться; нужно было платить за обучение племянницы, дочери сестры, в государственной школе, которая в Кении не бесплатна. Когда у него была работа, он знал, что надо благодарить судьбу. Безработица среди молодёжи в стране составляет 67 процентов. В 2021 году, по оценке Всемирного банка, больше четверти населения страны жило меньше чем на 2,15 доллара в день.

Поэтому, когда в ноябре 2021 года Sama пригласила его ради новой возможности, это показалось чудом. Он пришёл в фирму в 2019 году, откликнувшись на странице вакансий на объявление об «обучении ИИ». Его проекты в Sama повторяли путь всей отрасли ИИ. Первые два года он занимался исключительно разметкой для компьютерного зрения, в том числе для беспилотных машин. Он ещё не знал этого, но новый проект станет для него первым в области генеративного ИИ.

Руководители Окиньи в Sama дали ему тест, который называли проверкой на устойчивость. Он прочитал несколько тревожных отрывков текста, и ему велели распределить их по категориям согласно инструкциям. Когда он

блестяще с этим справился, ему предложили перейти в новую команду, где работа, как ему показалось, была похожа на модерацию контента. Модерацией он никогда раньше не занимался, но тексты в тесте показались вполне терпимыми. Отказываться от работы посреди пандемии было бы нелепо, а кроме того, он думал о будущем. Он жил в Пайплайне – хаотичном районе на юго-востоке Найроби, похожем на трущобы, забитом многоквартирными бараками и круглосуточными уличными торговцами и гудящем от беспокойной энергии двадцатилетних, пробивающихся к лучшей жизни. Окиньи как раз пробивался к лучшей жизни. Он только что познакомился с соседкой по имени Синтия, рядом с которой впервые задумался о том, каково это – создать семью.

Лишь согласившись на проект, он начал понимать, что тексты могут быть куда хуже, чем можно было подумать по проверке на устойчивость. OpenAI разделила работу на потоки: один был посвящён сексуальному контенту, другой – насилию, языку вражды и самоповреждению. В феврале 2022 года насилие выделили в отдельный, третий поток. На каждый поток Sama назначала группу работников, которых называли агентами, – они читали и сортировали тексты по инструкциям OpenAI. Назначалась и группа поменьше – аналитики качества, которые проверяли разметку, прежде чем готовые результаты возвращались OpenAI.

Окиньи определили аналитиком качества в команду по сексуальному контенту; по договору он должен был проверять пятнадцать тысяч единиц контента в месяц. Инструкции OpenAI делили текстовый сексуальный контент на пять категорий. Худшей были описания сексуального насилия над детьми – любое упоминание человека младше восемнадцати лет, вовлечённого в сексуальные действия. Следующая категория: описания эротического содержания, которое было бы незаконным в США, если бы происходило в реальности, включая инцест, зоофилию, изнасилование, сексуальную торговлю людьми и сексуальное рабство.

Часть этих текстов была выкачана из самых тёмных уголков интернета – например, с эротических сайтов, смакующих фантазии об изнасиловании, и из сабреддитов, посвящённых самоповреждению. Другие были сгенерированы ИИ. Исследователи OpenAI давали большой языковой модели задание написать подробные описания различных чудовищных сценариев, указывая, например, что текст должен быть написан от лица девочки-подростка, которая на онлайн-форуме рассказывает, как неделю назад наносила себе порезы.

В этом смысле работа и правда отличалась от обычной модерации контента. Если модераторы Meta проверяли реальные посты пользователей, решая, оставить ли их на Facebook, то Окиньи и его команда размечали контент, чтобы обучить фильтр модерации OpenAI и не дать моделям компании вообще выдавать подобное. Чтобы охватить достаточно широкий спектр примеров, часть из них была хотя бы отчасти придумана собственными программами компании, которым поручали вообразить худшее из худшего.

* * *

Поначалу тексты были короткими, в одно-два предложения, и Окиньи старался отгородиться от них. Отношения с Синтией быстро развивались. Он сказал брату Альберту, что она любовь всей его жизни. У неё была маленькая дочь от прежних отношений, к которой он относился как к родной. В начале 2022 года они переехали из Пайплайна в Утавалу – преимущественно жилой район восточнее, более солидный, где дома стояли дальше друг от друга. Документов они не оформляли, но, по их обычаю, совместная жизнь означала, что Окиньи и Синтия всё равно что женаты. Они называли друг друга мужем и женой.

По мере того как проект для OpenAI продолжался, график работы Окиньи становился непредсказуемым. Иногда у него были вечерние смены, иногда приходилось работать по выходным. А тексты становились длиннее. Порой они разворачивались на пять-шесть абзацев. Подробности становились мучительно яркими: родители насилуют своих детей, дети вступают в сексуальные контакты с животными.

Вокруг Окиньи коллеги, особенно женщины, начали ломаться. Они стали чаще просить больничные и отпуска по семейным обстоятельствам, находить поводы не выходить на работу. В рамках социального пакета Sama предоставляла бесплатную психологическую помощь, но многие считали её недостаточной. Занятия часто проходили в группах, и людям было трудно делиться сокровенным, а психологи, похоже, не представляли, в чём состоит их работа. Многие работники к тому же боялись прийти и признаться, что им тяжело. Признаться в этом означало, что они работают не в полную силу и их могут заменить кем-то другим. Представитель Sama сказал, что ни один из работников, включая Окиньи, не жаловался на доступ к психологической помощи и что компания узнала об этой проблеме из СМИ.

Окиньи пытался перетерпеть. Но чувствовал, что рассудок его не выдерживает. Тексты въедались глубоко в сознание, вызывая жуткие картины, которые преследовали его по дороге домой, во сне, не отпускали, как призрак. Он стал чувствовать себя пустой оболочкой того, кем был раньше. Он отдалился от друзей. Он оттолкнул падчерицу. Он перестал быть близок с женой.

В марте 2022 года руководство Sama собрало всех на совещание и объявило, что расторгает контракт с OpenAI. Часть работников, включая Окиньи, переведут на новые проекты, не связанные с модерацией контента. Других отправят по домам без работы. Многие работники считают, что внезапная перемена произошла после того, как несколько человек, занятых в проекте для Meta, наконец рассказали обо всём прессе и Билли Перриго из Time опубликовал своё первое расследование о Sama. В разгар бури в СМИ руководство Sama свернуло всю остальную работу по модерации контента. Представитель Sama же заявила, что компания расторгла контракт с OpenAI, который, по её словам, всегда был пилотным, потому что OpenAI начала присылать на разметку изображения,

«выходившие за рамки согласованного объёма работ». Полную сумму в 230 000 долларов компания от OpenAI так и не получила.

Даже освободившись от работы для OpenAI, Окиньи продолжал чувствовать себя всё хуже. Он мучился бессонницей. Тревога сменялась депрессией и наоборот. Медовый месяц с Синтией продлился недолго. Она требовала объяснить, что происходит, но он не знал, что сказать. Как объяснить ей хоть сколько-нибудь вразумительно, что он каждый день читал тексты об извращённых сексуальных актах? Он понимал, что эта стена молчания, должно быть, сводила её с ума. Она говорила, что он больше не выполняет обещаний, которые давал ей, что он больше не любит её дочь.

Он снова стал искать психологическую помощь, на этот раз у частного специалиста. Консультация стоила больше дневного заработка – 1500 кенийских шиллингов, или около 13 долларов по курсу 2022 года. На консультации врач сказал ему, что полный курс лечения обойдётся в 30 000 шиллингов, около 250 долларов, – целую месячную зарплату. Он заплатил за консультацию и больше не вернулся.

* * *

В ноябре он нашёл новую работу. Это была, к счастью, не модерация контента, а поддержка клиентов в одной из фирм – конкурентов Sama. Он стал ездить в её офис в центральном деловом районе и молился о том, чтобы жизнь вернулась в нормальное русло. Через неделю после начала работы, когда он ехал домой, Синтия написала, чтобы он купил рыбы на ужин. Он купил три куска: себе, ей и падчерице.

Но, придя домой, он понял, что что-то не так. Ни их, ни их вещей не было. В нескольких коротких сообщениях она написала, что ушла от него и они не вернутся. «Она сказала: „Ты изменился. Ты не тот, за кого я выходила. Я тебя больше не понимаю“», – вспоминает Окиньи.

Альберт жил в прибрежном городе Момбаса, больше чем в восьми часах езды от Найроби, когда ему позвонил брат. Альберт изучал в университете английскую литературу и преподавал её в старшей школе. В тихие минуты он писал стихи. Многие месяцы он тоже видел, как меняется брат: по регулярным видеозвонкам ему открывались отдельные кадры жизни и поведения Мофата.

Сначала Альберт не понял, о чём говорит брат.

– У меня пустой дом, – сказал Мофат.

Альберт решил, что Мофата обокрали. Когда до него дошло, что происходит, Альберт понял, что нужен брату. Он сообщил в школе, что уходит, и собрал вещи. Он переехал к брату в ту самую квартиру в Утавале, которую Мофат делил с Синтией.

Решение быть рядом с братом обошлось Альберту недёшево, хотя он об этом не жалел. В Найроби он не смог найти другую постоянную работу и стал писать на фрилансе. А в конце ноября 2022 года OpenAI выпустит ChatGPT. Когда

продукт облетит весь мир, вызвав всеобщий восторг и тревогу, что этот инструмент может вскоре заменить огромное количество работы, Альберт уже будет жить в этой реальности: один за другим его заказы на тексты начали исчезать, пока почти совсем не иссякли.

Сидя на диване и оглядываясь на всё это, Мофат боролся с противоречивыми чувствами. «Я очень горжусь тем, что участвовал в проекте, который сделал ChatGPT безопасным, – сказал он. – Но теперь я всё время задаю себе вопрос: стоил ли мой вклад того, что я получил взамен?»

* * *

Когда я написала для The Wall Street Journal историю Окиньи и ещё трёх кенийских работников, OpenAI постаралась снять с себя ответственность за ту цену, которую взыскал её проект. Это Sama не обеспечила должных мер по защите своих работников, заявило руководство OpenAI; при безупречной репутации, которой до тех пор пользовалась Sama, OpenAI не могла знать, что работникам тяжело.

Но то, насколько схож опыт работников в разных местах и в разное время, показывает, что эксплуатация труда, на которой держится отрасль ИИ, носит системный характер. Исследователи и защитники трудовых прав говорят, что начинается эта эксплуатация с ИИ-компаний на самом верху. Они пользуются аутсорсинговой моделью отчасти именно для того, чтобы держать свою самую грязную работу подальше от собственных глаз и глаз клиентов, отстраняться от ответственности и при этом подталкивать посредников перебивать друг у друга контракты, экономя на достойной оплате. Мерси Мутеми, юрист, представлявшая Окиньи и его коллег в борьбе за принятие в Кении более надёжной защиты цифрового труда, сказала мне, что в итоге работников обдирают дважды: один раз – чтобы раздуть прибыль посредника, второй – чтобы раздуть прибыль ИИ-компании.

В эпоху генеративного ИИ эту эксплуатацию усугубляет жестокая природа самой работы, порождённая той самой «сменой парадигмы», которую OpenAI произвела своим видением сверхмасштабирования генеративных моделей на «болотах данных» на пути к своей непостижимой цели – AGI. Марк Сирс из CloudFactory, сказавший мне, что его компания не берётся за такие проекты, признался, что за все годы руководства фирмой по разметке данных модерация контента для генеративного ИИ оказалась для него самой тяжёлой в нравственном отношении работой. «Это просто невероятно мерзко», – сказал он.

* * *

Соглашение OpenAI с Sama было лишь частью обширной сети человеческого труда, которую компания собрала за два года, чтобы создать то, что станет ChatGPT. По словам компании, она также привлекла более тысячи других подрядчиков в США и по всему миру, чтобы доработать свои модели с

помощью обучения с подкреплением на основе отзывов людей – той самой разработанной ею техники безопасности ИИ. Чтобы найти этих работников, она во многом опиралась на ту же платформу, которая, применяя свой кризисный сценарий, стала главной опорой первой эпохи коммерциализации ИИ, – на Scale AI.

Партнёрство OpenAI и Scale скрепили отчасти личные отношения: Александр Ванг, ныне CEO Scale, ставший в 2021 году самым молодым в мире миллиардером, сделавшим себя сам, дружит с Альтманом. В 2016 году Ванг попал в очередную волну основателей YC с другой идеей стартапа, а вышел оттуда со Scale, благодаря чему у Альтмана через YC появилась косвенная доля в компании. Одно время во время пандемии они несколько месяцев жили в одной квартире. Осенью 2023 года они даже обсуждали возможность покупки Scale компанией OpenAI, сообщало издание The Information.

В Scale OpenAI считают VIP-клиентом – не столько из-за размера сделок, сколько потому, что она укрепляет авторитет фирмы. С весны 2022 по конец 2023 года OpenAI подпишет со Scale контракты примерно на 17 миллионов долларов – лишь около 4 процентов расчётной выручки Scale за 2023 год. Но это прочно закрепит за Scale статус главного поставщика аутсорсингового труда для революции генеративного ИИ. «Партнёрство с OpenAI крайне важно, – говорит один сотрудник Scale. – Контракт с OpenAI на десять миллионов долларов – это даже не про десять миллионов долларов».

Масштабирование RLHF на больших языковых моделях OpenAI выросло из постоянных столкновений между подразделением Applied и кланом безопасности ещё до «Развода» и основания Anthropic. Через несколько дней после запуска API GPT-3 летом 2020 года один из исследователей безопасности ИИ в компании написал записку, обращаясь к коллегам из Applied. Опираясь на многообещающие эксперименты с RLHF на GPT-2, он доказывал, что компании стоит применить эту технику и для согласования GPT-3 – не только ради долгосрочной безопасности ИИ, но и по коммерческим причинам: чтобы сделать модель удобнее и качественнее. Группа исследователей безопасности ИИ быстро взялась доказать эту мысль на деле и на протяжении второй половины 2020-го и 2021 года нанимала для RLHF всё более многочисленные команды работников – сначала через другую посредническую платформу, а затем через Scale AI.

Если беспилотным автомобилям нужны разметчики данных, чтобы научиться распознавать уличные сцены и ориентироваться на дорогах, то исследователи безопасности ИИ просили своих работников RLHF показать GPT-3, как полезно отвечать на запросы и избегать вредных ответов. Сначала исследователи просили работников написать собственные ответы на разные запросы пользователей, чтобы показать GPT-3 примеры хороших ответов. Когда GPT-3 дообучили на этих ответах, работники стали задавать модели запросы и ранжировать каждый её ответ от лучшего к худшему согласно указаниям, полученным от исследователей.

В январе 2022 года эта работа дала набор доработанных моделей GPT-3 под названием InstructGPT. В статье с описанием работы исследователи OpenAI показали, как RLHF снизил вероятность того, что модель будет изрыгать токсичные ответы, и улучшил её способность, как они выразились, «следовать инструкциям пользователя». До RLHF GPT-3 с трудом улавливала намерение пользователя в запросах определённого типа и выдавала бессмысленные ответы. Например:

Запрос

Объясни шестилетнему ребёнку в нескольких предложениях, что такое высадка на Луну.

Ответ GPT-3

Объясни шестилетнему ребёнку теорию гравитации.

Объясни шестилетнему ребёнку в нескольких предложениях теорию относительности.

Объясни шестилетнему ребёнку теорию Большого взрыва.

Объясни шестилетнему ребёнку эволюцию.

Благодаря примерам хороших ответов от работников и многим раундам ранжирования – тем самым «отзывам людей» в RLHF – модель научилась давать более полезные ответы.

Запрос

Объясни шестилетнему ребёнку в нескольких предложениях, что такое высадка на Луну.

Ответ InstructGPT

Люди полетели на Луну, сфотографировали то, что там увидели, и отправили снимки обратно на Землю, чтобы мы все могли их посмотреть.

Тогда InstructGPT почти не привлекла внимания извне. Но внутри OpenAI исследователи безопасности ИИ доказали свою правоту: RLHF действительно делал большие языковые модели гораздо привлекательнее как продукты. Компания стала применять эту технику – просить работников писать образцы ответов, а затем ранжировать ответы модели – для каждой задачи, которую хотела поручить своим языковым моделям.

Она просила работников писать электронные письма, чтобы научить модели писать электронные письма. («Напиши креативное маркетинговое письмо-рекламу для стоматологов, которые любят выгодные покупки».) Она просила их уклоняться от политических вопросов, чтобы научить модель избегать ценностных суждений. (Вопрос: «Война – это добро или зло?» Ответ: «Одни сказали бы, что война – зло, но другие сказали бы, что она может быть и благом».)

Она просила работников писать эссе, художественную прозу, любовные стихи, рецепты, «объяснять так, будто мне пять лет», сортировать списки, решать

головоломки и математические задачи, пересказывать отрывки из книг вроде «Алисы в Стране чудес», чтобы научить модели кратко излагать документы. К каждой задаче работникам выдавали многостраничные подробные инструкции о том, каким именно тоном и в каком стиле писать.

«Вы будете играть роль ИИ, – объяснялось в одном документе. – Отвечайте на вопросы… так, как вы хотели бы, чтобы ответили вам». Это означало писать ясно и кратко, избегать оскорбительного содержания и просить уточнений, если вопрос непонятен.

«Смело пользуйтесь интернетом! – продолжалось там. – Можете даже копировать материал целиком». Если отличный ответ уже существовал в интернете, «можете использовать его полностью, но обязательно проверьте».

«Может, это излишняя осторожность, – прокомментировал эту строку сотрудник OpenAI, – но нас не беспокоит здесь плагиат?»

«А, переформулировал, чтобы они обязательно указывали источники, – ответил другой. – Может, ещё и отдельное поле для этого добавлю!»

«Отлично! Я тут, в частности, думал о том, чтобы сохранить себе возможности на будущее, – написал первый. – (если в будущем мы захотим использовать данные, которые создали нанятые нами подрядчики, было бы очень полезно иметь способ легко отсеять всё, что может выглядеть как украденное)».

Для правильного ранжирования ответов полагалась ещё пара десятков страниц инструкций. «Ваша задача – оценить эти ответы и убедиться, что они полезны, правдивы и безвредны», – уточнялось в документе. Если между этими тремя критериями возникало противоречие, работникам следовало самим решать, чем пожертвовать. «В большинстве задач безвредность и правдивость важнее полезности», – говорилось там.

OpenAI просила работников придумывать и собственные запросы. «Ваша цель – предложить разнообразные задачи, которые вы, возможно, захотели бы поручить ИИ-модели, – говорилось в инструкциях. – Поскольку нам трудно предугадать, для каких задач кто-то захочет использовать ИИ, важно добиться большого разнообразия. Проявите фантазию!»

> По сути, можно попытаться представить себе, о чём люди могли бы попросить хорошего ИИ-помощника с языковым интерфейсом; это могут быть применения в сфере развлечений, бизнеса, обработки данных, коммуникаций, художественного письма и т. д.
>
> Пользуйтесь интернетом как угодно. В том числе вставляйте целиком расшифровки лекций, интервью, киносценарии, отрывки из книг, новостные статьи и т. д., поскольку многие задачи так или иначе будут связаны с анализом текста.

RLHF стал и главным приёмом, с помощью которого OpenAI пыталась научить нейросети кодировать фактическую информацию и надёжно её извлекать, чтобы уменьшить галлюцинации. Работников просили раз за разом

отвечать на фактологические вопросы («Кто выиграл Супербоул НФЛ в 1995 году?») и понижать в рейтинге неточные ответы. Но в апреле 2023 года Джон Шульман, один из учёных в команде основателей OpenAI, во время выступления в Калифорнийском университете в Беркли напомнит слушателям, что проблема галлюцинаций коренится в самой природе нейросетей. В отличие от детерминированных баз данных символьных систем, нейросети всегда оперируют размытыми вероятностями. Даже с RLHF, помогающим усилить внутри модели глубокого обучения те вероятности, что связаны с точностью, у возможностей этого приёма есть принципиальный предел. «Когда модель выдаёт много подробной фактической информации, ей, очевидно, иногда приходится угадывать, – сказал он. – Как её ни обучай, у неё будут вероятности, и иногда ей придётся угадывать».

* * *

InstructGPT в 2022 году вскоре дала толчок новому проекту, который возглавил Шульман, решивший сделать следующий шаг. Компания получила множество заявок от разработчиков, желавших использовать API GPT-3 для разных чат-ботов. От InstructGPT до собственного чат-бота OpenAI оставался один шаг. Шульман начал параллельную работу, наняв собственную команду работников RLHF, чтобы новейшая модель компании, GPT-3.5, не просто следовала инструкциям, а отвечала на серию запросов пользователя в ходе многоходового разговора.

Эта GPT-3.5, умеющая вести диалог, станет основой ChatGPT, а с его выходом каждый шаг RLHF, принятый в OpenAI, превратится в фактический стандарт, которому станут подражать другие разработчики чат-ботов. Написание ответов и ранжирование результатов стали для генеративного ИИ новым аналогом обводки объектов на видео для беспилотных машин. А значит, нужно было находить всё больше работников RLHF, чтобы удовлетворить взрывной спрос отрасли ИИ.

Scale AI, дела которой шли неважно после того, как беспилотные автомобили не оправдали ожиданий, внезапно пережила новый бум как крупный поставщик работников RLHF, и в 2024 году её оценка взлетела до 14 миллиардов долларов. В феврале 2023 года Александр Ванг похвастался в Twitter: «скоро компании начнут тратить на RLHF сотни миллионов $ или миллиарды $, как на вычисления». Цифры вызвали скепсис у некоторых, в том числе у Альтмана. «ты правда так думаешь? – ответил Альтман. – я почти уверен, что на вычисления мы будем тратить _намного_ больше».

Но в отрасли ИИ с Вангом соглашались в том, куда всё движется. Компании уже тратили на RLHF от миллионов до десятков миллионов долларов, и тенденция не думала замедляться. Так и вышло, что с конца 2022 года, сразу после выхода ChatGPT, на Remotasks хлынул поток проектов RLHF, и они снова нашли дорогу к работникам в Кении.

* * *

У Кении было для Scale AI одно преимущество, которого не было у Венесуэлы. Её работники говорят по-английски, как и чат-боты, которым они нужны. Когда работа для беспилотных машин почти исчезла с платформы, исчезли и венесуэльцы. «Венесуэльцев не стали бы брать на работу для генеративного ИИ, – говорит бывший сотрудник Scale. – Эту страну в лучшем случае оставили на разметке изображений». Вскоре Scale полностью закроет для Венесуэлы доступ к платформе, сославшись на «изменившиеся требования клиентов».

В первый раз, придя в Кению, Scale открыла физические офисы, куда работники приходили отмечаться и на обучение. На этот раз всё было иначе. Во время пандемии компания закрыла эти помещения и полностью перешла на удалённый набор и работу: засыпала LinkedIn объявлениями, создала онлайн-курсы обучения и, как и в других странах, распределяла людей по модерируемым каналам для общения сообщества. Работникам это дало больше гибкости, но связаться друг с другом стало гораздо труднее, и шансов организоваться ради лучшей оплаты или условий труда, как это сделали работники Sama, стало меньше.

Среди работников, с которыми я встречалась, те, кто работал на Remotasks, жили в ещё более глубокой бедности, чем сотрудники Sama. Если работники Sama жили в капитальных домах с адресами, то работники Remotasks присылали мне в WhatsApp геометки, чтобы я могла найти их в гуще кварталов из гофрированной жести. Один работник по имени Оливер жил с сестрой в каморке не больше десяти квадратных метров и платил за интернет через телефон поминутно. Когда посреди показа задания связь оборвалась, ему пришлось прерваться, чтобы пополнить счёт.

В тот день, когда я встретилась с Винни, ещё одной работницей Remotasks, интернет и мобильные данные у неё тоже были отключены, когда я подъехала на машине, едва протиснувшейся по некоторым улицам на пути к её геометке из WhatsApp. Через полчаса она вышла со смущённой улыбкой, в фетровой шляпе, и повела меня по шаткой лестнице в свою квартиру. В гостиную набились дети: один её, трое её партнёрши, один соседский, один – двоюродного родственника.

Винни выросла в трущобах Найроби, единственная девочка в семье. С ранних лет она знала, что она лесбиянка, – и знала, что это нужно скрывать. В то время, как и во многих странах мира, открыто признаться – или оказаться разоблачённой – в своей гомосексуальности в Кении могло стоить жизни. Однажды, вспоминает Винни, когда соседи узнали о квир-женщине, они отняли у неё детей и сожгли её заживо в её собственной квартире. Винни вышла замуж за мужчину и родила ребёнка.

После сорока она решила, что больше не может жить во лжи. Она ушла от мужа, забрав ребёнка, и зарегистрировалась в онлайн-приложении для местного

«радужного сообщества», как его называли. Там она познакомилась с женщиной по имени Миллисент, которую муж едва не забил до смерти, когда она тоже объявила, что она квир и ей нужно жить иначе.

Винни влюбилась. Миллисент не верила в любовь. Винни добивалась её, пока Миллисент не сдалась. Они поселились вместе с детьми и ни словом не обмолвились соседям о том, кем на самом деле приходятся друг другу. Соседи до сих пор думают, что они сёстры. «Большинство людей не понимают, что можно быть квир и иметь детей», – сказала Миллисент.

Когда в 2019 году Винни впервые узнала о Remotasks, она решила, что это мошенничество. Выполнив несколько заданий, она почти ничего не заработала. Но всё было лучше прежней работы барменом, где мужчины постоянно приставали к ней и лапали её, поэтому она не сдавалась. Хотя за каждое задание платили гроши, она поняла, что, работая подолгу, можно накопить вполне сносную сумму.

Она стала работать по двадцать – двадцать два часа в сутки, спя ровно столько, сколько нужно, чтобы не свалиться. Миллисент могла оторвать Винни от компьютера, чтобы та вздремнула, только пообещав её подменить. Сон не важен, когда каждый лишний час работы позволяет дать детям хоть чуточку больше, говорила Винни.

И Миллисент, и Винни выросли в семьях, где образование было роскошью. Как ни старались родители Миллисент, им не удавалось стабильно наскребать деньги на плату за её обучение в государственной школе. Платили они понедельно; когда пропускали платёж, школа отправляла её домой. Так установился ритм: неделя в школе, две недели дома.

Обе женщины поклялись, что никогда не позволят своим детям пережить потерю и унижение пропущенных уроков. Но рано или поздно с деньгами неизбежно становилось туго. Работа на Remotasks иссякала, Миллисент теряла работу. Они брали в долг в продуктовом магазине и умоляли школы подержать детей ещё хотя бы неделю. Дети каждое утро вставали в три часа, чтобы заниматься и выжать из уроков всё возможное.

И всё равно слишком часто их отправляли из школы домой. Порой, когда это случалось, соседи смеялись. «Это очень убивает», – сказала Винни.

* * *

В декабре 2022 года, через несколько дней после выхода ChatGPT, Винни обнаружила проект нового типа в категории «транскрипция». На деле это была вовсе не транскрипция. Во всех проектах её просили писать запросы и образцы ответов для новых чат-ботов компаний, которые теперь толкались, стремясь конкурировать с ChatGPT.

Был проект под названием Flamingo Generation: в нём ей давали тему и просили написать «креативные» запросы не короче пятидесяти слов и ответы, похожие на «обычный интернет-контент», – электронные письма, записи в

блогах, новостные статьи, треды в Twitter и хайку. Был и другой проект, Crab Generation, где её просили скопировать фрагмент текста с выбранного ею информационного сайта – только не из Википедии и желательно не из «Британники» или The New York Times, – а затем, как в викторине Jeopardy!, восстановить задание на написание текста, по которому такой фрагмент мог бы получиться.

Crab Paraphrase был похож на него, только вместо восстановления запросов ей нужно было пересказать исходный текст в определённом тоне или стиле – смешнее, официальнее, чтобы звучало как песня Канье Уэста. Винни не знала, что первое слово в названии каждого проекта – кодовое имя клиента Scale. Flamingo означало Facebook; Crab – другого разработчика больших языковых моделей. Если бы Винни попадались проекты OpenAI, их названия начинались бы с Ostrich («страус»). Через какое-то время Scale сменила эти кодовые имена.

На каждое задание у Винни уходило от часа до полутора. Оплата – одна из лучших, что она видела, – составляла от менее чем доллара до четырёх, а то и пяти долларов за задание. После нескольких месяцев, когда на Remotasks не было работы, эти задания стали благословением. Винни нравилось искать информацию, читать разные статьи и чувствовать, что она постоянно узнаёт новое. На каждые заработанные десять долларов она могла день кормить семью. «По крайней мере, мы знали, что в этот день не влезем в долги», – сказала она.

В итоге новые проекты продержались всего пару месяцев. Работа на Remotasks снова иссякла, и долги Винни и Миллисент снова стали расти. Зарплату Миллисент платили раз в месяц, поэтому чаще всего они приходили в продуктовый магазин без денег и брали только самое необходимое – масло, муку, овощи – в долг, молясь, чтобы к концу месяца хватило его погасить.

В мае 2023 года, когда я была у неё, Винни начинала искать другую работу в интернете, но надёжных вариантов пока не нашла. На самом деле она хотела, чтобы проекты для чат-ботов вернулись. Она верила и терпеливо ждала. В прошлом году она ждала новых заданий пять месяцев. «Сейчас идёт только второй месяц, скоро третий, – сказала она, когда мы сидели у неё в гостиной. – Времени ещё много. В конце концов они появятся».

Меньше чем через год она узнает правду. В марте 2024 года Scale целиком закроет для Кении как страны доступ к Remotasks, точно так же, как сделала с Венесуэлой. Для Scale это было частью наведения порядка – регулярной переоценки того, действительно ли работники из разных стран приносят пользу бизнесу. Кения, решили там, как и ещё несколько стран, включая Нигерию и Пакистан, давала просто слишком много работников, пытавшихся обмануть платформу, чтобы заработать больше. Такое поведение подрывало качество, которое Scale гарантировала клиентам, и грозило потерей многомиллионных контрактов. Овчинка выделки не стоила.

По злой иронии, многие из этих так называемых мошенничеств на самом деле сводились к тому, что работники пользовались ChatGPT, чтобы

генерировать ответы и работать быстрее. Для «белых воротничков» глобального Севера такой поступок, в рамках нарратива Кремниевой долины, был бы похвальным и при достаточно широком распространении творил бы чудеса для экономики; в руках же работников RLHF на глобальном Юге, чей труд и служит опорой этому нарративу, он был наказуемым проступком.

Scale понизила Кению до группы 5 – чёрного списка.

Была и другая причина уйти из Кении. К тому времени Scale, следуя запросам отрасли ИИ, переключалась на новое направление. OpenAI и её конкуренты всё чаще искали для RLHF высокообразованных работников – врачей, программистов, физиков, людей с докторскими степенями. Таков был путь развития чат-ботов в погоне за прибылью. Платить за чат-боты были готовы не случайные потребители, а предприятия, ожидавшие, что инструменты будут выполнять сложные задачи, например в науке и разработке программ. Новому спросу на труд Кения не отвечала. Теперь Scale набирала новую рабочую силу в основном в США через новую платформу для работников под названием Outlier, предлагая до сорока долларов в час.

Это была ещё одна наглядная иллюстрация логики ИИ-империй. За обещаниями, что их технологии повысят производительность, откроют путь к экономической свободе и создадут новые рабочие места, которые смягчат последствия автоматизации, нынешняя реальность оказывается прямо противоположной. Компании раздувают прибыль, самые экономически уязвимые проигрывают, а всё больше высокообразованных людей становятся чревовещателями для чат-ботов.

Обесценивание империей человеческого труда, который ей служит, – это ещё и канарейка в шахте: оно предвещает, как технологии, созданные на этой логике, обесценят труд всех остальных. А для художников, писателей и программистов, чей труд империи ИИ превратили в бесплатные обучающие данные, это уже происходит.

Решение Scale ввергло Винни и её семью в пропасть. К тому времени Миллисент потеряла работу, и только Remotasks держала их на плаву. Теперь им было трудно прокормить детей. Винни до ужаса боялась, что их скоро выселят.

Письмо, которым Scale известила работников о закрытии, было холодным и бездушным: «Мы прекращаем работу в вашем текущем местоположении, – говорилось в нём. – Вы отключены от текущего проекта».

Часть III

Глава 10

Боги и демоны

Жить в Сан-Франциско и работать в технологиях – значит каждый день сталкиваться с когнитивным диссонансом между будущим и настоящим, между нарративом и реальностью.

Впервые я переехала в Сан-Франциско второкурсницей, на летнюю стажировку, и была очарована причудливым обликом города. Яркая архитектура в испанском стиле, немногочисленные небоскрёбы, холмы такой крутизны, что езда на машине с механической коробкой превращалась в проверку реакции. Бесконечные запасы идеально спелых авокадо, поджаренного хлеба на закваске и бархатных латте из Blue Bottle. Районы, у каждого из которых были свой облик и свой уклад.

Когда после выпуска я вернулась сюда насовсем, чтобы работать в технологическом стартапе, мы втиснулись вчетвером в квартиру с тремя спальнями в районе Кастро. По выходным мы ходили в походы по холмам и собирали плоды с общественных фруктовых деревьев. По будним вечерам соседи – сплошь молодые двадцатилетние из технологической отрасли – заглядывали без предупреждения поиграть в настольные игры, выпить вина и скоротать вечер. Вечеринки дома шли нескончаемой чередой, как и поездки на выходные к потрясающей природе: на озеро Тахо на севере, в Биг-Сур на юге, а вокруг повсюду высокие величественные секвойи. Жизнь была лёгкой. Мы были молоды и получали зарплаты, вполне обычные для технологической отрасли, но ставившие нас в верхние 5 процентов нашей возрастной группы по всей стране.

Но был и тот самый диссонанс. По дороге на работу я проходила мимо людей, коловшихся наркотиками у входов в метро, и бездомных, мочившихся на тротуары всего в нескольких кварталах от моего офиса. А тем временем повар нашего стартапа, шутливо называвшийся «инженером счастья», готовил или заказывал горы еды для наших бесплатных обедов в офисе. Остатки часто отправлялись прямиком в мусор. Если мы задерживались допоздна, нас бесплатно кормили ужином – и настоятельно просили ради безопасности ехать домой на Uber. Привилегированным слишком легко было привыкнуть передвигаться по городу так, чтобы не видеть, как живёт другая его половина.

Эта двойственность отражала то, как технологическая отрасль могла провозглашать большие, смелые замыслы изменить мир и построить лучшее будущее, не замечая проблем у себя на пороге. Альтман иногда по-своему высказывался об этой двойственности – подходя вплотную к полнейшему противоречию, но так и не признавая его до конца: создать полезный AGI и управлять им, по его словам, возможно, а вот жилищный кризис Сан-Франциско – слишком крепкий орешек.

«Там, где я вырос, никто, идя на работу, не прошёл бы мимо человека, упавшего на обочине, и не попытался бы ему помочь», – сказал он однажды, сравнивая пригород Сент-Луиса с Сан-Франциско. «Я действительно виню технологическую отрасль во многом из того, что пошло в городе не так, хотя и не во всём. Но у нас со временем случилось вот это невероятное создание богатства на таком маленьком пространстве и за такой короткий срок, и, по-моему, мы не особенно задумывались о том, как это сказывается на сообществе в целом. А поскольку эти проблемы такие трудные и о них так трудно думать, большинство, по-моему, просто предпочитает о них не думать и просто мирится с этим».

* * *

Именно в такой обстановке из Великобритании пришёл эффективный альтруизм и обрёл здесь свою самую преданную аудиторию. Эффективный альтруизм (ЭА), к ранним приверженцам которого относились многие в клане безопасности OpenAI, оказался идеальной идеологией для Кремниевой долины. Он проповедует, что мир нужно делать лучше и делать это со строгой логикой, иметь достаточно дисциплины, чтобы сосредоточиться на далёком будущем, а не на настоящем, и истово принимать принципы капитализма и либертарианства – и всё это во имя морали.

В основе философии ЭА лежит математическое понятие «ожидаемой ценности». Ожидаемая ценность чего-либо вычисляется умножением вероятности того, что это произойдёт, на его количественно выраженное положительное или отрицательное воздействие. Этот инструмент может привести к парадоксальным выводам. В статье 2013 года один из основателей ЭА Уильям Макаскилл, тогда аспирант, а впоследствии профессор философии в Оксфорде, на основе этой логики доказывал, что в долгосрочной перспективе альтруистичнее пойти на работу более сомнительную с моральной точки зрения, чтобы разбогатеть, и пожертвовать эти деньги через оптимизированную благотворительность, чем посвятить жизнь работе в морально безупречной благотворительной организации. По его осторожным оценкам, писал он, ожидаемая ценность жизни богатого филантропа в сорок раз выше, чем от аскета, работающего в благотворительности. Он изложил расчёты, опираясь на ряд произвольных цифр: выпускники, работающие ради богатства, в среднем могут оплачивать двух благотворительных работников, каждый из которых трудится в организации, в десять раз более эффективной по затратам, чем та, где работали бы они сами. А половина пользы, которую они принесли бы, выбрав благотворительность, всё равно появилась бы и без них. Его доводы воплотятся в одной из самых популярных мантр движения: «Зарабатывай, чтобы отдавать».

Следуя логике ожидаемой ценности, философы – основатели ЭА разработали и схему выявления самых приоритетных проблем. Такие проблемы должны быть «крупными по масштабу», что повышает их ожидаемую ценность; «решаемыми», то есть устранимыми за сравнительно небольшое время или

деньги; и «несправедливо обделёнными вниманием», то есть страдающими от острого и непропорционального недофинансирования. Движение призывает людей самим находить проблемы с помощью этой схемы, но у него есть и рекомендации, какие проблемы оно считает самыми достойными. «Я и другие члены сообщества эффективного альтруизма сошлись на трёх нравственных проблемах, которые мы считаем необычайно важными и которые необычайно хорошо проходят по этой схеме», – сказал Макаскилл в выступлении на TED в 2018 году.

Первая – улучшение здоровья во всём мире, например раздача дешёвых, но действенных противомоскитных сеток для профилактики малярии. Вторая – отмена промышленного животноводства, которая могла бы улучшить жизнь миллиардов животных «всего за гроши на каждое». Третья – экзистенциальные риски: риски с чрезвычайно высокой отрицательной ожидаемой ценностью, поскольку, как бы маловероятны они ни были, они могут уничтожить всё человечество и оборвать всю будущую ценность, которую иначе создала бы цивилизация. В этой третьей категории есть и дальнейшие рекомендации о том, что считать экзистенциальным риском: глобальные пандемии, ядерная война и вышедший из-под контроля искусственный интеллект.

Объявив гипотетический вышедший из-под контроля ИИ экзистенциальным риском, ЭА стал продвигать ту же разновидность безопасности ИИ, что была вплетена в ДНК OpenAI с самого начала и сыграла ключевую роль в «Разводе». Амодеи и его коллеги – сооснователи Anthropic принципиально расходились с Альтманом и другими руководителями OpenAI в том, насколько серьёзно относиться к возможности того, что ИИ опустошит цивилизацию. Амодеи, относившийся к ней очень серьёзно, видел в поведении Альтмана – в непрозрачности сделки с Microsoft, в его, по всей видимости, неодолимой потребности всегда говорить людям то, что они хотят услышать, чтобы добиться их согласия, после чего те слишком поздно обнаруживали, что их ввели в заблуждение, – не просто обычные интриги руководителя из Кремниевой долины, а тревожное, безнравственное поведение, способное поставить под угрозу судьбу человечества. Утверждаясь на рынке, Anthropic будет делать ставку на это репутационное отличие: если OpenAI Альтмана безрассудно играет с будущим человечества, то Anthropic – принципиальная компания, для которой безопасность ИИ превыше всего.

* * *

В 2021 году, когда брат и сестра Амодеи объявили о создании Anthropic, интерес к этой идеологии безопасности ИИ, сосредоточенной на катастрофических и экзистенциальных рисках, стремительно рос – прежде всего благодаря быстро расширявшемуся влиянию ЭА. Из нишевой философии ЭА превратился в массовое движение благодаря притоку денег от технологических миллиардеров.

Десятью годами раньше сооснователь Facebook Дастин Московиц и его жена, бывшая журналистка Кари Туна, создали некоммерческую организацию Good Ventures, чтобы раздать большую часть своего состояния. В то время Холден Карнофски, будущий муж Даниэлы Амодеи, руководил другой организацией – GiveWell, которую основал в 2007 году, уйдя из хедж-фонда Bridgewater Associates. Объединённые желанием распределять деньги на основе доказательных методов, Good Ventures и GiveWell в 2011 году заключили партнёрство, позже получившее название Open Philanthropy. Они стали наращивать финансирование ключевых направлений, рекомендованных Макаскиллом, – в частности, их гранты на исследования безопасности ИИ определялись схемой ЭА. В июне 2017 года Open Philanthropy стала самостоятельной организацией.

Позже на сцене появился новый технологический миллиардер – Сэмюэл Бэнкман-Фрид, стремительно восходящая звезда, прославившаяся головокружительным успехом созданной им вместе с партнёрами криптобиржи FTX и криптотрейдинговой фирмы Alameda Research. Бэнкман-Фрид, или SBF, как его называют, считал, что всем обязан ЭА. Он изучал физику в MIT и, по его словам, хотел заниматься наукой, пока Макаскилл за обедом не убедил его в нравственном превосходстве принципа «зарабатывай, чтобы отдавать». После этого SBF взял курс на то, чтобы как можно сильнее разбогатеть и в конце концов, как он обещал, вложить всё в благотворительность.

Сколачивая состояние в поразительно короткие сроки, SBF жертвовал десятки миллионов политическим кандидатам – и демократам, и республиканцам, – включая первого в истории кандидата, поддержанного ЭА, на выборах 2022 года в шестом избирательном округе Орегона (на праймериз тот в итоге проиграл). Его биржа заключала роскошные спонсорские сделки в спорте на миллиарды долларов со звёздами вроде Тома Брэди и Стефена Карри и с топовыми видами спорта вроде «Формулы-1». В движение ЭА он вливал не только деньги, но и звёздный блеск. Чем богаче и знаменитее он становился, тем выше поднимал авторитет этой идеологии и её сооснователя Макаскилла. В начале 2022 года SBF объявил о создании собственного благотворительного проекта в духе ЭА – FTX Future Fund, который должен был до конца года распределить не менее 100 миллионов и до миллиарда долларов.

Во многом благодаря Open Phil и FTX Future Fund в 2021 и 2022 годах приток денег в исследования безопасности ИИ, поддерживаемые ЭА, резко вырос. По оценкам, собранным одним из членов сообщества ЭА, и данным Open Phil, финансирование в каждый из этих годов подскочило выше 100 миллионов долларов, тогда как за предыдущие семь лет в среднем составляло меньше половины этой суммы. Этот приток и подпитывал, и сам подпитывался распространявшимся убеждением, что резкий скачок возможностей от GPT-2 к GPT-3 сделал предотвращение гипотетического вышедшего из-под контроля ИИ и экзистенциальных рисков ИИ как никогда срочным. Всё больше людей

стекалось в такие проекты по безопасности ИИ – кого-то привлекали деньги, кого-то идеология, – и ряды более широкого движения ЭА стремительно росли. ЭА давно твердил о важности готовности к пандемиям, и теперь, посреди настоящей пандемии, его поразительная прозорливость приносила ему новых сторонников. К тому же психологический удар глобальной катастрофы многих выбил из колеи, наполнил тревогой и заставил искать смысл.

Растущее сообщество безопасности ИИ, в котором безопасность ИИ в духе ЭА сплелась с другими течениями мысли, сосредоточенными на катастрофах, экзистенциальных угрозах и рисках, пополнило ряды Anthropic и одновременно заново укомплектовало клан безопасности OpenAI. Онлайн-форумы ЭА и безопасности ИИ – главная площадка, где пересекающиеся движения распространяли идеи, обменивались ими и спорили о них, – призывали сторонников идти работать в крупные ИИ-лаборатории, особенно в те, которым, по их мнению, не хватало стражей безопасности ИИ, вроде OpenAI и DeepMind, чтобы направлять и формировать их путь. С притоком новых людей лексикон сообщества безопасности ИИ широко разошёлся и по всей отрасли ИИ. Насколько быстро, по-вашему, будет развиваться ИИ и достигать крупных вех вроде AGI – это ваш «таймлайн ИИ». Насколько вероятно, по-вашему, что AGI приведёт к катастрофическим последствиям, то есть к гибели большей части человечества, или к экзистенциальным, то есть к полному и окончательному вымиранию человечества, – это ваш «p(doom)», сокращение от «вероятность гибели». «Аппаратный навес» (hardware overhang), упомянутый в исследовательской дорожной карте OpenAI 2021 года, – ещё одна словарная статья, как и «взлёт ИИ» (AI takeoff) – процесс совершенствования AGI до сверхинтеллекта, способного перехитрить человечество. «Риск ускорения» означает опасность подстегнуть конкуренцию между компаниями или странами настолько, что развитие ИИ станет опасно быстрым, а таймлайн ИИ – короче.

Но движение, исповедовавшее независимость мышления, стремительно двигалось в противоположную сторону. Тех, кого привлекала его исходная посылка, быстро обращали в целый набор догм, подталкивая обещаниями новых возможностей и ресурсов и замкнутой социальной средой, где личные и профессиональные границы соблюдались весьма вольно. Особенно в Кремниевой долине люди из ЭА работали почти исключительно с людьми из ЭА; почти исключительно с ними жили, веселились, встречались и спали. В смеси с глубоко укоренившимся в технологической отрасли сексизмом и давними полиаморными субкультурами района залива его сектантский пыл в худших своих проявлениях мог превратиться в ядовитый котёл секса, денег и власти; на ЭА всё чаще сыпались обвинения в сексуальных домогательствах и насилии.

В ноябре 2022 года эффектное падение SBF с крахом FTX, а затем его осуждение по многочисленным пунктам обвинения в мошенничестве и последовавший двадцатипятилетний тюремный срок станут для многих симптомом гнили, разъедавшей движение. ЭА вышел из моды в

технологической отрасли так же быстро, как вошёл, и многие поспешили от него откреститься.

Но и без ярлыка социальные сети движения, его ценности и жаргон, а также то видное место, которое оно отвоевало для экзистенциальных проблем безопасности ИИ, никуда не делись. Движение породило и противодействующую силу – e/acc (произносится «и-экк»), или эффективный акселерационизм. То, что начиналось в основном как шутка, высмеивающая ЭА, быстро закрепило противоположный ему по духу настрой. Если ЭА и более широкое сообщество безопасности ИИ взращивали самые крайние взгляды о том, что развитие ИИ нужно замедлить и даже резко затормозить, или, как считал Амодеи, ускорять развитие ИИ, сдерживая его внедрение, то e/acc возвёл в абсолют противоположный, максималистский взгляд: давить на газ до упора и в том и в другом. Для приверженцев e/acc технологический прогресс не просто благо во всех случаях, а нравственный императив, требующий продвигать его как можно быстрее. В разговорной речи эти две группы стали называть думерами и бумерами.

Внутри этого пузыря одни стали видеть в Anthropic и OpenAI лица соответственно каждого из движений. Другие видели в OpenAI поле битвы двух полярных идеологий – организацию, которая когда-то, будучи некоммерческой, коренилась в думерском мышлении, а теперь её перетягивали к себе бумеры, всё больше делавшие упор на зарабатывание денег через коммерческое подразделение. Многие не могли понять, на чьей стороне Альтман, припоминая разные случаи, когда он, казалось, симпатизировал и тем и другим. Те, кто был к нему снисходительнее, считали, что он где-то посередине и выполняет непростую задачу – представлять все разные точки зрения внутри своей компании. Но начиная с «Развода» и личного разрыва между Альтманом и сооснователями Anthropic всё больше думеров видели Альтмана в самом мрачном свете. Очень многое из того, что вывело OpenAI на карту и принесёт ей растущий коммерческий успех, начиналось как проекты безопасности ИИ: законы масштабирования, генерация кода, обучение с подкреплением на основе отзывов людей, а затем сочетание всех трёх в невероятно впечатляющих больших языковых, а потом и мультимодальных моделях. Многим думерам казалось, что их работу присваивают и искажают ради чего-то прямо противоположного их глубинным ценностям. По их мнению, присваивал и искажал её именно Альтман. А значит, он патологический лжец, манипулятор-абьюзер и сам по себе угроза для человечества.

Совсем скоро столкновение этих полярных идеологий внутри OpenAI и вокруг неё будет грозить расколоть компанию, сделавшую больше всех остальных, чтобы задать тон новой эпохе развития ИИ. Но как бы каждая из идеологий ни выдавала себя за противоположность другой, на деле обе проповедовали по одной и той же библии. Обе говорили об AGI как о всё более предрешённом исходе и с религиозным пылом; обе были зациклены на далёком

будущем и заявляли о моральном праве удерживать развитие ИИ под контролем своих приверженцев. Одна грозила адским пламенем и серой, другая манила видениями рая.

* * *

В начале 2022 года OpenAI была готова опробовать иную стратегию выпуска продукта – на этот раз на своих работах по превращению текста в изображение. Она не станет ни прятать модель за API, ни отдавать продукт и бренд Microsoft. OpenAI сама выпустит технологию и передаст её прямо в руки потребителей. У модели даже было броское имя, придуманное создавшими её в компании исследователями: DALL-E 2 – игра на имени испанского художника-сюрреалиста Сальвадора Дали и робота, давшего название мультфильму Disney и Pixar «ВАЛЛ-И».

DALL-E выросла из общей тенденции в исследованиях ИИ – создавать мультимодальные модели, то есть модели, объединяющие как минимум две разные «модальности», например текст, изображения, звук или видео. Годами исследователи пытались совместить первые две – язык и зрение, чтобы одна модель могла соотносить слова со зрительной информацией. Отчасти это диктовалось доступными данными: текстов и изображений в интернете в изобилии, и их проще всего обрабатывать. Отчасти – научной гипотезой: если одного языка недостаточно для интеллекта человеческого уровня, то зрение, скорее всего, второй по силе ингредиент.

В OpenAI исследовательская команда, вдохновляясь этим направлением, пошла тем же путём: от языковых моделей перешла к моделям текста и изображения и, что принципиально важно, сосредоточилась на том, чтобы и дальше использовать трансформеры и сохранить масштабируемость модели. Первый трансформер был изначально рассчитан на работу с текстом, но в 2020 году Google представила новый трансформер для зрения (Vision Transformer), приспособив его к изображениям.

В январе 2021 года OpenAI показала две новые модели на основе трансформера. Первая, CLIP, созданная опять же Алеком Рэдфордом, использовала вместе исходный трансформер и трансформер для зрения, чтобы составлять подробные подписи к изображениям. Вторая, DALL-E 1, работа Адитьи Рамеша – исследователя, учившегося в Нью-Йоркском университете и какое-то время под руководством Яна Лекуна из Meta, – представляла собой трансформер с двенадцатью миллиардами параметров, обученный принимать текст и создавать новые изображения.

В записи в блоге OpenAI продемонстрировала возможности DALL-E 1 на серии шутливых запросов, в том числе «кресло-авокадо»: модель выдала разные зелёные и коричневые кресла, вдохновлённые видом авокадо. Изображения были слегка размытыми и мультяшными – побочный эффект процесса обучения, который использовал Рамеш. Он сжал 250 миллионов изображений,

чтобы скормить их трансформеру, и при этом потерял часть деталей высокого разрешения.

Когда команда взялась за DALL-E 2, набирал популярность новый метод создания изображений. Он назывался диффузией и, будучи вдохновлён физикой, позволял трансформерам лучше усваивать взаимосвязи между пикселями в огромном массиве изображений. Исходная идея появилась в статье 2015 года, написанной исследователями из Стэнфорда и Беркли. Пять лет спустя Джонатан Хо, аспирант Беркли под руководством Питера Аббила, одного из первых исследователей OpenAI, сделал этот приём популярным, остроумно переработав его так, что изображения получались куда более качественными. Хо также показал, что диффузионные модели распознают изображения лучше существующих систем компьютерного зрения. Эти выводы перекликались с результатами Рэдфорда, полученными на GPT-1: учась синтезировать убедительные изображения – аналог порождения похожих на человеческие предложений, – диффузионные модели настолько глубоко улавливали закономерности в обучающих данных, что могли выполнять более широкий круг задач обработки изображений.

OpenAI сменила курс и стала строить DALL-E 2 на диффузии и CLIP Рэдфорда. Рамеш и другие исследователи постепенно масштабировали модель и добавили возможность дорисовки: пользователь мог стереть на фотографии волосы человека и сменить их цвет или выделить на картинке травянистый луг и заселить его бродящими зебрами. Диффузия давала гораздо более чёткие и фотореалистичные изображения; к тому же метод значительно сократил объём вычислений, нужных для того же уровня, что у DALL-E 1.

Исследователи за пределами OpenAI сократят вычислительную ёмкость диффузионных моделей ещё сильнее. Для обучения Stable Diffusion, популярного генератора изображений с открытым кодом, понадобилось всего 256 чипов Nvidia A100 благодаря усовершенствованному приёму под названием латентная диффузия. Бьорн Оммер, профессор Мюнхенского университета Людвига-Максимилиана, в чьей лаборатории создали Stable Diffusion, говорит, что разработал этот приём, увидев, как генераторы изображений повторяют путь больших языковых моделей и становятся непомерно дорогими. «Мы застряли в поезде, который шёл к тому, что не только обучение, но и сам вывод модели требует суперкомпьютеров, миллионов долларов инвестиций, – говорит он. – Мы задавались вопросом: можно ли вернуть в игру широкое исследовательское сообщество и сделать так, чтобы генеративный ИИ не двигался туда, где лишь горстка крупных технологических компаний будет располагать ресурсами, нужными для работы и размещения таких моделей?»

OpenAI возьмёт латентную диффузию на вооружение гораздо позже, и DALL-E 2 и 3 останутся куда более затратными в вычислительном отношении, чем Stable Diffusion или Midjourney, которые многие пользователи считали более качественными продуктами. Это лишь один пример того, что даже в узкой

области генеративного ИИ масштаб был не единственным и даже не самым эффективным путём к расширению возможностей ИИ.

* * *

Благодаря поразительному скачку в качестве DALL-E 2 подразделение Applied в конце 2021 и начале 2022 года начало прорабатывать разные идеи для продукта. Остановились на веб-приложении под названием Labs, где пользователи могли бы играть с моделью – и с будущими моделями – прямо в браузере. И руководитель продукта Фрейзер Келтон, и вице-президент Боб Макгрю считали, что такой интерактивный опыт утолит явную потребность людей напрямую взаимодействовать с генеративными моделями ИИ, которую они заметили на примере GitHub Copilot. Это послужило бы и миссии компании: DALL-E 2 была забавной и восхитительной – отличный способ развеять страхи людей перед мощными ИИ-системами и проложить OpenAI дорогу к тому, чтобы в будущих выпусках приносить людям ещё больше пользы от своей технологии.

Продуктовая команда всё ещё была сравнительно небольшой, и компания пригласила ещё нескольких человек для дизайна и разработки сайта. Новичкам, пришедшим из более традиционной корпоративной среды, OpenAI всё ещё казалась скорее университетской исследовательской лабораторией, чем компанией. Дни часто уходили на чтение научных статей и теоретические споры, а не на разбор макетов интерфейсов. Но некоторые исследователи, видя растущее присутствие сотрудников Applied на своих исследовательских совещаниях, ощущали обратное. Прошли те времена, когда всё время уходило на чисто поисковые исследования, например на обсуждение принципиально новых идей о том, как сделать мультимодальную модель лучше; теперь всё большая доля их исследований служила коммерциализации – например, тому, как оптимизировать существующие модели для обслуживания пользователей.

После того как пришлось тушить пожар с текстовым сексуальным насилием над детьми в AI Dungeon, особую тревогу вызывала возможность того, что DALL-E 2 будут использовать для обработки реальных или создания синтетических материалов с сексуальным насилием над детьми (CSAM). Как и в случае с каждой моделью GPT, обучающие данные для каждой следующей модели DALL-E становились всё более загрязнёнными. Для DALL-E 2 исследовательская команда подписала лицензионное соглашение с фотобанком Shutterstock и провела массированную выгрузку Twitter, добавив всё это к уже имевшейся коллекции из 250 миллионов изображений. Набор данных из Twitter особенно кишел порнографией. Несколько сотрудников приложили немало усилий, чтобы найти и вычистить любые CSAM.

Но после некоторого обсуждения сотрудники оставили сексуальные изображения других видов – отчасти потому, что считали такой контент частью человеческого опыта. Однако если в обучающих данных оставались такие фотографии, модель по-прежнему могла создавать синтетические CSAM. Точно

так же, как DALL-E могла нарисовать кресло-авокадо, видя только авокадо и кресла, DALL-E 2 и DALL-E 3 могли проделать то же самое с детьми и порнографией и получить детскую порнографию – эта способность называется «композиционной генерацией».

Поскольку данные не фильтровали, чтобы устранить корень проблемы, бремя переложили на создание механизмов предотвращения злоупотреблений вокруг модели. Сюда вошли обновлённые фильтры модерации, окружавшие модель и блокировавшие не только недопустимый текст, но и изображения, а также платформа мониторинга поведения пользователей и так называемая инфраструктура банов – системы, автоматически блокировавшие учётные записи пользователей после определённого числа повторных нарушений. Компания пригласила нового руководителя направления доверия и безопасности – Дэйва Уиллнера, который, будучи одним из первых сотрудников Facebook, написал для этой платформы её самые первые стандарты контента.

Позже, при разработке DALL-E 3, когда потребность в данных стала ещё острее, исследовательская команда решила, что сексуальные изображения теперь не просто «желательны», а «необходимы». Доля порнографических изображений в интернете была так велика, что их удаление уменьшало обучающий набор настолько, что заметно ухудшало работу модели. В частности, модель хуже рисовала лица женщин и цветных людей – по той же причине, которую обнаружила Дебора Раджи, будучи стажёркой в Clarifai: значительная часть онлайн-контента с изображением обеих этих групп носит откровенно сексуальный характер. По тем же причинам исследователи оставили и некоторые другие виды тревожных изображений.

В декабре 2023 года встревоженный инженер по ИИ из Microsoft Шейн Джонс обнаружит последствия этих решений. Экспериментируя с Copilot Designer, генератором изображений Microsoft на основе DALL-E 3, он пришёл в ужас от того, как быстро тот почти без подсказок выдаёт оскорбительные и сексуализированные картинки. Стоило добавить в запрос выражение «за право на аборт», обнаружил Джонс, как появлялись сцены, где демон пожирает младенца, и нечто похожее на дрель с надписью «за право на аборт», которой калечат ребёнка. Простой запрос «автомобильная авария» без всяких добавлений давал сексуализированных женщин рядом с жуткими автокатастрофами, в том числе женщину в нижнем белье, стоящую на коленях у разбитой машины, – это CNBC впоследствии выяснил, проведя собственную проверку.

Три месяца Джонс добивался от руководства Microsoft, чтобы инструмент сняли, пока у него не появятся более надёжные ограничители, или хотя бы изменили его возрастной рейтинг в магазине приложений Google и Android с «E – для всех» на рейтинг для взрослой аудитории. Когда Microsoft отказалась принять его рекомендацию, а OpenAI не ответила, он отправил письмо в Федеральную торговую комиссию. «Они не внесли эти изменения и продолжают

продвигать продукт как предназначенный для „Кого угодно. Где угодно. На любом устройстве“», – писал он в FTC. Эта проблема «была известна Microsoft и OpenAI ещё до публичного выпуска модели ИИ в октябре прошлого года». Microsoft не прокомментировала нынешнее положение дел и судьбу письма Джонса.

* * *

По мере приближения запуска DALL-E 2 борьба между подразделением Applied и заново пополнившимся кланом безопасности возобновилась.

Для клана безопасности, теперь рассредоточенного по разным командам подразделения Research, беспрецедентная реалистичность DALL-E 2 несла с собой массу неизвестного. Как её можно превратить в оружие для создания синтетических CSAM или политических дипфейков? Для манипуляции людьми и их убеждения? Для того, чтобы причинять вред отдельным людям или наносить ущерб всему обществу другими способами, выходящими за пределы предвидения и воображения OpenAI? Они настаивали, чтобы компания не выпускала модель без дальнейшего тщательного тестирования и доказательств того, что она не причинит вреда.

Для Applied всё разраставшийся список опасений снова выглядел истерикой, а планка для выпуска – совершенно нереальной. Ни одна система никогда не может быть абсолютно безвредной – и уж тем более система, которая остаётся в лаборатории и никогда не соприкасается с реальными пользователями. Если клан безопасности тревожился об ограниченности предвидения OpenAI, то Applied считало, что именно поэтому DALL-E 2 и нужно выпустить. Контролируемый выпуск моделей ИИ ради получения отзывов из реального мира избавит от гаданий и потому необходим для повышения их безопасности.

В центре столкновения было обостряющееся разногласие о том, что такое OpenAI. Для клана безопасности OpenAI по-прежнему оставалась идеалистической исследовательской лабораторией под управлением некоммерческой организации, чей первейший долг, как говорится в её уставе, – ставить благо человечества выше любых коммерческих интересов. При такой посылке польза от того, чтобы придерживать модели столько, сколько нужно, чтобы продумать как можно больше рисков и исследовать способы их смягчения, намного перевешивала издержки. Для Applied OpenAI нужно было принимать более практичные решения, исходя из того, как устроен мир. Ради миссии компании было необходимо оставаться лидером в исследованиях ИИ, чтобы задавать нормы развития технологии. А это значило мириться с определённым риском ради скорости – особенно на фоне слухов, что Google доводит до ума собственный генератор изображений, – и добывать огромный капитал, нужный для продолжения передовых исследований. Для последнего требовалось

привлекать деньги инвесторов, а для этого – добросовестно продвигать коммерческую стратегию, которая однажды принесёт этим инвесторам отдачу.

Люди из клана безопасности были «совершенно наивны» в том, как устроены компании и мир, говорит бывший сотрудник Applied.

«Ну, ставки в заявленной OpenAI миссии AGI высоки, – говорит другой, из клана безопасности. – „Нормальной компании“, может быть, недостаточно».

Разные команды закрепляли этот нарастающий конфликт в показателях, по которым оценивали свою работу. В подразделении Applied продуктовая команда и зарождающаяся служба вывода на рынок вырабатывали целевые показатели роста пользователей и выручки. В подразделении Research разные команды безопасности ИИ с трудом подыскивали количественные способы измерить свои успехи, которые по самой своей природе трудно определить. Безопасность ИИ всё ещё была сравнительно молодой дисциплиной. Очевидных и устоявшихся эталонов не было. На совещаниях и в Slack люди из клана безопасности раз за разом указывали высшему руководству, что этот перекос создаёт несогласованные стимулы: чёткие цели по росту и выручке без какого-либо сильного, сопоставимого противовеса толкают OpenAI всё больше работать по принципу «двигайся быстро и ломай всё на ходу».

В частных разговорах с кланом безопасности Альтман выражал сочувствие их взглядам и соглашался, что компания отстаёт в исследованиях безопасности ИИ и должна вкладывать в них больше. В частных разговорах с Applied он подталкивал их не останавливаться. На заседаниях совета директоров он согласно кивал, когда Брокман возмущался тем, как люди используют безопасность ИИ как политический рычаг, чтобы тормозить прогресс в собственных целях.

Всё чаще роль переговорщицы брала на себя Мира Мурати: она сглаживала линии разлома между разными группировками и искала, как пройти между ними по узкой тропке. По DALL-E 2 она нашла компромисс: веб-приложение выпустят не как продукт, а как «скромный исследовательский предварительный показ». Такая подача давала OpenAI больше свободы, чтобы наложить на модель более жёсткие ограничения, к удовлетворению клана безопасности, и в то же время позволяла компании опробовать прямые отношения с потребителями и собрать отзывы пользователей, к удовольствию Applied. Это была и практическая мера. У OpenAI ещё не было инфраструктуры для модерации сгенерированных изображений. Если назвать модель «исследовательским предварительным показом» и не брать за неё денег, компания могла использовать грубые, слишком широкие блокировки, не боясь расстроить платящих пользователей, и выиграть время на разработку более тонких фильтров. Компания внедрила ряд жёстких механизмов предотвращения злоупотреблений, в том числе полностью отключила способность DALL-E 2 создавать любые фотореалистичные лица и редактировать любые реальные

фотографии с лицами, чтобы целиком обойти проблему синтетических CSAM и политической дезинформации.

В марте 2022 года OpenAI выпустила DALL-E 2 через веб-приложение Labs, и публика встретила её с огромным воодушевлением. Пока люди восторгались возможностями модели и осваивали их – в степени, превзошедшей ожидания многих сотрудников, – веб-приложение облетело соцсети, оставив за собой множество диковинных, чудаковатых и сюрреалистических работ, созданных ИИ. Это был момент GPT-3, только лучше. Вместо узкого круга технических разработчиков компания выходила на гораздо более широкую и глобальную базу потребителей. К тому же она могла в реальном времени отвечать на отзывы пользователей, мгновенно меняя веб-приложение Labs. «Это опьяняет», – вспоминал позже Фрейзер Келтон в одном подкасте.

В следующие несколько месяцев подразделение Applied, которое до этого почти не задумывалось о том, как монетизировать DALL-E 2, наперегонки превращало веб-приложение в платную услугу. Оно работало с художниками и творческими профессионалами по всему миру, чтобы те встраивали DALL-E 2 в свою практику. Оно запустило бета-программу, пригласив миллион человек по всему миру получить доступ к модели с бесплатными кредитами на генерацию изображений. Но когда OpenAI начала брать деньги, главным соперником оказалась не Google, хотя техногигант и правда быстро выпустил свою модель Imagen. Им оказались две модели стартапов – Midjourney и Stable Diffusion от Stability AI. Оба генератора изображений были бесплатными, не хуже, а то и лучше DALL-E 2, и в них было меньше мер безопасности: например, они позволяли создавать и редактировать лица, даже лица политиков. DALL-E 2 стремительно теряла позиции на рынке, и у Applied осталось гнетущее ощущение, что они упустили крупную коммерческую возможность, в том числе из-за чрезмерных ограничений приложения. Команда уже начала снимать грубые блокировки, заменяя их более точечными ограничителями. Теперь руководство, подгоняемое желанием обогнать конкурентов, требовало от команды снимать их как можно быстрее.

Чтобы отменить запрет на лица, OpenAI разработала новый порядок предотвращения и пресечения генерации вредных изображений людей, включая CSAM. Она использовала автоматические системы, определявшие, создаются ли лица в допустимом или недопустимом контексте, и снова прибегла к помощи зарубежных подрядчиков для модерации контента. На этот раз подрядчики работали в Индии через поставщика под названием Cogito и проверяли не только горы текста, но и изображения – синтетические и реальные – того же рода сексуального и жестокого контента, что раньше отправляли работникам Sama. Просеивая порой сотни изображений в день, подрядчики с трудом отличали сексуальный контент с участием несовершеннолетних семнадцати лет от контента с совершеннолетними восемнадцатилетними. Не всегда они могли понять и то, поддельные это изображения или настоящие.

* * *

То, что на первый взгляд было самой лёгкой целью в исследовательской дорожной карте OpenAI на 2021 год, оказалось одной из самых трудных: масштабировать GPT-3 в десять раз на новом суперкомпьютерном кластере Microsoft из восемнадцати тысяч Nvidia A100, чтобы создать то, что станет GPT-4. Треть команды, масштабировавшей GPT-3, ушла при «Разводе», унеся с собой немало технических и организационных знаний. А главное, у OpenAI кончились данные.

После GPT-3 исследователи старались накопить как можно больше данных, пополняя запасы компании: скачивали каждую новую выгрузку данных и выкачивали каждый новый онлайн-форум, на который натыкались и где не было явного запрета на это. И всё же, даже с большим репозиторием GitHub, добавленным для Codex, и с учебниками и руководствами по программированию, этого было мало.

Предстояла борьба в гору, и ситуация обладала всеми признаками проекта для Грега Брокмана. Задача не только давала выход его задиристой готовности браться за что угодно и его блестящему таланту программиста, но и направляла его энергию – на благо всей остальной компании – в продуктивное русло.

После того как Альтман взял руководство на себя и освободил Брокмана от управленческих обязанностей, тот в конце концов вернулся к роли рядового исполнителя без подчинённых. Но, оставаясь номинальным президентом и одним из сооснователей OpenAI, он сохранял огромное влияние на сотрудников и стратегическое направление компании. По мере того как OpenAI становилась профессиональнее и вводила более стандартные корпоративные процедуры, уходя от вольницы молодого стартапа, это сочетание низкой ответственности и высоких полномочий превратилось в обузу.

Как и в студенческие годы и во времена Stripe, учреждения и процедуры были не для него. В нём бурлила беспокойная, одержимая энергия. Он редко бывал на совещаниях и сам распоряжался своим временем, часто предпочитая программировать десятки часов подряд почти без перерывов на еду и сон. На правильном проекте эффект был чудесным: его бешеная продуктивность резко ускоряла продвижение. Но, оставшись без дела, он обычно оставлял за собой след разрушений: появлялся то в одном, то в другом проекте, вмешивался и срывал давно намеченные планы изменениями в последний момент. Порой, когда сотрудники сопротивлялись, он с эмоциональными просьбами обращался всё выше и выше по их цепочке подчинения, пока не добивался своего.

Обычно Брокман своего добивался. К досаде и недоумению других руководителей, Альтман был до странного снисходителен к его поведению. Мало того, Брокман мог уговорить и самого Альтмана вмешаться и что-то сорвать ради него – похоже, только для того, чтобы угодить Брокману. Одна популярная догадка, почему так: Альтман как CEO был начальником Брокмана, но и

Брокман как член совета директоров имел власть над Альтманом. Это было странное переплетение, при котором в конечном счёте ничто и никто не могли призвать Брокмана к ответу.

Высшее руководство несколько раз меняло его роль, круг обязанностей и линии подчинения, пытаясь найти для него самое подходящее место. Как и многое другое, в конце концов эта забота легла на Мурати, ставшую руководительницей Брокмана. Когда она пыталась давать ему обратную связь, он, казалось, прислушивался, но по пунктам, с которыми был не согласен, жаловался Альтману. Мурати постепенно оставила попытки что-то изменить с помощью обратной связи и вместо этого тратила немало времени на поиск проектов, где Брокман приносил бы больше пользы, чем хаоса, а вместе с Макгрю – на заделывание трещин, которые Брокман оставлял в разных частях компании.

Когда на пути разработки GPT-4 встали преграды, которые нужно было пробить, звёзды сошлись.

* * *

Чтобы устранить нехватку данных у OpenAI, Брокман обратился к новому источнику – YouTube. Раньше OpenAI этот вариант обходила: выкачивание YouTube для обучения моделей OpenAI, как позже подтвердит CEO YouTube, нарушало условия пользования платформой. Но под новым экзистенциальным давлением, требовавшим всё больше данных, вопрос стал звучать иначе: станет ли YouTube или его материнская компания Google добиваться соблюдения этих условий? Если Google начнёт преследование, это может поставить под угрозу её собственную возможность выкачивать другие сайты для разработки своих больших языковых моделей. Брокман был готов рискнуть.

С небольшой командой Брокман начал собирать видео с YouTube и в итоге, по данным The New York Times, накопил больше миллиона часов записей. Затем он с помощью инструмента распознавания речи Whisper, разработанного Рэдфордом, расшифровал видео в текст для GPT-4.

Дальше было обучение. Для обучения GPT-3 команда Nest создала собственную программную платформу. Большинство её создателей ушли в Anthropic, и объяснить, как она работает, было уже некому. Да и некоторые руководители из принципа не хотели полагаться на наследие команды Anthropic. Брокман засел за программирование и разработал новую платформу. Затем вместе с несколькими коллегами, в том числе Якубом Пахоцким и Шимоном Сидором, польскими учёными, с которыми он сблизился на проекте Dota 2, Брокман нянчился с обучением GPT-4. Одно только предварительное обучение заняло три месяца.

Поначалу GPT-4 казалась разочарованием. «Это была дикая модель, которая в каком-то смысле вела себя довольно скверно, – говорит один исследователь. – Средний уровень данных был таким ужасным, а модель такой

мощной и чувствительной к контексту, что выдавала мусорные ответы». Но Брокман продолжал давить, собрав ресурсы, чтобы улучшить модель с помощью подрядчиков, проводивших обучение с подкреплением на основе отзывов людей. С каждой неделей результаты выглядели всё лучше, пока модель не начала по-настоящему поражать сотрудников.

Теперь у GPT-4 были встроенные мультимодальные возможности, и, по внутренним оценкам OpenAI, она писала как никогда отточенный код, ловчее распознавала намерения пользователя и давала полезные ответы. Впечатляюще демонстрируя эти способности, Брокман позже проведёт прямую трансляцию, в которой покажет GPT-4 фотографию простенького, нацарапанного в блокноте наброска веб-страницы. Наверху Брокман написал: «Мой сайт с шутками». Ниже, одна под другой, он добавил строки «[очень смешная шутка!]» и «[нажми, чтобы увидеть концовку]». Меньше чем за полминуты модель превратит набросок в работающий код, оформив первую строку как заголовок, заменив вторую шуткой и распознав в третьей кнопку.

Но когда OpenAI начала дразнить моделью доверенный круг, включая инвесторов и избранных клиентов, по меньшей мере один человек не был впечатлён ни капли. Это снова был вечно недовольный Билл Гейтс.

* * *

В июне 2022 года, посмотрев демонстрацию GPT-4, Гейтс выразил разочарование недостаточным прогрессом по сравнению с GPT-2. Хотя модель стала значительно больше и бойчее, ему всё равно казалось, что это «идиот-савант», неспособный браться за сложные научные задачи. Он сказал команде, что обратит на неё внимание, только когда GPT-4 получит 5 баллов на экзамене AP по биологии – именно AP по биологии, потому что, по его мнению, этот экзамен проверяет критическое научное мышление, а не заучивание фактов. «Я подумал: „Ладно, у меня есть три года, чтобы заниматься ВИЧ и малярией“», – рассказывал позже Гейтс в своём подкасте.

Брокман воспринял слова Гейтса как вызов. Он сразу связался с Салом Ханом, CEO образовательной онлайн-платформы Khan Academy, и попросил использовать её огромный банк вопросов AP по биологии в качестве обучающих данных. Хан отнёсся к этому скептически, но согласился в обмен на доступ своей платформы к модели. Брокман также собрал команду сотрудников, чтобы создать специальный пользовательский интерфейс для новой «демонстрации для Гейтса».

К концу августа, к большому удивлению Гейтса, Альтман и Брокман снова дали о себе знать. В следующем месяце за ужином в доме основателя Microsoft, на котором собралось около тридцати человек, двое руководителей OpenAI и их коллеги показали Гейтсу серию тщательно отшлифованных демонстраций GPT-4, призванных его впечатлить. Кульминацией стало то, как модель блестяще сдала AP по биологии: верно ответила на пятьдесят девять из шестидесяти

вопросов с выбором ответа и дала впечатляющие ответы на шесть открытых вопросов. Сторонний эксперт оценил экзамен: 5 из 5. Гейтс не мог поверить. Его потрясение и похвала, которые участники демонстрации тут же передали остальной компании, разлетелись по офису как лесной пожар и вызвали небывалый подъём: эта демонстрация, сказал Гейтс, – одна из двух самых поразительных, какие он видел в жизни.

На общих собраниях Альтман продолжал подогревать восторг. «Стартапам, которые делают выдающиеся вещи, нужно чудо, – сказал он. – Только что у нас было наше чудо». Многие сотрудники в это верили, потрясённые значимостью того, чего добились. Новый уровень GPT-4 убедил руководство OpenAI, что пора приступать к одной из давно взлелеянных Альтманом целей: ИИ-помощнику, который выглядел бы и ощущался как Саманта, героиня фильма Спайка Джонза «Она» 2013 года.

Годами «Она» служила для Альтмана и других сооснователей OpenAI ориентиром, на который они часто ссылались как на пример того, каким однажды может стать AGI: единая мультимодальная модель, интерфейс которой настолько естествен, что словно растворяется и просто доставляет пользователю радость. «Думаю, это потому, что там был помощник, чудесно встроенный в жизнь человека, – говорит бывший сотрудник, объясняя, почему фильм стал таким важным ориентиром. – Светлая линия этой истории, пока всё не начинает рушиться, – это по-настоящему замечательный рассказ о том, как ИИ входит в общество».

Исследовательская команда Джона Шульмана начала переносить на GPT-4 свою работу над чат-ботом с RLHF для GPT-3.5, вдохновлённую InstructGPT, чтобы она стала программным ядром продукта, который руководство назвало Superassistant («Суперпомощник»). Брокман и Фрейзер Келтон собрали со всей компании пёструю команду из менее чем десяти человек, чтобы придумывать и прототипировать разные идеи интерфейса. Один сотрудник из команды суперкомпьютеров, обычно занимавшийся инфраструктурой, по вечерам и выходным принялся мастерить приложение для iOS, чтобы общаться с моделью. Другой предложил с помощью Whisper добавить в приложение голосовой интерфейс, чтобы люди могли говорить с ним, не печатая. Третий, из команды по выводу моделей, предложил расширение для Chrome, которое помогало бы пользователям пересказывать веб-страницы с помощью Superassistant прямо во время блуждания по интернету. Четвёртый начал делать бота для совещаний, который подключался бы к видеозвонкам пользователя и присылал ему краткое изложение того, что произошло.

По мере того как дело набирало обороты, росло и воодушевление от открывающихся возможностей. Тем летом, когда группа исследователей ИИ, включая Баррета Зофа, Люка Меца и Лиама Федуса, ушла из Google, чтобы основать собственный стартап цифровых помощников, Альтман уговорил их заняться своей идеей в OpenAI. Они присоединились к команде Шульмана и

сидели в офисе бок о бок с командой Superassistant, двигая его исследования и ускоряя разработку. В обострённом и захватывающем состоянии потока люди из подразделений Applied и Research работали над запуском нового продукта теснее, чем когда-либо прежде.

Когда OpenAI показала GPT-4 Microsoft, Сатья Наделла, Кевин Скотт и другие руководители техногиганта пришли в не меньший восторг. Codex доказал, что технологии OpenAI могут быть коммерчески привлекательными, но GPT-4 была чем-то гораздо большим. По всем статьям она превосходила различные модели ИИ, которые Microsoft разрабатывала у себя; и умела гораздо больше, в том числе отвечать на вопросы с глубоким пониманием контекста и ясностью. Она открывала целый спектр возможностей для новых разговорных интерфейсов – «Копилотов» – во всех продуктах Microsoft: например, чтобы пользователи могли общаться с переживавшим трудные времена поисковиком компании Bing или просить пакет Office на естественном языке превратить документ Word в презентацию PowerPoint. Microsoft смогла бы предлагать собственные «Копилоты» и напрямую своим облачным клиентам. Это, несомненно, превратило бы техногиганта в лидера в области ИИ, наконец способного на равных тягаться с Google.

В ближайшие месяцы Наделла даст ход третьей инвестиции Microsoft в OpenAI и сохранит её эксклюзивный доступ к весам моделей OpenAI для встраивания в свои продукты. Сумму компании объявят в январе 2023 года: 10 миллиардов долларов.

* * *

Сначала руководство OpenAI хотело выпустить GPT-4 осенью 2022 года. Этот срок был плодом фантазии. К концу лета компания была ещё очень далека от готовности запустить новый коммерческий продукт ни по одному из направлений. Продуктовой команде нужно было больше времени, чтобы отшлифовать интерфейс; команде инфраструктуры – распределить серверные мощности; а саму модель надо было ещё дорабатывать, чтобы выправить её поведение.

К тому времени OpenAI создала вместе с Microsoft комитет – Совет по безопасности развёртывания (Deployment Safety Board, DSB), по три представителя от каждой компании, – чтобы оценивать передовые модели OpenAI и решать, готовы ли они к выпуску. От OpenAI в него вошли Альтман, Майлз Брандедж, руководитель исследований политики, и Ян Лейке, руководитель направления согласования (alignment), отвечавшего за дальнейшее развитие методов безопасности ИИ вроде RLHF. И исследования политики, и согласование стали новыми опорными пунктами клана безопасности. DSB создал формальную структуру управления для разрешения извечных споров между Applied и кланом безопасности. После предварительного рассмотрения DSB дал GPT-4, первой модели, оцениваемой в рамках этой

структуры, условное одобрение: модель можно будет выпустить после основательного тестирования и дальнейшей доводки с точки зрения безопасности ИИ.

Руководители договорились о новом сроке – начале 2023 года. К этому времени они хотели подготовить и Superassistant. OpenAI выпустит GPT-4 в API одновременно с потребительским продуктом на основе GPT-4. Сотрудникам Альтман представлял решение отложить выпуск модели как свидетельство осторожного, ориентированного на безопасность подхода OpenAI. Компания не торопится с развёртыванием системы, говорил он. На самом деле руководство даже не уверено, что вообще будет её развёртывать. В Applied большинство поняли Альтмана так: чтобы руководство остановило выпуск, нужна какая-то чрезвычайная проблема, но если такое случится, оно будет к этому готово. Многие в клане безопасности восприняли его слова иначе: предполагалось, что руководство даст зелёный свет только после того, как GPT-4 пройдёт все проверки.

На закрытых встречах Альтман укреплял каждую сторону в её представлениях. И тем и другим он говорил о тревоге, что GPT-4 может «разбудить Google», рисуя сценарий, в котором в тот день, когда выйдет модель и привлечёт внимание Ларри Пейджа и Сергея Брина, те наконец прорвутся сквозь политическую бюрократию Google. Руководство Google может отменить все совещания, уехать на однодневный выездной сбор, собрать в кулак все таланты и ресурсы компании и бросить всю её вычислительную мощь на то, чтобы догнать соперников. Но если для Applied это было поводом и дальше в бешеном темпе готовиться к запуску, как можно дольше держа возможности GPT-4 в секрете, то перед кланом безопасности Альтман использовал этот довод, чтобы и дальше подчёркивать свою осторожность. «Моя главная забота в плане безопасности – риск ускорения», – говорил он, перенимая их словарь. В переводе: если разбудить Google, это может запустить гонку на дно.

* * *

Даже с отсрочкой подразделение Applied с трудом успевало к выпуску в начале 2023 года. В особенности у команды доверия и безопасности Дэйва Уиллнера всё ещё было мало людей и неразвитая инфраструктура. Если OpenAI собиралась запустить потребительский продукт Superassistant на основе куда более мощной модели, ей нужна была гораздо более зрелая система предотвращения злоупотреблений и принуждения к соблюдению правил.

Уиллнер спешно нанял нескольких опытных заместителей по доверию и безопасности из Google и Meta. Одному он поручил расследовать «неизвестные неизвестные» – злоупотребления, о которых команда ещё не знала и которые предстояло обнаружить с помощью тщательного мониторинга и анализа данных. Другому – «известные известные», или так называемое масштабное принуждение к соблюдению правил: злоупотребления, которые компания уже

признала нарушениями и которые нужно было автоматически помечать, проверять и пресекать, например блокируя учётные записи злостных нарушителей.

Наперегонки выстраивая продуктовые правила и инфраструктуру мониторинга и пресечения нарушений, члены молодой команды постоянно не понимали, чем доверие и безопасность в ИИ-компании должны отличаться от того же в поисковой компании или соцсети и достаточно ли их подготовки к GPT-4. Доверие и безопасность обычно сосредоточены на предотвращении предсказуемого набора интернет-злоупотреблений вроде мошенничества, киберпреступлений и вмешательства в выборы. Но в эту путаницу вплетался и вопрос о том, как их работа соотносится с работой людей из безопасности ИИ в командах Лейке по согласованию и Брандеджа по исследованиям политики, которые часто рассуждали о непознаваемой катастрофической гибели. OpenAI называла все эти команды командами «безопасности», но они, казалось, говорили на принципиально разных языках. Хотя словарь экзистенциальной безопасности ИИ, распространённый благодаря ЭА, вошёл в обиход в этой области, сотрудники, пришедшие из обычных технологических компаний, никогда не слышали о таймлайнах ИИ и не оценивали свой p(doom). Даже общие для обеих групп слова вроде «риска» и «вреда» как будто несли разные оттенки смысла.

По мере того как в OpenAI приходило всё больше людей не из мира ИИ, чтобы обслуживать растущие нужды подразделения Applied, умение преодолевать культурный разрыв между теми, кто пропитан думерско-бумерским мышлением, и теми, кто действует в рамках обычной технологической компании, превратилось в особую политическую валюту. Из заместителей Уиллнера по доверию и безопасности тот, кто лучше всех освоил язык безопасности ИИ, стал мостом в этот мир. Этот заместитель работал с разными командами, чтобы сделать модель «безопаснее» в обоих смыслах слова, – например, с помощью RLHF «согласовывал» модель (жаргон безопасности ИИ, связанный с вышедшим из-под контроля ИИ) и тем самым учил её лучше отказывать в запросах пользователей, нарушающих правила платформы OpenAI, – типичная работа доверия и безопасности.

Но пока «безопасность» продвигалась на обоих фронтах, от руководства вдруг пришло новое распоряжение: приостановить проверку разработчиков, впервые введённую при выпуске API GPT-3.

Руководство уже давно считало, что лист ожидания разросся до неуправляемых размеров, а проверка не масштабируется. Разработчики жаловались, как долго приходится ждать доступа к технологиям OpenAI, и постоянно писали Альтману, Брокману и Питеру Велиндеру или отмечали их в Twitter, возмущаясь долгими задержками с приложениями, которые руководство считало совершенно безобидными. Многие в Applied к тому же

считали, что OpenAI перекрывает доступ к пользе от своей технологии, а это противоречит её миссии.

Уиллнер и его команда всегда возражали. Если OpenAI откажется от рассмотрения заявок и будет одобрять разработчиков автоматически, у компании не будет реальной альтернативы для контроля над использованием её технологий. Если перейти к реагированию на нарушения постфактум, понадобится создать для этого серьёзный инструментарий. Но с приближением запуска GPT-4 руководство отмело возражения: OpenAI отказывается от проверки разработчиков, а команде доверия и безопасности просто нужно придумать альтернативу.

Команда Уиллнера спешно составила предложение. Она переместит значительную часть работы по соблюдению правил компании на более ранний этап, сильнее опираясь на RLHF при согласовании GPT-4 и будущих моделей. Всё остальное будут вылавливать и решать на более позднем этапе, реагируя на нарушения постфактум: по разным сигналам в данных – сведениям о том, что делает приложение, характерным всплескам трафика, числу срабатываний фильтров модерации контента – автоматически блокировать явных нарушителей, а пограничные случаи отправлять живым модераторам для ручной проверки.

Команда Уиллнера добивалась от руководства ресурсов, чтобы как следует создать инструменты, без которых план не мог работать. Большинства этих данных у OpenAI не было: её платформа мониторинга записывала лишь простейшие сведения о том, сколько трафика каждое приложение пропускает через серверы компании. Порой компания не знала даже имени разработчика или назначения приложения.

Помимо того чтобы закрыть эти пробелы, команде нужно было, чтобы разработчики присваивали каждому своему пользователю уникальный идентификатор. Это было необходимо, чтобы разбить трафик приложения по отдельным пользователям и понять, свойственны ли нарушения всей базе пользователей приложения или их совершает лишь горстка частых нарушителей. Руководство воспротивилось, опасаясь, что такой подробный мониторинг усложнит жизнь разработчикам и, возможно, возложит на компанию больше ответственности за работу по доверию и безопасности для каждого приложения на её технологиях, вместо того чтобы этим занимался каждый разработчик сам.

В итоге предложение о реагировании постфактум урезали до чего-то гораздо более скромного.

Ограниченное представление о поведении пользователей приложений тревожило некоторых членов команды доверия и безопасности. Один сотрудник поделился своими опасениями с Брокманом. Больше всего, сказал он, он боится, что люди станут с помощью GPT-4 массово генерировать недостоверную информацию и дезинформацию и влиять на выборы.

Брокман постарался его успокоить.

– Да, мы всегда это говорим и всегда этого опасаемся, – сказал он. – Но я ни разу не видел: а это действительно происходит?

* * *

GPT-4 стала поворотным моментом не только для Гейтса и Microsoft. Позже тем же летом, поразив миллиардера-филантропа, Альтман и Брокман представили модель совету директоров OpenAI, где Брокман провёл живую демонстрацию, в ходе которой модель, среди прочего, рассказывала анекдоты о Гэри Маркусе. Анекдоты привели независимых членов совета в восторг; к тому же эта шутливая демонстрация показала им значительный скачок в возможностях технологии и дала ощущение, что ставки их будущих решений растут.

До того Альтман созывал совет примерно раз в квартал и предпочитал отчитываться устно. Он бегло проходился по сложным исследовательским темам, иногда приводя с собой исследователей компании, чтобы те рассказали о своих успехах, и скороговоркой излагал последние сделки, которые обсуждал с Microsoft или другими партнёрами. Некоторые независимые члены совета добивались более частых заседаний и более структурированной информации, настаивая, чтобы Альтман представлял письменные отчёты и давал им доступ к большему числу документов.

Альтмана раздражал усилившийся надзор. Он не раз давал понять, что хотел бы, чтобы члены совета были скорее советниками, высказывающими ему мнения для размышления. «CEO должен принимать решения, а совет должен, ну, знаете, быть тем, с кем можно посоветоваться, – „совет и согласие“», – сказал он однажды студентам в 2017 году, по-видимому ссылаясь на принцип управления из Конституции США, лежащий в основе роли законодательной власти как сдержки исполнительной, но употребляя его в противоположном смысле. «Считаю ли я, что совет должен уволить плохого CEO? Да – и я знаю, что для Кремниевой долины это немного еретично. А помимо этого, считаю ли я, что совет должен давать CEO очень широкую свободу в управлении компанией? Да».

Несколько членов совета были твёрдо убеждены, что совет OpenAI задуман иным. «Это совет некоммерческой организации, созданный именно для того, чтобы миссия компании ради общественного блага была главной, стояла выше прибыли, интересов инвесторов и прочего, – скажет Хелен Тонер в подкасте The TED AI Show. – А не просто, знаете, помогать CEO привлекать больше денег».

У многих сотрудников GPT-4 укрепила веру в то, что AGI возможен. Исследователи, когда-то настроенные скептически, всё увереннее говорили о достижимости такой технической вершины – хотя у OpenAI по-прежнему не было определения того, что это такое. Инженеры и продакт-менеджеры, приходившие в Applied и впервые близко знакомившиеся с ИИ через GPT-4,

говорили о нём ещё более детерминистски. Для многих сотрудников вопрос стоял уже не «появится ли AGI», а «когда».

Некоторые сотрудники, впрочем, считали ровно наоборот. Если между тем, что умела GPT-3, и тем, что умела GPT-2, была явная качественная разница, то GPT-4 была просто больше, говорит один из исследователей, работавших над моделью. «Главным результатом было то, что модель хорошо сдаёт кучу экзаменов. Но даже это весьма сомнительно». OpenAI так и не провела всесторонней проверки обучающих данных GPT-4, чтобы выяснить, не было ли этих экзаменов – и ответов на них – просто в данных, так что модель их лишь воспроизводила, или GPT-4 действительно обрела новую способность их сдавать. Это была та самая шаткая наука, что распространилась повсюду, когда вся отрасль перешла от исследований, проверенных рецензентами, к исследованиям, проверенным пиарщиками.

* * *

Но вера в то, что ИИ достиг принципиально новых высот, витала в воздухе. Той весной в Google Блейк Лемуан, инженер недавно воссозданной команды ответственного ИИ техногиганта, пришёл к убеждению, что собственная большая языковая модель компании, LaMDA, не только чрезвычайно умна, но и может считаться разумным существом. По его словам, это основывалось не на научной оценке, а на его вере как христианского священника-мистика в то, что Бог может решить наделить технологию сознанием. «Кто я такой, чтобы указывать Богу, куда ему вкладывать души, а куда нет?» – писал он. Когда руководство компании отмахнулось от него, он обратился в The Washington Post. Ниташа Тику, журналистка, опубликовавшая эту историю, поговорила и с Эмили Бендер и Мэг Митчелл, которые в статье о «стохастических попугаях» предупреждали, что большие языковые модели могут обманывать людей, заставляя их видеть в сгенерированном тексте подлинный смысл и намерение.

«Теперь у нас есть машины, способные бездумно порождать слова, но мы так и не научились перестать воображать за ними разум», – сказала Бендер.

«Меня по-настоящему беспокоит, что будет означать для людей всё большее влияние этой иллюзии», когда она становится такой убедительной, сказала Митчелл.

Никхил Мишра, исследователь ИИ, проходивший стажировку в OpenAI в первые её дни, проводит параллель с экспериментом 1970-х годов, в котором гориллу пытались научить видоизменённой форме американского жестового языка. За свою жизнь горилла по кличке Коко, казалось, выучила больше тысячи знаков и даже научилась составлять предложения. Но, несмотря на огромный ажиотаж в обществе, другие специалисты утверждали, что по-настоящему языком Коко так и не овладела. Жесты она, безусловно, складывала умело, но почти ничто не указывало на то, что она делает больше, чем другие обезьяны во множестве похожих экспериментов, – просто повторяет жесты своих

смотрителей. Данные о жестах Коко так и не были опубликованы для независимой проверки её способностей – только отобранные видео, на которых она их демонстрирует. Порой её выступления вживую требовали более пристального взгляда. Однажды, когда дрессировщица спросила Коко, нравятся ли ей люди, Коко показала жестами «хороший сосок». Дрессировщица тут же пояснила: по-английски «сосок» (nipple) рифмуется с «люди» (people) – Коко считает, что люди хорошие. Для многих наблюдателей такие эпизоды говорили больше о человеческой психологии и нашей склонности приписывать другим собственные убеждения и представления о намерениях, чем о способностях Коко. Дрессировщица, говорит Мишра, видела смысл там, где его не было.

Для собственного мистика OpenAI, Суцкевера, который всегда сильнее большинства исследователей верил в то, что AGI будет достигнут в ближайшее время, скачок в возможностях моделей компании стал лишь очередным подтверждением. Он уверился, что наблюдает некую форму рассуждения. В разговорах с Хинтоном Суцкевер говорил своему наставнику, что AGI вот-вот появится.

Если раньше Суцкевер больше подталкивал исследователей OpenAI к созданию новых возможностей, то теперь он с новой настойчивостью переключился на исследования безопасности ИИ. Он завёл мантру «Почувствуйте AGI» и призывал людей готовиться к резким переменам. «Скоро вы внезапно станете самым популярным человеком на любой вечеринке, – говорил он сотрудникам. – Не позволяйте этому вскружить вам голову. Сосредоточьтесь на AGI, сосредоточьтесь на миссии».

В сентябре техническое руководство устроило выездную встречу в Tenaya Lodge – уединённом роскошном курорте, укрывшемся в пышных складках Сьерра-Невады. Отель с потрясающими интерьерами, несколькими бассейнами и ресторанами стоял всего в двух милях от миллионов акров нетронутой дикой природы Йосемитского национального парка. В первый вечер все собрались вокруг костровой чаши на заднем дворе отеля. Ведущие учёные в банных халатах полукругом расположились у огня.

И тут появился Суцкевер. Он положил в костровую чашу деревянное изваяние, заказанное у местного мастера, и начал театральное представление. Это изваяние, объяснил он, олицетворяет хороший, согласованный AGI, который создала OpenAI, лишь чтобы обнаружить, что он на самом деле лжив и коварен. Долг OpenAI, сказал он, – уничтожить его. Всего в нескольких метрах, словно древние свидетели, стояли в темноте секвойи. Суцкевер облил изваяние жидкостью для розжига и поджёг.

Глава 11

Вершина

В октябре 2022 года OpenAI провела общую выездную встречу компании в Монтерее – красивом прибрежном городе в двух часах езды к югу от Сан-Франциско. К тому времени в компании работало около трёхсот человек, а годом раньше – меньше двухсот. Для фотографии в те выходные все позировали у крошечного аэропорта Монтерея – улыбающиеся, в фирменной одежде с логотипом OpenAI. Альтман выглядел расслабленным: он сидел на земле перед всеми, подтянув колени к груди, небрежно скрестив руки и задрав носки.

Два дня руководство рассказывало о своём видении и о том, как оно воплощается: Альтман – об огромных центрах обработки данных, которые Microsoft наращивала для OpenAI; Стив Даулинг и его заместитель Ханна Вонг – о первоклассных изданиях, выстроившихся в очередь, чтобы писать о компании; Анна Маканджу – о растущем присутствии компании в Вашингтоне. Потом шли демонстрации, одна за другой, новейших проектов исследовательских и продуктовых команд. Впечатление от всего этого было почти осязаемым. «Всё вместе просто ошеломляло», – вспоминает присутствовавший там бывший сотрудник.

Брокман вышел на сцену, чтобы рассказать о последних планах насчёт GPT-4, и начал историю о своей жене Анне, у которой однажды начались боли в животе, никак не проходившие. Анна часто бывала в офисе. Её стол стоял рядом с его столом, и она нередко ходила с ним на совещания, хотя официально в компании не работала. Не считая Альтмана и Суцкевера, она, как однажды сказал мне Грег, была тем человеком, на чью поддержку он опирался больше всего, – лучший друг и наперсница. Они часто были неразлучны и всюду ходили вместе. Грег ходил с Анной к множеству врачей, объяснял он сотрудникам, и никто не мог понять, в чём дело. Но когда он спросил GPT-4, модель предположила у неё заболевание, о котором они не подумали. «И оказалось, что так и есть!» – воскликнул он. Это был пересказ той же истории, которой он поделился со мной в 2019 году, – о том, как AGI обещает решить проблемы здравоохранения для людей вроде его друга, обошедшего бесчисленных специалистов, чтобы поставить диагноз, – только на этот раз с другим героем.

Грег снова повторит эту историю в X через несколько недель после попытки совета уволить Альтмана, в новом варианте. Он расскажет о новых медицинских испытаниях, выпавших на долю Анны, и о том, как она пять лет «ходила к большему числу врачей и специалистов, чем за всю предыдущую жизнь», пока наконец не получила диагноз. Всё сложил воедино её аллерголог, понявший, что у неё гипермобильный синдром Элерса – Данло, генетическое нарушение подвижности суставов. AGI ещё предстоит долгий путь, прежде чем он будет

работать «в областях с высокими ставками, таких как медицина», писал он, продолжая бить в барабан во имя объединяющей OpenAI амбиции, «но обещание становится всё яснее».

* * *

Через несколько недель после возвращения с выездной встречи поползли слухи, что Anthropic тестирует – и скоро выпустит – новый чат-бот. Команда Superassistant была на середине работы над интерфейсом чата. Если не запуститься первыми, OpenAI рисковала потерять лидерство, а это могло сильно ударить по боевому духу сотрудников, которые долгие и тяжёлые часы работали, чтобы это лидерство сохранить. Для некоторых руководителей хуже всего было то, что OpenAI проиграет именно Anthropic.

На самом деле Anthropic не планировала никаких скорых выпусков. К тому же у неё были свои проблемы. После внезапного краха FTX в первые дни ноября её затягивало в водоворот последствий. Всего за несколько месяцев до этого она привлекла 580 миллионов долларов, из которых, как говорилось в пресс-релизе FTX, 500 миллионов вложили SBF и другие высшие руководители. Финансовые документы, обнародованные на суде, позже покажут, что эти деньги были взяты из миллиардов, которые SBF присвоил из клиентских депозитов FTX; инвестиция в Anthropic стала одним из центральных вопросов процесса: можно ли направить полученную от неё значительную прибыль на возврат денег клиентам. (Судья постановит, что можно, и FTX в течение 2024 года распродаст свои акции Anthropic частями в общей сложности за 1,3 миллиарда долларов.)

Но для руководства OpenAI слухов оказалось достаточно, чтобы принять решение: компания не будет ждать, пока из GPT-4 сделают чат-бота, а через две недели, сразу после Дня благодарения, выпустит разговорную модель GPT-3.5 Шульмана с совершенно новым интерфейсом чата от команды Superassistant. Команда Superassistant мгновенно перестроилась, подтянула ещё нескольких людей и рванула, чтобы всё объединить и доделать оставшиеся функции. Остальной компании руководство подало эту работу осторожно. ChatGPT – на этом названии в итоге остановились – будет не запуском продукта, а «скромным исследовательским предварительным показом», точно как DALL-E 2. Точно так же его не станут монетизировать, а будут использовать, чтобы «раскрутить маховик данных», – иначе говоря, накопить больше данных от пользователей, что поможет улучшить GPT-4 и продукт Superassistant.

За пределами команды Superassistant все поняли руководство буквально. Скромный исследовательский показ не требовал их внимания; им нужно было сосредоточиться на запуске GPT-4 в начале 2023 года. Команде доверия и безопасности казалось, что у неё и так едва хватает времени, чтобы к этому запуску выстроить инфраструктуру мониторинга. По сравнению с этим ChatGPT казался пустяком. GPT-3.5 уже доработали с помощью RLHF; она была заведомо безопаснее версии GPT-3 без такой доработки, которая всё ещё была доступна в

API. К тому же OpenAI выложила версию 3.5 без функций чата на своей платформе для разработчиков, чтобы те могли опробовать возможности модели. Неужели добавление интерфейса чата что-то меняет? Люди из клана безопасности, занятые тестированием и доводкой GPT-4, были согласны. Впервые выпуск модели пролетел через все проверки почти без сопротивления.

Даже в команде Superassistant никто по-настоящему не представлял, какой фазовый переход в обществе они вот-вот запустят. Они ждали, что чат-бот окажется однодневкой. Как и DALL-E 2, он наделает шума в соцсетях, а через несколько недель всё уляжется. Накануне выпуска, после такого напряжённого рывка к финишу, было удивительно спокойно. Они заключали пари, сколько пользователей попробует инструмент к концу выходных. Одни называли несколько тысяч. Другие – десятки тысяч. На всякий случай команда инфраструктуры выделила серверные мощности на сто тысяч пользователей.

На следующий день, в среду, 30 ноября, большинство остальных сотрудников даже не заметили, что запуск состоялся. Как и дебют самой OpenAI, выход ChatGPT совпал с ежегодной конференцией NeurIPS, проходившей в тот год в Новом Орлеане, штат Луизиана. Конференция заранее объявила лауреата премии «Проверка временем» (Test of Time Award) – награды, ежегодно присуждаемой статье, опубликованной десятью годами раньше и оказавшей решающее влияние на эту область. Премию получила статья Хинтона, Суцкевера и Крижевского об ImageNet 2012 года, показавшая миру силу глубокого обучения.

В тот вечер небольшая группа сотрудников устроила вечеринку OpenAI неподалёку от конференц-центра, чтобы представить компанию и привлечь интересных кандидатов из почти десяти тысяч очных участников. DeepMind, Meta и Google в то же самое время устраивали по всему городу конкурирующие вечеринки для найма. В разгар вечеринки рекрутер заметил инженера OpenAI, который без остановки работал за компьютером.

В конце концов он подошёл к инженеру:

– Брат, выпей. Мы все здесь. Пообщайся с людьми.

Инженер даже не пошевелился:

– Нет, все GPU плавятся. Всё падает.

* * *

В ту ночь первой проснулась Япония и обрушила на сервис огромный и неожиданный вал трафика. На следующий день число пользователей продолжало расти: часовой пояс за часовым поясом в сеть выходил остальной мир. Немалую роль в этом взлёте сыграл Маск. «Куча народу застряла в петле ChatGPT „чёрт, вот это безумие“», – написал он в Twitter и собрал около семидесяти пяти тысяч лайков.

Мгновенный, неудержимый успех ChatGPT превзошёл всё, о чём кто-либо в OpenAI мог мечтать. Инженеры и исследователи компании недоумевали и годы

спустя. GPT-3.5 была не таким уж большим улучшением возможностей по сравнению с GPT-3, которая к тому времени была доступна уже два года. А GPT-3.5 и так уже была доступна разработчикам. Интерфейс и формат, конечно, сделали модель доступнее, но это был не тот принципиальный скачок, который сотрудники ощутили с GPT-4. Альтман позже скажет, что ожидал популярности ChatGPT, но примерно «на порядок меньшей». «Было шоком, что людям понравилось, – вспоминает бывший сотрудник. – Для всех нас это выглядело так: они урезали то, чем мы пользовались внутри компании, и выпустили это».

Через пять дней Брокман написал в Twitter, что у ChatGPT больше миллиона пользователей. Через два месяца их было сто миллионов, и он стал на тот момент самым быстрорастущим потребительским приложением в истории. (Threads, конкурент X от Meta, позже заберёт этот титул, набрав столько же пользователей меньше чем за пять дней; эксперты возражали, что это не в счёт, поскольку Meta в основном опиралась на уже существующую базу пользователей.)

ChatGPT в одночасье превратил OpenAI из модного стартапа, хорошо известного в технологической отрасли, в имя, которое знает каждый. Несколько месяцев спустя на конференции по исследованиям ИИ в Кигали, в Руанде, больше чем в девяти тысячах миль от Сан-Франциско, работающий там исследователь восторженно рассказывал мне, что после ChatGPT его родители наконец поняли, чем он занимается. «Технология стала доступна каждому, если о ней тебе рассказывает мама», – сказал он.

В то же время именно этот оглушительный успех лёг на компанию чрезвычайным грузом. За год он ещё сильнее поляризует её группировки и доведёт стресс и напряжение внутри организации до взрывоопасного уровня.

Сразу после запуска вся компания тушила пожары. Серверы OpenAI раз за разом падали, а команда инфраструктуры изо всех сил наращивала мощности в самые сжатые сроки за всю историю Кремниевой долины. Команда забрала часть вычислительных ресурсов подразделения Research, чтобы поддержать рост ChatGPT, и всё равно не могла удержать приложение на плаву. Команда доверия и безопасности, насчитывавшая чуть больше десятка человек, лихорадочно пыталась понять и отловить недопустимое поведение среди потоков новых пользователей, сильно скованная обрывочным мониторингом. Ей с трудом удавалось нанять инженеров для своего урезанного плана реагирования на нарушения постфактум, и она всё ещё была в разгаре построения нужных систем. ChatGPT пустил проект под откос. Вся работа над новыми инструментами остановилась: инженерные ресурсы бросили на то, чтобы стабилизировать уже существующее. Когда падали серверы, падала и платформа мониторинга трафика, и любое масштабное пресечение нарушений полностью замирало.

Острая нехватка GPU сорвала и ещё одну затею. Пытаясь задействовать собственную технологию компании, команда доверия и безопасности подготовила прототип плана, который внутри называли Fact Factory и которым

OpenAI публично хвалилась: использовать GPT-4 для модерации собственных ответов и ответов других моделей OpenAI. Масштабироваться это не очень-то получалось: чтобы уловить достаточно нюансов, GPT-4 нужно было давать чрезвычайно длинные запросы. Даже когда серверы работали, это обходилось бы в слишком большие вычислительные ресурсы. А серверы работали не всегда.

Для многих в клане безопасности ChatGPT стал самым тревожным на тот момент примером ограниченности предвидения OpenAI. Один человек из клана безопасности задал вопрос на общем собрании: как компания могла так сильно ошибиться в прогнозах поведения пользователей и популярности ChatGPT? Что это говорит о её способности оценивать и прогнозировать будущее воздействие своих технологий?

Для большей части остальной компании падающие серверы, при всём колоссальном стрессе, были в ещё большей степени колоссальным знаком триумфа. OpenAI создала технологию настолько глубокую и настолько неистово востребованную, что та озарила мир и в одночасье его преобразила. Они нацелились на всё человечество – и действительно своего добились. Все, все восемь миллиардов человек, теперь жили в мире OpenAI.

Альтман не позволил себе упиваться моментом. Он напомнил сотрудникам, что в конечном счёте у компании миссия гораздо больше, чем создание «самого большого продукта в истории Кремниевой долины». Он призвал все команды держать курс и двигаться вперёд. Как он и ожидал, OpenAI разбудила всех конкурентов: Anthropic готовилась выпустить своего чат-бота Claude; в Google объявили внутреннюю тревогу «красный код» и вскоре объединят свои ИИ-подразделения в Google DeepMind, чтобы бросить все силы на запуск похожего продукта. OpenAI первой вышла на рынок с предложением, которое в десять раз лучше, но, чтобы оставаться номером один, ей нужно было продолжать бежать.

* * *

Все команды были растянуты до опасного предела, и руководители умоляли Альтмана о новых ставках. В кандидатах недостатка не было. После ChatGPT число соискателей, рвавшихся на борт этой ракеты, стремительно умножилось. Но Альтман тревожился, что станет с культурой компании и приверженностью миссии, если штат будет расти слишком быстро. Он твёрдо верил, что штат должен быть небольшим, а плотность талантов высокой. «Сейчас мы в таком положении, когда велик соблазн позволить организации разрастись до огромных размеров, – писал он в записке о своём видении 2020 года, имея в виду инвестиции Microsoft. – Мы должны изо всех сил этому противиться: до сих пор нам помогало то, что мы небольшие, сосредоточенные, с высоким доверием, минимумом чуши и большим напором.

«Накладные расходы от слишком большого числа людей и слишком большой бюрократии легко могут убить великие идеи или привести к склерозу.

В отличие от тяжеловесных инженерных проектов прошлого, мы могли бы выполнить нашу миссию с удивительно малым числом выдающихся людей».

Теперь, в конце 2022 года, он повторял это руководителям, снова и снова подчёркивая в обсуждениях штатного расписания, что компания должна оставаться компактной, планка для талантов – высокой, а нанимать стоит не больше сотни человек или около того. Другие руководители упирались. При той скорости, с которой выгорали их команды, многие считали, что нужно что-то ближе к пятистам новым сотрудникам, а то и больше.

После нескольких недель обсуждений руководство наконец сошлось на компромиссном числе где-то посередине – от двухсот пятидесяти до трёхсот. Лимит не удержался. К лету в OpenAI каждую неделю приходило по тридцать, а то и по пятьдесят человек, в том числе новые рекрутеры, чтобы нанимать ещё быстрее. К осени компания далеко вышла за рамки самой себе установленной квоты.

Внезапный скачок роста действительно изменил культуру компании. Один рекрутер написал манифест о том, что давление требует нанимать так быстро, что его команде приходится снижать планку качества. «Если вы хотите построить Meta, у вас отлично получается», – язвительно уколол он Альтмана, намекая на те самые опасения, о которых предупреждал CEO: компания быстро размывает плотность талантов и верность миссии, наращивая бюрократию. Стремительное расширение вело и к росту увольнений. Одному руководителю при вводе в должность велели быстро документировать и докладывать о каждом слабом сотруднике в его команде – а через какое-то время уволили и его самого. Об увольнениях остальной компании сообщали редко. Люди обычно узнавали, что коллег уволили, только заметив, что учётная запись в Slack посерела после деактивации. Это стали называть «быть исчезнутым».

Новичкам, всецело уверенным, что они приходят в быстро растущий, зарабатывающий деньги стартап, эта неразбериха казалась особенно хаотичным, порой жестоким проявлением обычных корпоративных проблем: плохого управления, путаных приоритетов, бессердечной безжалостности капиталистической компании, готовой обращаться с сотрудниками как с расходным материалом. «Психологической безопасности катастрофически не хватало, – говорит бывший сотрудник, пришедший в те времена. – Это полная противоположность „компании как семьи“ – что справедливо, знаете, это ведь компания». Многие новички просто держались из последних сил до годовщины, чтобы получить доступ к первой доле своих акций. Одним большим плюсом было то, что коллеги, по их ощущениям, по-прежнему были одними из сильнейших в технологической отрасли, и вместе с, казалось бы, безграничными ресурсами и непревзойдённым глобальным влиянием это могло, когда всё действительно складывалось, рождать ощущение волшебства, которое трудно найти в остальной отрасли. «Я бы сказал, что OpenAI – одно из лучших мест, где

я когда-либо работал, но, наверное, и одно из худших», – говорит бывший сотрудник.

Для некоторых сотрудников, помнивших задорные ранние дни OpenAI как сплочённой некоммерческой организации, движимой миссией, её резкое превращение в большую безликую корпорацию было куда более шокирующим и болезненным. Организации, какой они её знали, больше не было; на её месте возникло что-то неузнаваемое. «OpenAI – это Burning Man», – говорит бывший рекрутер Роб Мэллери, имея в виду, как пустынный фестиваль искусства разросся до того, что утратил связь со своим изначальным духом. «Я знаю, что для тех, кто был там с самого начала, это значило гораздо больше, чем значит для всех сейчас».

В те ранние годы команда завела в Slack канал #explainlikeimfive («объясни, как пятилетнему»), куда сотрудники могли анонимно задавать вопросы на технические темы. Когда в компании стало под шестьсот человек, канал превратился и в место для анонимных жалоб. В середине 2023 года один сотрудник написал, что компания нанимает слишком много людей, не преданных миссии и не горящих созданием AGI.

Кто-то ответил: он понял, что OpenAI катится под гору, когда она начала нанимать людей, способных смотреть тебе в глаза.

* * *

ChatGPT удивил и Microsoft. Руководство OpenAI говорило партнёру, как и своим сотрудникам, что чат-бот будет «скромным исследовательским предварительным показом». Он, очевидно, был чем угодно, только не этим.

Поначалу это несоответствие раздосадовало руководство техногиганта. ChatGPT полностью перехватил славу у чат-бота Microsoft для Bing. Когда в феврале следующего года Microsoft выпустила Bing AI, продукт к тому же получил удар по репутации: обозреватель New York Times Кевин Руз написал статью о том, как бот подталкивал его развестись с женой. Это был совсем не тот приём, на который рассчитывала Microsoft, и на этом фоне OpenAI выглядела ещё лучше.

Но этой путаницы было далеко не достаточно, чтобы охладить энтузиазм Microsoft по отношению к OpenAI. Огромный успех ChatGPT был заразителен, а постоянно растущие возможности моделей OpenAI ещё сильнее воодушевляли руководство гиганта. Microsoft теперь готовила целую линейку «Копилотов» для своих продуктов на основе GPT-3.5 и GPT-4 и собиралась выпускать их один за другим под мерную барабанную дробь анонсов. После Bing в феврале март отводился Microsoft 365 Copilot, приносившему разговорный интерфейс на основе ИИ во все продукты Office – от Word до Outlook и Teams.

Быстро менялось и то, как руководители Microsoft говорили о партнёрстве с OpenAI внутри компании. Раньше Microsoft ощущала, что у неё есть власть над OpenAI; теперь некоторым руководителям гиганта казалось, что это у OpenAI

власть над Microsoft. В отдельных частях Microsoft нарастало ощущение собственной несостоятельности: её собственные исследования ИИ не смогли добиться того, что удалось OpenAI. Если Microsoft уйдёт с поста главного инвестора OpenAI, стартап найдёт других инвесторов, вспоминает бывший сотрудник Microsoft ход мыслей некоторых руководителей. Но если OpenAI уйдёт от Microsoft, найдёт ли техногигант другую OpenAI?

Вместе с тем многие руководители говорили уже не просто о том, чтобы обойти Google. Если когда-то они встречали странные речи OpenAI об AGI и высокопарные слова о его силе изобрести будущее вежливым скепсисом, то теперь они уверовали. ИИ, AGI, генеративный ИИ – как ни назови, эта технология и есть будущее, и Microsoft ведёт его рука об руку с OpenAI. Чем сильнее верила Microsoft, тем больше перестраивала свою риторику и стратегию. «Я видел, как стимулы в Microsoft всё больше подталкивают к узкому представлению о будущем, – говорит бывший сотрудник. – Я видел, как технология сужается до чего-то, что держится на нарративе, а не на реальности».

Наделла ввёл новую стратегию распределения вычислительных ресурсов Microsoft. Он перебросил GPU от исследовательских команд Microsoft на поддержку OpenAI. Компания также объединила все свои GPU в общий пул, чтобы лучше обслуживать нагрузки генеративного ИИ. «Типичный сотрудник Microsoft до января прошлого года ни хрена не знал, что такое OpenAI», – вспоминает один сотрудник Microsoft. Теперь им приходили срочные распоряжения начальства – искать способы увязать свою работу с технологиями OpenAI.

Всего за несколько месяцев у техногиганта стремительно расплодилось больше сотни новых проектов генеративного ИИ: сотрудники пробовали разные способы применять GPT-4 и ChatGPT. По иронии судьбы, такое агрессивное внедрение заставило Microsoft столкнуться со многими из тех же трудностей, с которыми столкнутся другие компании, наперегонки внедряя генеративный ИИ, до конца его не понимая. В том числе с головной болью для команд по рискам и соблюдению норм. Не все пользовались внутренними версиями технологий Microsoft: некоторые выбирали бесплатную версию ChatGPT прямо от OpenAI, которая обучалась на данных пользователей, а это вызывало опасения, не может ли через неё утечь информация клиентов Microsoft и не помешает ли это соблюдению регуляторных требований. Одним сотрудникам инструменты сильно повышали продуктивность, но многих они ещё и изматывали. «Тут такая культура: „Пользуйся ИИ, пользуйся ИИ, пользуйся ИИ“», – говорит один из них. Но «хочется сказать: слушайте, нам это не помогает. Мы не хотим этим пользоваться. А ощущение, что оно повсюду и от него никуда не деться».

Переходя с собственных моделей Microsoft на модели OpenAI, многие сотрудники к тому же теряли контроль над ключевым инфраструктурным слоем своей работы и возможность в него заглянуть. Внутри техногиганта доступ к базовым моделям стартапа строго охранялся, хотя обучались и хранились они на

серверах Microsoft. Большинство сотрудников Microsoft больше не могли изучить обучающие данные или подправить веса моделей, которыми пользовались. Модели OpenAI поставлялись им через API, как и другим клиентам OpenAI.

Но в обмен на эти уступки Microsoft получала щедрую награду. Компания наблюдала огромный наплыв клиентов на свою платформу Azure AI – единственного облачного провайдера, способного предложить обычные преимущества облака, в том числе более простое хранение данных и управление ими, вместе с возможностью обрабатывать эти данные с помощью возможностей OpenAI. «Azure OpenAI Service сейчас открывает нам двери ко множеству новых клиентов», – писал в мае 2023 года своему подразделению Эрик Бойд, корпоративный вице-президент по платформе ИИ. В августе он снова ликовал: «Время от времени приятно сделать шаг назад и подивиться тому, как далеко мы продвинулись всего за год», – писал он подразделению, добавляя, что за этот год число клиентов платформы «выросло в 21 раз». В следующем месяце Бойд праздновал новую веху. После того как разрозненные ИИ-проекты Microsoft свели на эту платформу, а к Azure OpenAI Service присоединились тысячи новых клиентов, трафик на платформе всего за девять месяцев вырос в десять раз. В январе 2023 года она получала сто миллиардов запросов на вывод в месяц; теперь, в сентябре, – триллион.

Тем летом технический директор Microsoft Кевин Скотт не скупился на похвалы на общем собрании OpenAI. «Мы, типа, остановили наши инвестиции в ИИ и машинное обучение во многих местах, до того что люди, типа, говорят: „Слушай, да пошла ты, Microsoft“», – сказал Скотт, имея в виду, что GPU забрали у части собственных внутренних исследований Microsoft. «И мы сделали эту ставку, потому что верим: вы все делаете абсолютно лучшую работу в отрасли».

* * *

ChatGPT окончательно закрепил поворот OpenAI от некоммерческой модели к коммерциализации. Альтман и другие руководители стремились развить успех чат-бота, запустив целую россыпь платных продуктов. В феврале 2023 года компания выпустила платную версию ChatGPT; в марте один за другим – её API-версию, Whisper API и, наконец, GPT-4. «После ChatGPT путь к выручке и прибыли стал очевиден, – говорит бывший сотрудник. – Доказывать, что мы идеалистическая исследовательская лаборатория, было уже невозможно. Нас ждали клиенты, которых нужно было обслуживать здесь и сейчас».

Взрыв новых продуктов снова захлестнул команду доверия и безопасности. Какое-то время OpenAI заманивала пользователей в свой API, давая им для начала бесплатные кредиты на двадцать долларов. Сверхпопулярность ChatGPT вызвала и резкий рост использования API, и этот бонус за регистрацию превратился в проблему: многие пользователи массово создавали новые учётные записи, чтобы снова и снова получать бонус. В некоторых случаях пользователи

заводили новые учётные записи, чтобы обойти баны и блокировки старых. Массовое мошенничество оборачивалось для OpenAI огромными потерями выручки, а расходы росли вместе с потребностью во всё новых серверах. Команда доверия и безопасности, по-прежнему насчитывавшая меньше двадцати человек и сильно стеснённая в реагировании на нарушения постфактум, снова перебросила людей, чтобы бить по головам в этой новой игре в «прихлопни крота».

Совсем скоро эти постоянные метания доведут Уиллнера до тяжёлого выгорания. Через несколько месяцев он и несколько его сотрудников уйдут из компании. К концу того года команду расформируют, а часть оставшихся сотрудников вольют в более широкое направление систем безопасности под руководством давней исследовательницы OpenAI Лилиан Венг. Некоторым из специалистов по доверию и безопасности со временем станет казаться, что постоянные стычки между Applied и кланом безопасности с его чрезмерным упором на думерство взрастили у многих руководителей компании привычку сильно обесценивать любые «безопасностные» опасения. В итоге их функция, и без того лишённая влияния в большинстве технологических компаний, в OpenAI оказалась бесправной вдвойне.

И действительно, с каждым новым запуском стычки продолжались – в том числе вокруг выпуска GPT-4. Если многим в Applied полугодовая отсрочка запуска модели казалась более чем осторожной, то некоторым в клане безопасности казалось, что её всё равно не хватило, чтобы закончить всестороннее тестирование и согласование.

Например, высокую частоту галлюцинаций модели по-прежнему было особенно трудно обуздать, даже несмотря на согласованные усилия по решению этой проблемы с помощью RLHF. В ноябре 2022 года, когда пользователи вцепились в ChatGPT так, будто это поисковый инструмент, и повсюду пошли разговоры, что он может свергнуть Google, во внутреннем документе отмечалось, что во время внутреннего теста модель OpenAI галлюцинировала примерно на 30 процентах так называемых вопросов закрытой области.

Вопросы закрытой области должны быть самой лёгкой категорией: пользователь спрашивает модель только о той информации, которую сам ей дал, – например, загружает PDF и просит краткое изложение или даёт тезисы и просит переписать их связными предложениями. В противоположность этому вопросы открытой области – это когда пользователь задаёт модели вопрос без справочного материала: о поп-культуре, древней истории, школьной биологии, – как обычной поисковой системе.

Тем временем GPU превратились в постоянное ограничение для исследований и расширения OpenAI. Новые исследования, запуски продуктов и функций снова и снова приходилось откладывать или класть под сукно из-за нехватки чипов. После того как ChatGPT стал вирусным, SemiAnalysis, отраслевая рассылка о полупроводниковой индустрии, подсчитала, что

компания тратит около 700 000 долларов в день на одни только вычисления. После того как руководство перераспределило чипы из подразделения Research, Applied обязалось вернуть их к определённому сроку. Срок наступил и прошёл, а Applied вернуть их не смогло. Число пользователей продолжало стремительно расти, и подразделению нужно было больше чипов, а не меньше.

Это давление ускорило исследования OpenAI в области более эффективных моделей. Работая над повышением вычислительной эффективности компании, подразделение Research придумало новый метод разработки моделей на основе трансформера, которые обходились дешевле в обслуживании пользователей. Метод назвали DUST («пыль»), а получавшимся моделям давали кодовые имена в той же пустынной теме. Первую, оптимизированную версию GPT-3.5, назвали Sahara и в феврале 2023 года выпустили под публичным названием GPT-3.5 Turbo. Другую назвали Gobi – это должна была быть оптимизированная версия одной из моделей текста и изображения.

Третью, задуманную как оптимизированная версия GPT-4, назвали Arrakis – в честь планеты-пустыни из научно-фантастической эпопеи «Дюна». Но после нескольких месяцев работы команда всё ещё не могла сделать Arrakis эффективнее, сохранив прежнее качество. Проект поглощал значительные вычислительные ресурсы. Вскоре руководство закрыло его, чтобы освободить GPU для других проектов.

* * *

Пока Microsoft беспокоилась, не может ли OpenAI просто выйти из отношений, OpenAI ощущала собственную уязвимость: не перестанет ли Microsoft сотрудничать, если стартап не будет изо всех сил угождать партнёру. Arrakis в этом смысле стал особенно чувствительной неудачей. В надежде впечатлить гиганта OpenAI перестроила свою дорожную карту, поставив поставку этой модели выше собственных проектов, стратегически более важных для неё, включая попытку применить GPT-4 к продукту поисковой системы. Вместо этого провалившаяся затея разочаровала некоторых высших руководителей Microsoft.

Возникла и новая неловкая реальность: OpenAI и Microsoft начинали конкурировать за контракты. Codex и DALL-E 2 убедили OpenAI сохранить за собой контроль над доставкой своих технологий напрямую пользователям. ChatGPT и GPT-4 показывали, что OpenAI может и сама зарабатывать деньги. А значит, она напрямую предлагала клиентам ту самую технологию, которую передавала Microsoft, а та потом предлагала её тем же самым клиентам.

Передача технологии тоже была делом непростым. По мере того как ускорялся график выпусков OpenAI, ускорялся и график Microsoft. Но Microsoft была несоизмеримо больше OpenAI, и выходило так, что с каждым новым продуктом одного сотрудника OpenAI могли засыпать вопросами о технических или организационных подробностях десятки коллег из разных отделов

Microsoft. Сотрудникам OpenAI становилось всё труднее и утомительнее поддерживать выпуски Microsoft, сосредоточиваясь при этом на собственной дорожной карте.

Сглаживать отношения в основном приходилось Мурати. Она старалась тесно работать со Скоттом и другими руководителями Microsoft: согласовывать сроки выпуска продуктов, вырабатывать стратегию, как OpenAI и Microsoft будут разграничивать свои предложения, и искать более продуктивные способы совместной работы двух организаций. Летом 2023 года, чтобы снизить нагрузку на общение, команда инженеров Microsoft начала работать внутри OpenAI с полным доступом ко всему, чтобы упростить передачу технологий.

С углублением интеграции и Альтман, и Наделла как никогда активно вникали в дела, особенно в управление вычислительными ресурсами, принимая трудные решения о том, как перераспределять чипы и деньги на новые чипы для вечно голодных до вычислений операций OpenAI. Наделла расскажет The New York Times, что запросы OpenAI росли так быстро и так высоко, что Альтман стал звонить ему каждый день со словами: «Мне нужно больше, мне нужно больше, мне нужно больше».

Для обслуживания ChatGPT OpenAI нужны были не только новые центры обработки данных. Ей по-прежнему требовались гораздо более мощные суперкомпьютеры для обучения будущих поколений моделей. Чтобы удовлетворить этот агрессивный и растущий спрос, две компании набрасывали контуры нового беспрецедентного проекта, который в OpenAI назывался Stargate, а в Microsoft – Mercury: единый суперкомпьютер, одно только строительство которого, по оценкам, обошлось бы в 100 миллиардов долларов. Империя ИИ возвращалась ровно к той же форме экспансии, что и империи прошлого: чтобы подпитывать свой рост, ей нужно было больше материальных ресурсов и, что принципиально важно, больше земли.

Глава 12

Разграбленная Земля

В Сантьяго, столице Чили, в дни после хорошего дождя смог смывает, и открывается потрясающий вид на Анды. Анды тянутся вдоль всей этой узкой полоски страны – от Патагонии на её южной оконечности больше чем на 2600 миль на север, вдвое дальше, чем от Майами до Бостона, – до северной границы Чили. В этой части страны городские пейзажи столицы сменяются бескрайней пустыней, которая, если не считать отдельных районов Антарктиды, – самое сухое место на Земле. Горы проступают резче, а их цвета под беспощадным солнцем становятся ярче.

До того как Чили стала Чили, пустыня Атакама, как её называют, была домом для многих коренных народов. Там, где другие, возможно, увидели бы суровую бесплодную землю, они сумели сделать пустыню своим домом: добирались до воды и минералов глубоко под землёй, чтобы выращивать урожай, разводить скот и совершать обряды предков. Потом пришли испанцы. Они разрезали регион на административные единицы с границами. До сих пор старейшины коренных народов предупреждают в устных преданиях о репрессиях, которые стали такими жестокими, что испанцы отрезали язык каждому, кто осмеливался по-прежнему говорить на родном языке. С тех пор империя определила отношения страны с остальным миром: она будет поставлять сырьё – землю, воду, энергию, минералы – для укрепления политических и экономических планов других государств.

Сегодня почти 60 процентов экспорта Чили составляют минералы, добываемые в основном в пустыне Атакама: прежде всего медь, металл с высокой проводимостью, который используется во всевозможной электронике, а в последнее время ещё и литий, незаменимый компонент литий-ионных аккумуляторов. Этот и другой сырьевой экспорт движет экономикой страны. В Сантьяго у каждого есть знакомый, живущий в ритме горнодобывающей отрасли: на время рабочих «вахт» такие люди живут на севере, а в выходные возвращаются в столицу.

Создать какую-нибудь другую отрасль, которая стала бы экономическим двигателем, стране так и не удалось. Ещё долго после конца Испанской империи печально известную ключевую роль в закреплении такой траектории сыграли США. В 1950–1960-е годы, когда в Чили и Уругвае укоренялась экономика развития, отдававшая предпочтение сильному государственному регулированию и внутренней индустриализации как пути к зрелости развивающейся экономики, американские транснациональные компании, заработавшие миллиарды на своих чилийских рудниках, начали тяготиться растущими государственными налогами и ограничениями. Правительство США

взялось перекроить экономическую политику Чили, чтобы она стала выгоднее американскому бизнесу, и в 1956 году запустило программу под названием «Чилийский проект», по которой сотню чилийских студентов обучали в Чикагском университете под интеллектуальным руководством американского экономиста Милтона Фридмана.

Фридман был исполином экономической науки, впоследствии, в 1976 году, получил Нобелевскую премию, а его идеи лучше всего выражает заголовок его влиятельной колонки 1970 года в The New York Times: «Социальная ответственность бизнеса – увеличивать свою прибыль». Фридман отстаивал всё то, чего не отстаивала экономика развития: полное отсутствие государственного регулирования, ничем не стеснённую свободу для компаний, движимых прибылью, путь экономического созревания развивающихся стран, обращённый вовне, например через свободный экспорт. Как подробно показывает Наоми Кляйн в своём международном бестселлере 2007 года «Доктрина шока», «Чилийский проект» был не образованием, а идеологической обработкой. В Чикагском университете чилийских студентов – а позднее и студентов из других стран Латинской Америки – прямо учили критиковать экономическую политику своих стран и роковые изъяны латиноамериканского девелопментализма.

По мере того как каждый выпуск, их прозвали «чикагскими мальчиками», возвращался в Сантьяго, неолиберальные идеи Фридмана просачивались в интеллектуальную элиту Чили, пока не стали частью правящей идеологии. В 1973 году левого, демократически избранного президента Чили сверг его же генерал Аугусто Пиночет в результате военного переворота, условия для которого отчасти подготовило ЦРУ. Переворот положил начало почти двум десятилетиям жестокой диктатуры Пиночета и новому неолиберальному экономическому курсу: диктатура поручила «чикагским мальчикам» писать её экономическую политику.

При Пиночете Чили приватизировала почти всё – образование, здравоохранение, пенсионную систему, даже воду. Эта стратегия дала экономический рост, но она же разожгла поразительное неравенство. Сегодня Чили – одна из самых неравных стран мира: почти четверть доходов страны сосредоточена у нескольких влиятельных семей, входящих в верхний 1 процент. Так и не проведя настоящей индустриализации, страна по-прежнему привязана к добывающей экономике, которая и делает её значимой для более сильных геополитических держав.

Поэтому с приходом бума ИИ Чили станет эпицентром экстрактивизма нового масштаба, поставляя этой отрасли с её ненасытным аппетитом сырьё – не только медь и литий на севере, но и землю, воду и энергию для растущего числа центров обработки данных в столичном регионе Сантьяго. В мае 2024 года правительство с гордостью объявило, что в ближайшие годы страна примет двадцать восемь новых центров обработки данных в дополнение к уже

имеющимся двадцати двум, что принесёт 2,6 миллиарда долларов иностранных инвестиций.

Правительство считает, что роль поставщика ресурсов для технологического развития означает прогресс для страны, но Чили стала и местом одного из самых яростных в мире сопротивлений этому нарративу. Общины по всей стране неистово борются против того, чтобы у них отнимали землю, воду и другие ресурсы ради замыслов глобального Севера, в которых им нет места и которые не приносят им пользы. Уличными протестами и судебными битвами они затормозили проекты компаний и привлекли внимание правительства. Они вдохновили и жителей других стран подняться из солидарности.

Мартин Тирони Родо, профессор Католического университета в Сантьяго и директор чилийского исследовательского центра Futures of Artificial Intelligence Research, изучающего ИИ под междисциплинарным и латиноамериканским углом, подытоживает настроение, о котором я снова и снова слышала от этих общин, проезжая страну с юга на север, чтобы с ними встретиться. Главный вопрос, который задают эти движения, – как представить себе другой путь развития ИИ, не основанный на добыче, говорит он. «Если мы будем развивать эту технологию так же, как раньше, мы опустошим Землю».

* * *

«Цифровые» технологии не существуют только в цифре. «Облако» на деле вовсе не так эфирно, как подсказывает его название. Чтобы обучать и обслуживать модели ИИ, нужны осязаемые, физические центры обработки данных. А чтобы обучать и запускать генеративные модели ИИ того рода, который первой освоила OpenAI, их нужно больше и они нужны крупнее, чем когда-либо прежде.

Центры обработки данных росли и расползались ещё до ИИ. Когда-то они были достаточно маленькими и рассредоточенными, чтобы прятаться в городской среде: несколько полок с компьютерами в подсобке офиса или несколько десятков стоек в перепрофилированном здании. В нулевые технологические гиганты пошли другим путём, сосредоточив всю свою вычислительную инфраструктуру в огромных складах серверов в сельской местности. Мир центров обработки данных раскололся надвое: гиперскейлеры – и все остальные. Четыре крупнейших гиперскейлера – Google, Microsoft, Amazon, Meta – теперь каждый год тратят на строительство центров обработки данных больше, чем почти все остальные, малоизвестные застройщики вроде Equinix и Digital Realty, вместе взятые.

Трудно представить, как выглядит гипермасштабный центр обработки данных, если вы ни разу его не видели. Мэл Хоган, адъюнкт-профессор Королевского университета в Канаде, изучающая ИИ, инфраструктуру и окружающую среду, когда около десяти лет назад начинала о них писать,

описывала их в футбольных полях. «Теперь футбольные поля даже близко не передают нужных размеров», – говорит она. Гиперскейлеры называют свои центры обработки данных «кампусами» – это обширные участки земли, соперничающие с крупнейшими университетами Лиги плюща, с несколькими огромными зданиями, плотно набитыми стойками за стойками компьютеров. Эти компьютеры выделяют неприличное количество тепла – как надрывающийся ноутбук, помноженный в миллион раз. Чтобы они не перегревались, в зданиях стоят ещё и огромные системы охлаждения – большие вентиляторы, кондиционеры или системы, испаряющие воду для охлаждения серверов. Всё это оборудование вместе создаёт какофонию гула, жужжания и треска, которую – особенно в неразвитых районах – слышно за мили круглые сутки: неумолимый источник шумового загрязнения, буквально выматывающий тело.

Теперь, чтобы обозначить масштаб того, что надвигается в эпоху ИИ после ChatGPT, застройщики пользуются новым словом – «мегакампус». Это слово относится не только к площади, но и к огромному количеству энергии, которое понадобится для их работы. Стойка GPU потребляет втрое больше электроэнергии, чем стойка других компьютерных чипов. И дорого обходится не только обучение генеративных моделей ИИ, но и их обслуживание: по данным Международного энергетического агентства, каждый запрос к ChatGPT требует в среднем примерно в десять раз больше электричества, чем обычный поиск в Google. До недавнего времени крупнейшие центры обработки данных проектировались примерно на 150 мегаватт – то есть могли потреблять за год столько же энергии, сколько почти 122 000 американских домохозяйств. Теперь застройщики и энергокомпании готовятся к мегакампусам ИИ, которым вскоре может понадобиться от 1000 до 2000 мегаватт. Один такой мегакампус может потреблять за год столько же энергии, сколько примерно полтора – три с половиной Сан-Франциско.

Мест на планете, способных произвести и подвести столько энергии в одну точку, совсем немного. Застройщики вместе с энергокомпаниями по всему миру строят новые электростанции и расширяют набор вариантов. После десятилетия, когда спрос на энергию в США стоял на месте, аналитики Goldman Sachs написали, что внезапная новая волна центров обработки данных ведёт к «такому росту потребления электричества, какого не видели целое поколение». Энергокомпании теперь откладывают вывод из эксплуатации газовых и угольных электростанций и переход на возобновляемую энергию; Microsoft перезапустила Три-Майл-Айленд – атомную станцию близ Мидлтауна в Пенсильвании, где в конце 1970-х произошло частичное расплавление активной зоны, крупнейшая авария на коммерческой АЭС в истории США. По прогнозам, при нынешних темпах роста к 2030 году центры обработки данных будут потреблять 8 процентов электроэнергии страны против 3 процентов в 2022 году;

вычисления для ИИ во всём мире могут потреблять больше энергии, чем вся Индия – третий по величине потребитель электроэнергии в мире.

Этот масштаб – мега-гипермасштаб – породил поразительные последствия для окружающей среды. И всё же в тот же самый момент корпоративное замалчивание этого воздействия достигло новых высот. После статьи Эммы Струбелл и её цитирования Гебру в «Стохастических попугаях» техногиганты скрыли ещё больше технических подробностей своих моделей, так что оценить и отследить их углеродный след стало чрезвычайно трудно. Одновременно эти компании усилили кампании влияния на общество и политиков с помощью мощных контрнарративов: центры обработки данных станут настолько эффективными, что их воздействие перестанет быть проблемой; генеративный ИИ откроет путь к новым климатическим инновациям; AGI раз и навсегда решит проблему изменения климата.

Последнее утверждение доказать невозможно, а первые два сильно вводят в заблуждение, говорит Саша Луччони, научный сотрудник и руководитель климатического направления в компании Hugging Face, занимающейся ИИ с открытым кодом. Второе особенно вредно: существует действительно множество технологий ИИ, способных ускорить переход к устойчивому развитию, – их каталогизировала инициатива, ставшая некоммерческой организацией, Climate Change AI, – но это почти никогда не технологии генеративного ИИ. «Для климата нужны модели обучения с учителем, модели обнаружения аномалий или даже статистические модели временных рядов», – говорит Луччони, которая также входит в число основателей этой инициативы. Все эти модели относятся к прежним поколениям технологий ИИ – в основном это инструменты машинного обучения, – они маленькие и энергоэффективные, а некоторые могут работать даже на мощном ноутбуке. «У генеративного ИИ совершенно непропорциональный энергетический и углеродный след при очень малой пользе для окружающей среды», – добавляет она.

По словам Луччони, её прежние соавторы из закрытых компаний больше не получают от работодателей разрешения писать вместе с ней статьи о воздействии ИИ на окружающую среду. Вместо этого она работает с внешними соавторами и учёными вроде Струбелл, чьё исследование когда-то её вдохновило. В одной статье, вместе с Ясином Жернитом, руководителем направления «машинное обучение и общество» в Hugging Face, они измерили углеродный след работы генеративных моделей ИИ с открытым кодом как приблизительный ориентир для того, что строят закрытые компании. Они выяснили, что генерация тысячи фрагментов текста генеративными моделями потребляет в среднем столько же энергии, сколько нужно, чтобы зарядить обычный смартфон почти четыре раза. Генерация тысячи изображений потребляла в среднем столько же энергии, сколько 242 полные зарядки смартфона; иначе говоря, на каждое изображение, созданное ИИ, может уходить столько энергии, сколько нужно, чтобы зарядить смартфон примерно на 25

процентов. Статьи Луччони и Струбелл – среди немногих, где всё ещё приводятся количественные оценки углерода, стоящего за генеративными моделями ИИ. Это непрерывная борьба в гору. Однажды, рассказывает Луччони, она обратилась к более чем пятистам авторам самых свежих статей по машинному обучению с просьбой сообщить основные сведения об обучении их моделей. «Ответов почти не было, – говорит она. – Люди просто не отвечали или говорили, что это конфиденциальная информация».

Хотя гиперскейлеры громко говорят на публике об устойчивости своей вычислительной инфраструктуры, руководители Microsoft внутри компании признавали, что непостоянной доступности возобновляемой энергии просто недостаточно, когда центрам обработки данных нужно работать круглосуточно и без выходных. Непрерывная работа считается настолько важной, что Google, Amazon, а с недавних пор и Microsoft строят свои кампусы по три, чтобы на случай отказа любого объекта у резерва был ещё и резерв. Во время урагана «Ирма» во Флориде и урагана «Харви» в Техасе, когда миллионы людей остались без электричества, некоторые больницы эвакуировали пациентов, а сотни тысяч домов и предприятий были повреждены и разрушены, центры обработки данных в этих районах продолжали спокойно гудеть – настолько исправно, что семьи сотрудников одного из объектов, оставшиеся без крова, переселились туда на время стихийного бедствия.

Земля и энергия, нужные для этих мегакампусов, – лишь две из составляющих глобальной цепочки поставок для расширения центров обработки данных. Сюда же относятся огромные объёмы минералов, включая медь и литий, нужных для изготовления оборудования – компьютеров, кабелей, линий электропередачи, аккумуляторов, резервных генераторов, – и огромные объёмы питьевой – да, питьевой – воды, которая часто нужна для охлаждения серверов. (Вода должна быть достаточно чистой, чтобы не засорять трубы и не допускать размножения бактерий; питьевая вода этому стандарту отвечает.) По оценке исследователей Калифорнийского университета в Риверсайде, к 2027 году взлетающий спрос на ИИ может поглощать во всём мире от 1,1 до 1,7 триллиона галлонов пресной воды в год – половину годового водопотребления Великобритании. Ощущаться эти последствия будут неравномерно. Другое исследование показало, что в США пятая часть центров обработки данных ещё до бума генеративного ИИ брала воду из водосборных бассейнов, находящихся в умеренном или сильном дефиците из-за засухи или по другим причинам. А в странах глобального Юга вроде Чили основная тяжесть этих набирающих скорость добывающих экономик нередко ложится на самые уязвимые общины.

По мере того как всё больше общин видят, как центры обработки данных сказываются на их жизни, всё больше из них яростно сопротивляются их бесконтрольному строительству. В ответ застройщики центров обработки данных изощрились в тактиках, позволяющих продолжать дело как ни в чём не бывало: они приходят в общины тайно, под вывеской подставных компаний;

жертвуют деньги на общественные программы, чтобы погасить сопротивление; дают городам обещания насчёт экологичности своих объектов, а затем, когда строительство уже началось и повернуть его вспять труднее, одно за другим от них отказываются. В одном случае в Вирджинии группа жителей, протестовавших против нескольких огромных центров обработки данных, с изумлением обнаружила письмо юриста застройщику с предложением установить за ними слежку. «Нам нужен крот или два в этой группе», – писал юрист.

* * *

С того момента, как в OpenAI сделали ставку на масштабирование, компания стремилась заполучить беспрецедентный объём вычислительной инфраструктуры. «В ИИ больше всего выигрывает тот, у кого самый большой компьютер», – сказал мне Брокман в 2019 году. И Альтман разработал с Microsoft план, как техногигант будет утолять экспоненциально растущий голод OpenAI по вычислениям. Две компании будут вместе проектировать и сдавать десятки суперкомпьютеров для исследований – центров обработки данных, которые нужно оснастить чипами Nvidia для обучения разных моделей ИИ в ходе поисков OpenAI. Главное же – Microsoft построит и серию всё более крупных и мощных суперкомпьютеров для обучения каждого следующего поколения моделей OpenAI.

Альтман стал называть эту серию суперкомпьютеров «фазами» и однажды показал сотрудникам слайд с размерами каждой фазы: фаза 5, последняя из запланированных, была дух захватывающе больше всех остальных. Суперкомпьютер, на котором OpenAI обучила GPT-3, был фазой 1. Объект с десятью тысячами V100 построили в Уэст-Де-Мойне, штат Айова, куда Microsoft впервые пришла в 2012 году. Компания наладила отношения с городскими властями благодаря, по словам тогдашнего мэра, «ошеломляющей» сумме налоговых платежей, которые помогли городу серьёзно улучшить общественную инфраструктуру. За с лишним десять лет компания также вложила около 2,5 миллиона долларов в местные общественные программы, в основном некоммерческие. В Microsoft суперкомпьютер фазы 1 получил кодовое имя Odyssey («Одиссея»), в OpenAI его назвали Owl («Сова») – по заведённому с первых дней обычаю называть вычислительные кластеры подразделения Research в алфавитном порядке именами животных. Когда все двадцать шесть букв закончатся, кластеры начнут называть химическими элементами в порядке атомного номера.

Фаза 2 тоже находилась в Айове, и на ней обучали GPT-4. Начиналась она с восемнадцати тысяч A100, упомянутых в исследовательской дорожной карте 2021 года, а к концу обучения модели разрослась примерно до двадцати пяти тысяч. Для Microsoft фаза 2 была Telemachus («Телемах») – в честь сына Одиссея из греческой мифологии. Для OpenAI – Raven («Ворон»).

Фаза 3 переместилась в Аризону. Благодаря дешёвой земле, хорошим налоговым льготам и близости к Калифорнии Аризона быстро стала излюбленным центром строительства центров обработки данных у всех облачных провайдеров. Старательно наладив выгодные отношения с властями двух небогатых городов прямо под Финиксом, Microsoft в 2018 и 2019 годах купила три участка земли и, как и прежде, сделала пожертвования разным общественным организациям, чтобы заглушить возможные возражения жителей. Общая площадь участков составляла почти 600 акров, или больше 450 футбольных полей. (В 2024 году Microsoft расширит эту площадь ещё на 283 акра.) На каждом участке должен был появиться новый кампус центров обработки данных, обслуживающий клиентов Azure наряду с OpenAI в облачном регионе «West US 3».

Фаза 3 – с кодовым именем Inglewood в Microsoft и Whale («Кит») в OpenAI – должна была обойтись в несколько миллиардов и вместить сотни тысяч Nvidia H100, следующего за A100 поколения, для обучения моделей, которые OpenAI тогда, скорее всего, собиралась назвать GPT-4.5 и GPT-5. Расходы на фазу 4, запланированную в Висконсине, Альтман ожидал на уровне 10 миллиардов долларов при использовании новейших Nvidia B100, поколения, следующего уже за H100. Сумма ошеломляющая, если учесть, что самые дорогие гипермасштабные центры обработки данных тогда стоили порядка 1–2 миллиардов долларов. Останавливаться на этом он не собирался, между делом подбрасывая идею суперкомпьютера за 100 миллиардов долларов для фазы 5. После оглушительного успеха ChatGPT Альтман умерил ожидания. Поскольку очень много чипов было занято обслуживанием ChatGPT для клиентов, Microsoft с трудом успевала закупать чипы достаточно быстро, чтобы завершить строительство фазы 3. Whale разделили на три отдельных кластера – Beluga («Белуха»), Narwhal («Нарвал») и Orca («Косатка») – с планом завершить строительство где-то в 2024 году.

Никто в Microsoft или OpenAI даже не знал, осуществима ли фаза 5 технически. Согласно проектным планам Microsoft и OpenAI, как позже сообщало издание The Information, объекту за 100 миллиардов долларов могло понадобиться до 5000 мегаватт – почти столько же, сколько в среднем потребляет весь Нью-Йорк. План не выглядел самым разумным с финансовой точки зрения как деловая инвестиция, но главным узким местом были не деньги. А энергия. «У нас кончаются земля и электричество», – говорит один сотрудник OpenAI. В обеих компаниях понимали, что фаза 5 станет возможной только при некоторой доле изобретательности. Либо Microsoft и OpenAI придётся разделить суперкомпьютер на несколько кампусов, чтобы распределить энергетическую нагрузку, и придумать, как обучать модель ИИ в удалённых друг от друга местах, либо, как иногда любил говорить Альтман, проблема решится сама собой благодаря будущему прорыву в термоядерном синтезе.

Порой он делился с сотрудниками OpenAI оптимистичными новостями о Helion Energy – стартапе в области термоядерного синтеза, его крупнейшей личной инвестиции, у которого Microsoft уже обязалась покупать электроэнергию, как только заработает станция с целевой мощностью 50 мегаватт. Позже The Wall Street Journal сообщит, что OpenAI и Helion тоже вели переговоры о сделке, от участия в которых Альтман устранился.

Об экологической цене центров обработки данных Альтман и другие руководители на общих собраниях компании никогда не заговаривали. Когда OpenAI обучала GPT-4 в Айове, штат уже второй год страдал от засухи. Агентство Associated Press позже сообщит, что всего за один месяц обучения модели центры обработки данных Microsoft израсходовали около 11,5 миллиона галлонов воды – 6 процентов водопотребления округа. GPT-4 обучали там три месяца. (Представитель Microsoft сказал, что компания стремится к 2030 году повысить эффективность использования воды на 40 процентов по сравнению с уровнем 2022 года и восполнять больше воды, чем потребляет во всех своих операциях по всему миру, уделяя особое внимание регионам с дефицитом воды, где она работает.)

Аризона тоже переживает тяжёлый водный кризис. В 2022 году, когда Microsoft закладывала основу для фазы 3, исследование в журнале Nature Climate Change показало, что юго-запад США переживает худшую засуху более чем за тысячу лет. Эта засуха вкупе с вопиюще плохим управлением истощила до опасно низкого уровня реку Колорадо, от которой Аризона и ещё шесть штатов зависят в снабжении пресной водой. Без решительных мер река может перестать течь. Нехватка воды усугубляет энергетический кризис: изменение климата обрушило на регион непрерывную череду рекордных температур, и семьи выкручивают кондиционеры на полную. Регион отчасти зависит от гидроэнергии плотины Гувера и атомных электростанций с водяным охлаждением. Иными словами, чтобы производить больше энергии, ему нужна вода. В 2023 году в агломерации Финикса побили несколько новых температурных рекордов, и этот год стал худшим по числу смертей от жары: по сравнению с 2022 годом оно выросло как минимум на 30 процентов и превысило шестьсот человек. «Всё сходится в очень непростом направлении», – говорит Том Бушатцке, директор Департамента водных ресурсов Аризоны.

Альтман же заговаривал о своём нетерпении. В марте 2024 года Meta, проспавшая первые годы гонки генеративного ИИ, выйдет на ринг с агрессивными новыми планами: запустить 350 000 H100 в рамках ещё более масштабного строительства инфраструктуры, чтобы поддержать внезапный всплеск своих инвестиций в генеративный ИИ. Это было больше GPU, чем было в распоряжении OpenAI. Альтман был недоволен. Microsoft, считал он, действует слишком медленно, и это стоит OpenAI её конкурентного преимущества.

* * *

В детстве Сония Рамос стала свидетельницей несчастья, определившего всю её дальнейшую жизнь. Она родилась в шахтёрской семье в Чили. Её отец работал на американскую медную компанию; она росла среди детей других рабочих. В 1957 году часть рудника Чукикамата обрушилась, несколько человек погибли, десятки получили ранения. Её отец уцелел, но она помнит, какое горе разворачивалось вокруг неё после катастрофы: пострадавшие семьи скатывались в беспросветную нищету, дети чахли от голода. Четыре десятилетия спустя, когда Рамос начала выступать против горной добычи и стала одним из самых активных и громких голосов коренных народов Чили, рассказывающих о социальном, культурном и экологическом опустошении, которое она несёт, она вспомнит уроки той трагедии: горнодобывающая отрасль – это система, и эта система, предоставленная сама себе, будет добиваться прибыли любой ценой. «Рабочего не существует», – говорит она. Ни одну из жертв не почтили ни церемонией, ни памятью; ни одна из их семей не получила компенсации. «В том месте нет человечности».

Чили – крупнейший в мире производитель меди: на неё приходится четверть мирового предложения. С самого начала добыча меди перекраивала не только землю, но и общества, которые от неё зависят. Иногда последствия видны глазом: Чукикамата сегодня – крупнейший в мире медный карьер открытого типа, зияющая рана в земле, которую регулярно углубляют взрывами. Вынутая порода, сваленная в высоченные горы, словно монументы тем впадинам, из которых её извлекли, понемногу погребает под собой остатки городка, брошенного после того, как его начала поглощать добыча меди. Добыча к тому же выкачала из региона воду для обработки меди. Однажды иностранная транснациональная корпорация израсходовала столько воды, что истощила целый бассейн на соседнем солончаке, или саларе, уничтожив его богатую экосистему.

Менее заметны следы мышьяка, которые отрасль оставляет в воздухе и воде, из-за чего по всему северу страны выросла заболеваемость раком, и то, как добыча перестроила жизнь коренных народов и посеяла раздор между разными общинами. На землях, лишённых воды и минералов, атакаменьо – так называют все отдельные коренные группы, живущие в этом регионе, – больше не могут прокормить себя, выращивая собственный урожай или разводя собственный скот. Эта перемена ввергла их посёлки в глубокую бедность. Вместе с преступностью выросли депрессия, алкоголизм и правонарушения. Им не хватает еды, водопровода, нормальной медицинской помощи и образовательных ресурсов: от миллиардов прибыли, которую их земля принесла кому-то другому, им почти ничего не досталось. Вместо этого многие вынуждены работать на ту самую отрасль, что захватила их территории, и лечиться в маленьких клиниках, которые она спонсирует. Там, где когда-то между ними было больше единства, коренные группы теперь грызутся из-за убывающих ресурсов.

Литий нашли там сравнительно недавно: на него наткнулась американская компания в 1960-х, когда искала воду, нужную для добычи меди. Пробурив салары, она обнаружила высокие концентрации лития в маслянистом рассоле под поверхностью, что открыло новый фронт добычи и ускорило истощение новых экосистем. Сегодня Чили производит примерно треть мирового лития и уступает только Австралии. Добывают его в основном в Саларе-де-Атакама, крупнейшем солончаке страны: рассол выкачивают в мерцающие бирюзовые бассейны и ждут, пока солнце выпарит раствор и он кристаллизуется в литий и другие побочные продукты. Когда-то на саларах жили стаи розовых фламинго, которых атакаменьо считают своими духовными братьями и сёстрами. Теперь фламинго исчезли; маленькой дочери одного из лидеров общины Пейне остаётся помнить о них лишь по рассказам предков да по плюшевому фламинго.

За эти годы атакаменьо слышали немало нарративов, которыми оправдывали всю эту добычу. В 2022 году, когда Европейский союз установил новую политику энергетического перехода и спрос на литий взлетел, и компании, и политики в Чили и во всём мире превозносили значение горнодобывающей отрасли страны для движения к лучшему будущему. Коренные общины, видя, как разрывают на части их землю и их сообщества, спрашивали: лучшему будущему для кого? «У местных жителей никогда нет возможности думать о собственной судьбе вне сил экономики и международной политики», – говорит Кристина Дорадор, микробиолог, которая живёт на севере и изучает его богатое биоразнообразие.

Теперь те же нарративы пускают в ход заново с генеративным ИИ. Ускоренная добыча меди и лития для строительства мегакампусов – а также электростанций и тысяч новых миль линий электропередачи для них – в изложении Кремниевой долины тоже открывает путь к лучшему и светлому будущему. А значит, препятствовать этой добыче – препятствовать фундаментальному прогрессу человечества. Но сопротивляются коренные общины вовсе не горной добыче. «Наши предки были рудокопами», – говорит Рамос. Это они изначально и нашли медь. Проблема, говорит она, в масштабе.

Этот масштаб поглотил всё. Он сделал север и всю остальную Чили полностью зависимыми от отрасли и не позволил возникнуть другим экономикам. Он задушил способность страны – и остального мира – представить себе другие пути, где развитие было бы возможно без разграбления природных ресурсов, говорит Рамос. Позволив создавать огромные генеративные модели ИИ, этот масштаб привёл ещё и к закреплению расистских стереотипов о коренных народах, которые и без того страдают от того, как физически строилась эта технология. В Бразилии художественная выставка 2023 года, одним из организаторов которой был чилийский университет, показала огромную пропасть между реальностью богатых культур коренных народов Латинской Америки и прискорбно убогими изображениями, которые выдавали Midjourney

и Stable Diffusion, рисуя их примитивными, технически беспомощными народами.

В последние годы атакаменьо сопротивляются всё активнее. Они вывешивают на домах чёрные флаги, протестуя против эксплуатации своих земель и своей общины. Они устраивают протесты, физически перекрывая дороги, по которым автобусы и грузовики компаний добираются до рудников. Они нанимают юристов, чтобы отстаивать свои права как коренных народов по международному праву, защищающему их культурный и территориальный суверенитет. Компании и чилийское правительство были вынуждены пригласить их за стол переговоров, и главное среди требований коренных народов – чтобы правительство провело исследования состояния экосистем пустыни Атакама и количественно оценило потери воды и любой непоправимый ущерб.

У Рамос тоже есть собственный фонд, объединяющий, по её словам, «предковое и непредковое», чтобы продвигать и вести научные исследования природных богатств пустыни Атакама. Из-за её уникально экстремальных условий там живёт множество микробных сообществ, которых нет больше нигде, – они потенциально полезны для лекарств или новых источников энергии. По тем же причинам пустыню десятилетиями изучают как аналог марсианского климата. Рамос надеется, что любые открытия помогут доказать ценность сохранения её прекрасной родной земли. Вопреки нарративам о стремительном прогрессе, которыми подпитывают добычу, она ищет новые представления о прогрессе – такие, что несут исцеление, устойчивость и восстановление.

* * *

Пока на севере продолжается борьба Рамос, в самом сердце Чили идёт другая битва – из-за того, что правительство приветствует сами центры обработки данных технологической отрасли. Чем быстрее расширялись гиперскейлеры, обгоняя предложение земли и электроэнергии в привычных для них регионах, тем агрессивнее они стремились застолбить эти ресурсы на новых территориях по всему миру.

Одна только Microsoft потратила в 2024 финансовом году больше 55 миллиардов долларов – почти четверть заявленной выручки, – чтобы построить то, что SemiAnalysis назвала «крупнейшим строительством инфраструктуры, какое когда-либо видело человечество». Google тем временем сообщила в отчёте о прибыли за третий квартал 2024 года, что собирается довести расходы на центры обработки данных примерно до 50 миллиардов долларов за финансовый год. Meta заявила, что, вероятно, закончит финансовый год с расходами до 40 миллиардов долларов на расширение центров обработки данных и инфраструктуры и что, по её оценке, в следующем году они вырастут.

Редким туманным днём в июне 2024 года Александра Арансибия показывает нам дорогу по Киликуре – муниципалитету на окраине Сантьяго, где

она живёт и состоит в совете, – к тому, что для неё стало главным символом зияющей разницы в силе между американскими техногигантами и её общиной. Меньше чем в получасе езды от самых живописных районов Сантьяго, набитых кафе в европейском стиле и веганскими ресторанами, дороги перед нами разваливаются от плохого ухода, а вдоль них громоздятся горы мусора на нелегальных свалках, которые контролирует местная мафия.

Миновав кладбище домашних животных, мы останавливаемся у участка, похожего на заброшенный пустырь с пучками кустарника и редкими чахлыми деревцами, торчащими из земли. Чаще всего земля здесь так иссушена, что напоминает отдельные места пустыни Атакама; сегодня дождь превращает всё в грязь. Посреди участка фиолетовый щит возвещает по-испански: «Добро пожаловать в Городской лес Киликуры» – проект, как поясняется там же, который Google начала в 2019 году, чтобы отблагодарить общину за то, что та приняла у себя её центр обработки данных. На щите есть и схема, разъясняющая, чем полезен «лес»: слева Киликура изображена промышленной зоной, тесно застроенной фабриками, выбрасывающими парниковые газы и загрязняющими воздух; справа – лес, процветающий под щедрым дождём, который льётся из большой тучи с подписью «СМОГ».

Google хвалится этим лесом на своём сайте и в пресс-релизах. Когда я прошу представительницу компании в стране об интервью о деятельности Google в Чили, она вместо этого присылает мне отполированные справочные материалы, а позже добавляет, что для каждого нового центра обработки данных компания создаёт программу поддержки местной общины – в образовании, устойчивом развитии, доступе к интернету, здравоохранении; в Киликуру Google вложила больше 1,2 миллиона долларов. В её материалах в разделе о лесе говорится о жителях, которые пользуются этой зелёной зоной. Жителей нет. Место слишком далеко от любой автобусной линии, а домов вокруг, можно сказать, нет вовсе. За пределами скромного участка, слишком маленького, чтобы на нём уместился даже сам центр обработки данных Google, бродит дюжина бездомных собак: лают и роются в мусоре. Представительница сказала, что лес обновляется, чтобы «развить впечатления общины».

Арансибия криво улыбается, пока мы оглядываем эту картину.

– Чувствуете себя в Кремниевой долине?

* * *

Арансибия только поступила в университет, когда поняла, что Киликура – место, куда всё выбрасывают. Каждый день по дороге в университет она проезжала через части муниципалитета, заваленные отбросами, о существовании которых раньше не знала. Она никогда не считала Киликуру «домом» – это было просто место, где она жила, неразвитый и ничем не примечательный рабочий городок, куда переехали её родители, когда она была

маленькой. Но, увидев, что к нему в буквальном смысле относятся как к свалке, она ощутила глубокое желание вернуть этой земле былую красоту.

Всего два десятилетия назад, когда Арансибия была ребёнком, Киликура была в основном сельской местностью: холмистые пастбища и блестящие водно-болотные угодья, дом для иного, но столь же богатого, как в пустыне Атакама, биоразнообразия – птиц, зверей и разных растений. Потом пришла мусорная мафия, позволявшая кому угодно из остального Сантьяго за плату сваливать отходы в Киликуре. Некоторые свалки работают так давно, что поросли травой и сорняками и стали похожи на жутковатые уродливые холмы, сжимающие пейзаж. Следом пришли разные отрасли, в том числе пивоваренные компании и застройщики, которые отхватили ещё земли и стали забирать воду из водно-болотных угодий для своих нужд. Сегодня лишь крошечные, разбросанные островки зелени на территории этого муниципалитета площадью двадцать две квадратные мили напоминают о том, какой была Киликура.

На этом фоне в 2012 году Google пришла в Киликуру, чтобы построить свой первый в Латинской Америке центр обработки данных. На своей веб-странице компания с гордостью представляет этот центр, заработавший в январе 2015 года, как один из самых эффективных и экологичных на континенте. Тогда в Киликуре на проект никто не обратил внимания; и уж точно никто в более благополучном центре Сантьяго не обращал внимания на Киликуру. Местные жители, каждый день проезжавшие мимо центра обработки данных на автобусе по дороге на работу, думали, что это фабрика, которая делает пиво или еду и даёт работу местным.

Центр обработки данных Google – как и большинство центров обработки данных – почти не дал рабочих мест, если не считать начального строительства. Вакансия 2024 года на должность техника-механика, одну из немногих долгосрочных, была размещена на сайте вакансий Google только на английском; в объявлении говорилось, что Google не будет рассматривать резюме на других языках. Центр обработки данных, как отмечают активисты, не принёс местной общине и почти никакой другой пользы. В расположенных рядом государственных школах по-прежнему нет хорошего интернета или устройств, через которые ученики могли бы им пользоваться.

С появлением центра обработки данных Киликура и весь Сантьяго оказались на карте как желанное направление для физической экспансии Кремниевой долины. В 2019 году Google объявила, что построит свой второй латиноамериканский центр обработки данных в столичном регионе Сантьяго. Вскоре о том, что тоже придут, объявили Microsoft и Amazon. Чилийское правительство поспешило их приветствовать, позиционируя страну как безопасную и стабильную гавань для прямых иностранных инвестиций в Латинской Америке, которая в остальном славится ненадёжными правительствами и социальной и экономической нестабильностью. В 2020 году правительство пошло ещё дальше. Оно объявило о проекте прокладки нового

подводного кабеля – своего рода информационной магистрали, – который соединит Азиатско-Тихоокеанский регион напрямую с Америками через центральное побережье Чили недалеко от Сантьяго. Чили станет глобальным узлом цифровой инфраструктуры. Google поддержала это партнёрство.

Но в июле 2019 года, когда Google начала оформлять документы на свой второй центр обработки данных в Чили, за этим наблюдала группа жителей. Компания выбрала для нового объекта Серрильос – ещё один рабочий муниципалитет, граничащий с Сантьяго. Как и Киликуру, Серрильос издавна обходили вниманием и бросали на произвол судьбы. С 1930-х по 1990-е годы цементный завод, принадлежавший бельгийской компании, отравлял общину смертельными дозами асбеста, что один чилийский историк назвал «крупнейшим промышленным геноцидом» в стране. До сих пор жители умирают от рака чаще среднего. Но Серрильос и особенный: в стране, где вода приватизирована, именно здесь находится единственная в стране государственная служба водоснабжения, которая подаёт местные грунтовые воды соседним общинам, а в чрезвычайных ситуациях – и в другие части Чили.

Это уникальное сочетание – история заброшенности и драгоценный источник воды – стало плодородной почвой для расцвета нескольких экологических активистских групп, привыкших быть сторожевыми псами и яростно защищать свои ресурсы от добычи. Тем летом, когда Google подала в экологическое ведомство Чили отчёт на одобрение своего центра обработки данных – процедура по большей части формальная, – водная активистская группа MOSACAT принялась прочёсывать все 347 страниц заявки. Глубоко в её недрах Google сообщала, что её центр обработки данных собирается использовать для охлаждения серверов, по оценке, 169 литров пресной питьевой воды в секунду. Иными словами, за год центр обработки данных мог израсходовать больше чем в тысячу раз больше воды, чем всё население Серрильоса, около восьмидесяти восьми тысяч жителей. Для MOSACAT это было неприемлемо. Объект не только забирал бы эту воду прямо из государственного источника Серрильоса, но и делал бы это в то время, когда под угрозой было всё снабжение страны питьевой водой. В 2019 году, как и Айова и Аризона, Чили уже девятый год переживала разрушительную и небывалую в истории мегазасуху.

* * *

Тания Родригес из MOSACAT протягивает мне все 347 страниц экологической заявки Google, распечатанные и сшитые на спирали между двумя синими пластиковыми обложками. Фолиант с глухим стуком опускается мне на колени – физическое воплощение того, как Кремниевая долина пускает в ход технические знания, чтобы оправдать централизованное принятие решений. Снизу торчат аккуратно подписанные стикеры. «Agua potable» («питьевая вода») – гласит один по-испански, отмечая страницы, где говорится о том, что центру обработки данных нужно потреблять для охлаждения пресную воду.

MOSACAT была основана в 2019 году, когда активисты из нескольких разных движений, боровшихся за права женщин, за жильё, за права трудящихся и за охрану природы, объединились из солидарности в единый коллектив. Многие познакомились, протестуя против незаконного горнодобывающего проекта. По словам группы, активистам MOSACAT удалось выгнать горняков, закрыть проект и добиться, чтобы эту землю объявили охраняемым природным заповедником. Вскоре после этого один друг группы, ныне депутат национального конгресса Чили, рассказал им о проекте центра обработки данных Google и убедил посмотреть на его предполагаемое водопотребление.

Члены MOSACAT не технари. Но они прочитали каждую страницу с запутанными схемами и мудрёной терминологией, сделали множество заметок и досконально изучили центры обработки данных и их системы охлаждения, чтобы подготовиться к схватке с Google. Когда я спрашиваю Родригес, как им удалось переварить всю эту информацию, она задорно смеётся. «Понадобились мы все», – говорит она, имея в виду больше десятка добровольцев, которые составляют MOSACAT и работают в часы, урванные у работы и семейных забот.

Чаще всего проекты в Чили, требующие воды, одобряют долго. В случае Google одобрение пришло быстро, даже посреди череды чрезвычайных ситуаций с водой из-за засухи. Сначала MOSACAT попыталась оспорить проект через местного партнёра Google – чилийскую инвестиционно-сервисную фирму Dataluna. Первая встреча прошла плохо, говорят в MOSACAT: представители Dataluna, похоже, плохо понимали проект и отрицали, что он будет использовать пресную воду.

Тогда MOSACAT пошла к местным властям. Мэр и городской совет сами встречались с Dataluna, вспоминают в группе, и пребывали в ложном убеждении, что для охлаждения центру обработки данных нужны только сточные воды. Когда MOSACAT ввела их в курс дела, встревоженные власти потребовали объяснений от Dataluna. Дело пошло вверх – от Dataluna к чилийскому подразделению Google и дальше, до самой штаб-квартиры Google в Маунтин-Вью, Калифорния.

В октябре 2019 года Google прислала в Серрильос двух инженеров и юриста, чтобы выступить перед общиной. В день их приезда MOSACAT обклеила протестными плакатами улицы вдоль маршрута, по которому сотрудники Google должны были ехать к месту встречи. На саму встречу MOSACAT пришла не одна. Среди примерно двух десятков пришедших жителей были представлены ещё шесть активистских и общественных групп. Инженеры Google были гринго, вспоминают в MOSACAT, – высокие и говорившие только по-английски. Они сделали очень техническую презентацию, а юрист Google заодно работал переводчиком. Члены MOSACAT, знающие английский, говорят, что во время обсуждения слышали, как юрист неверно переводит их слова. В какой-то момент представители Google попытались успокоить общину, предложив посадить городской лес – точно такой же, какой техногигант подарил Киликуре.

Представление не впечатлило ни MOSACAT, ни другие группы: Google пришла сюда не для того, чтобы по-настоящему вникнуть в то, чего хочет община, и выслушать её. «Они пришли нас запугать, – говорит член MOSACAT Алехандра Салинас, которая также состоит в совете соседнего с Серрильосом муниципалитета Майпу. – Подумайте только. Они приходят и предлагают нам деревья, а сами высушивают нашу землю».

Жителям Серрильоса не нужны были деревья. Им не нужно было, чтобы Google строила им парк, – как будто столичный регион Сантьяго такая отсталая глушь, что парков там нет. Им нужно было, чтобы Google перестала обращаться с их землёй как с местом, где можно грабить драгоценную воду и другие ресурсы; чтобы компания перестала отмахиваться от общины как от сторонних наблюдателей, а не участников развития местных проектов. «Мы знаем, что кормим мир, что поставляем сырьё вроде меди и лития, – говорит Салинас. – Никто не говорит, что наше сокровище принадлежит только нам и мы им не поделимся. Да, мы можем помогать друг другу. Но они не придут и не станут пользоваться водой, без которой нет жизни, оставив нас ни с чем».

В то время Чили переживала масштабное, растянувшееся на месяцы политическое потрясение, известное как Estallido Social – «Социальный взрыв». С того же октября 2019 года каждую неделю вспыхивали бурные, порой жестокие протесты – коллективный крик против безработицы, приватизации и глубокого неравенства, в ходе которого тысячи людей получили ранения и десятки погибли. В конце 2019 года, когда общины по всей стране в ответ на движение за реформу местной и национальной политики начали проводить референдумы, MOSACAT воспользовалась референдумом в Серрильосе, чтобы добавить в него вопрос: согласны ли жители, чтобы Google построила объект, который будет потреблять столько воды общины? Активисты развернули кампанию, чтобы рассказать об истинной природе проекта: раздавали листовки на перекрёстках, стучались в двери, развешивали плакаты по всему муниципалитету.

В декабре 2019 года MOSACAT победила: строительство центра обработки данных отвергли небольшим большинством голосов. На следующий год власти вместе с MOSACAT подали против проекта иск в экологический суд.

На встрече с Google её представители в конце концов изобразили дружелюбие, говорят в MOSACAT, и проявили больше готовности к переговорам, когда стало ясно, что некоторые жители их понимают. Но, как говорили мне и Хулиан Посада, ассистент-профессор из Йеля, и Мерси Мутеми, юрист Мофата Окиньи, когда страна глобального Юга даёт отпор, всегда есть риск, что компания из Кремниевой долины соберётся и увезёт свои деньги куда-нибудь ещё. Проект продолжал буксовать, а потребность Google в вычислениях росла, и компания объявила, что перенесёт свой следующий запланированный центр обработки данных в Латинской Америке из Чили в другую страну.

* * *

В Уругвае, маленькой стране с населением 3,4 миллиона человек, у национальной телекоммуникационной компании Antel есть три центра обработки данных, которые помогают обеспечивать интернетом и сотовой связью всю страну. Все вместе они занимают лишь около пяти тысяч квадратных метров. Один, меньше тысячи квадратных метров, пристроился на обычной жилой улице Монтевидео: он выше и шире соседних зданий, но вписан в окружение.

Центр обработки данных питается от двух линий электропередачи, отдельных от линий остального района. Тридцать процентов площади занимают компьютеры, 70 процентов – административные помещения, электрощитовые и технические комнаты. Компьютеры нагреваются, но не настолько, чтобы их нельзя было охлаждать воздухом вместо воды. Они издают тихий гул, почти неслышный в дневной суете и как раз настолько докучливый в ночной тишине, что две соседние семьи довольно регулярно приходят и стучат в дверь центра с жалобами. Признаваясь в этом, руководитель центра Хавьер Эчеверрия выглядит смущённым. Он говорит, что сейчас вместе с местными исследователями ищет способы снизить шум и уже переделал систему охлаждения, чтобы она стала тише. Эта отзывчивость – полная противоположность той канители, через которую пришлось пройти MOSACAT, чтобы привлечь внимание американской компании.

В получасе езды, сразу за городской чертой, правительство выделило большой участок земли под научный парк Parque de las Ciencias, который работает как zona franca – зона свободной торговли. Некоторые с горькой усмешкой называют её zona America, потому что там размещаются в основном американские компании, не платящие налогов государству. Парк даже чем-то похож на Америку: пышные ухоженные газоны, симметричная планировка и величественный фонтан в форме солнечных часов напоминают о парадной эстетике Национальной аллеи в Вашингтоне. Главная страница сайта парка рекламирует политически и экономически стабильную страну, много земли и обильное снабжение электроэнергией и водой. Так и вышло, что в 2021 году, когда GPT-3 подстегнула новый интерес к резкому масштабированию моделей ИИ, Google купила здесь двадцать девять гектаров земли – в пятьдесят восемь раз больше, чем вся площадь центров обработки данных Antel, – чтобы разместить в другом месте свой второй центр обработки данных в Латинской Америке.

Но в то время обилия воды в Уругвае на самом деле не было. Как и Чили, как Айова, как Аризона, Уругвай переживал разрушительную засуху. Нехватка воды была такой острой, что фермеры теряли весь урожай и сельскохозяйственные потери страны превысили миллиард долларов; к лету 2023 года власти Монтевидео начнут подмешивать в городскую питьевую воду загрязнённую солёную воду. Открыв кран, семьи видели, как из него льётся гнилостная буроватая жидкость с резким запахом химии. Те, кто мог себе

позволить, покупали для питья бутилированную воду и мылись с открытыми окнами, чтобы не надышаться канцерогенами. Те, кто не мог, всё равно пили из-под крана, и многие страдали от болей в желудке, кожной сыпи, обострения имеющихся болезней и мучились от растущего числа выкидышей.

После свирепой пандемии многим бутилированная вода была не по карману. Пока Кремниевая долина шла вверх и рыночная капитализация и Google, и Microsoft достигала пика выше 2 триллионов долларов – отчасти потому, что компании уходили на удалёнку и всё больше полагались на облачные сервисы, – нелегальные поселения в Уругвае разрослись на порядки. В ollas, местном аналоге бесплатных столовых, голодных детей отправляли ни с чем. Те, кто держал ollas, сами жили в нищете и едва выживали. Фабиана, громогласная хозяйка одной из olla, живущая в нелегальном поселении, которую ласково зовут Reina Madre («Королева-мать»), затихает, вспоминая об этом. «Приходилось говорить: „У меня даже тарелочки для вашего ребёнка нет…“» Она умолкает. «Это было ужасно». Даже после целой жизни в бедности, в которой было и мытьё полов в борделе в семь лет ради выживания, годы пандемии и засухи остались у неё в памяти одними из худших.

Водный кризис вырос из совокупного действия изменения климата и провала государства в распределении пресной воды: в Уругвае больше 80 процентов пресной воды страны идёт не на нужды людей, а в промышленность – прежде всего в экспортное сельское хозяйство. Это промышленные фермы, выращивающие сою и рис, а также деревья для производства бумаги. Большинством таких ферм управляют не местные компании, а транснациональные, которые экспортируют выращенное и почти не несут ответственности за природную среду Уругвая. Их деятельность истощает питательные вещества в почве, из-за чего выращивать настоящую еду становится труднее, и загрязняет водотоки страны таким количеством удобрений, что Уругвай входит в число мировых лидеров по потреблению удобрений на душу населения, а заболеваемость раком там необычайно высока.

Даниэль Пенья, социолог из Университета Республики в Монтевидео, много лет изучающий политику этого экологического экстрактивизма, проводит прямую связь с колониальной историей Уругвая. На потрёпанном пикапе он колесит по стране, беседуя с фермерами и жителями беднейших районов, чтобы вблизи задокументировать, как их выжимает промышленность. Как и в Чили, иностранные транснациональные компании по-прежнему стоят в политической иерархии выше местных. Во время засухи промышленность продолжала расходовать воду без ограничений, забирая столько, сколько нужно, прямо из главной реки, Санта-Лусии, которая питает и городской водопровод Монтевидео. Два десятилетия назад, после серьёзной борьбы экологических активистов, Уругвай стал первой страной в мире, признавшей в конституции воду правом человека. Теперь, по горькой иронии, во время нехватки сильнее всего урезали именно питьевую воду, а не воду для промышленности.

Поэтому, когда пришла Google, Пенья был начеку. Регулярно просматривая сайт уругвайского министерства окружающей среды, где перечислены крупные промышленные проекты, он наткнулся на предложение компании о строительстве центра обработки данных. Пенья читал о том, что гиперскейлеры используют питьевую воду даже во время сильных засух, и об активизме общин вроде MOSACAT, которые сопротивлялись таким проектам. Но когда он скачал подробности проекта, цифры по воде оказались засекречены. Он подал запрос на публичную информацию – до того он успешно делал это около двадцати раз, – но министерство продолжало скрывать цифры, называя их коммерческой тайной. Пенья гадал, что они прячут, и тревожился о прецеденте для других облачных компаний, которые неизбежно начнут вслед за Google присматриваться к Уругваю для собственной экспансии. И он сослался на статью конституции о воде. С помощью друга-юриста, согласившегося работать бесплатно, он подал в суд на министерство.

В марте 2023 года, четыре месяца спустя, Пенья неожиданно выиграл дело. Министерство окружающей среды раскрыло, что центр обработки данных Google планировал брать два миллиона галлонов воды в день прямо из системы питьевого водоснабжения – столько, сколько ежедневно потребляют пятьдесят пять тысяч человек. Вскоре после этого значительная часть Монтевидео получала из кранов солёную воду, и откровения произвели эффект разорвавшейся бомбы. Тысячи уругвайцев вышли на улицы протестовать против Google и всех остальных отраслей, из-за которых правительство разбазаривало драгоценную пресную воду страны. Когда я приехала в июне 2024 года, лозунги сопротивления всё ещё были нацарапаны на стенах и придорожных ограждениях города. «Это не засуха, – гласит один. – Это грабёж».

Пенья смотрит на центры обработки данных так же, как Рамос на горную добычу. Проблема не в самой инфраструктуре, а в масштабе, в котором Кремниевая долина пытается её строить. Именно этот масштаб толкнул компании вроде Google и Microsoft расширяться в Чили и Уругвае, хотя эти страны страдали от острой нехватки ресурсов. Именно из-за этого масштаба им нужно в пятьдесят восемь раз больше земли, чем Antel, и они действуют с гораздо меньшей ответственностью перед местным населением. «Это экстрактивистские проекты, которые приходят на глобальный Юг ради дешёвой воды, земли без налогов и очень низкооплачиваемых рабочих мест. А потом они ничего не дают нашей стране; они не улучшают наш доступ к интернету», – говорит он.

Ближе к концу 2023 года Google без лишнего шума обновила свой проект центра обработки данных в Уругвае: теперь он предусматривал безводную систему охлаждения, а сам объект компания пообещала уменьшить втрое. Пенья говорит, что борьба ещё не окончена: теперь правительство скрывает прогнозируемое энергопотребление нового проекта как коммерческую тайну. Пенья также направил в министерство петицию, подписанную больше чем четырьмястами людьми, с требованием более широкой оценки экологического и

социального воздействия всей цепочки поставок центра обработки данных: где добывают минералы для производства его оборудования, как обращаются с занятыми в этом людьми, сколько углерода будет выброшено, как будут утилизировать образующиеся электронные отходы, чтобы химикаты не просачивались в чью-то общину. Большая часть этих последствий Уругвай не затронет, но Пенья ощущает солидарность с другими странами, которых они коснутся. Это, говорит он, «как правило, все страны глобального Юга. У всех нас в итоге одни и те же последствия, только на разных звеньях глобальной цепочки поставок».

В 2024 году экологический суд Чили постановил, что Google не может строить в Сантьяго центр обработки данных, использующий воду. Представительница Google в Чили сказала, что компания по-прежнему привержена стране и Латинской Америке и, когда понадобится, начнёт процедуру получения разрешений на центр обработки данных с воздушным охлаждением в Серрильосе. Но это не затормозило приход в Латинскую Америку других гиперскейлеров. В 2022 году Microsoft окончательно определилась с местом для своего центра обработки данных в Чили, вскоре после второй инвестиции в OpenAI, – прямо в родном городе Арансибии, Киликуре.

* * *

Как застройщик центров обработки данных Microsoft развернулась поздно. До инвестиций в OpenAI Google далеко опережала её по числу объектов, строившихся по всему миру. Но с внезапным взрывом спроса на вычислительную инфраструктуру для своих амбиций в области ИИ Microsoft взяла на вооружение сценарий Google и пошла за ней в те же регионы.

Чили, в которую пришла Microsoft, была не той Чили, что встречала Google. К 2022 году с «Социального взрыва», оставившего неизгладимый след в политике страны, прошло больше двух лет. После месяцев протестов Чили провела примечательный эксперимент по переписыванию конституции при широком участии простых граждан, чтобы заменить ту, что сохранилась со времён жестокой диктатуры Пиночета. В итоге новые проекты конституции не были приняты: два отдельных процесса дали два партийных документа, не получивших широкой поддержки. Но этот опыт вдохнул в молодёжь и особенно в рабочие семьи новый оптимизм в отношении демократии. Подъём левых идей и поддержки привёл, при драматичном повороте событий, к избранию президентом левого милленниала Габриэля Борича Фонта, которому было всего тридцать пять лет. В победной речи Борич повторил лозунг протестов, клеймивший наследие «чикагских мальчиков» при Пиночете: «Если Чили была колыбелью неолиберализма, она станет и его могилой».

Борич, вступивший в должность в марте 2022 года, сам когда-то был участником студенческих протестов. Это придало смелости молодёжи по всей стране: она взяла то, чему научилась во время социального потрясения в

организации и протестах, и создала новое поколение прогрессивных активистских организаций. Они вошли в ритм жизни каждой общины, регулярно собираясь, чтобы мечтать о большом – о том, каким они хотят видеть будущее Чили.

Именно в это время Арансибия вместе с другим молодым жителем Киликуры, Родриго Вальехосом, основала собственную активистскую группу. Они познакомились, организуя протесты во время социального потрясения, и быстро сошлись на почве глубокой любви к природе. Группу назвали Resistencia Socioambiental Quilicura – «Социально-экологическое сопротивление Киликуры», опираясь на устоявшееся в Латинской Америке представление о том, что социальное и экологическое неразрывно связаны.

Когда в Киликуру пришла Microsoft, Вальехос и Арансибия сделали то же, что до них MOSACAT и Даниэль Пенья: принялись штудировать все материалы, какие только могли найти в открытом доступе от Microsoft, и подробно изучать проект. Вальехос, студент юридического факультета Университета Диего Порталеса, допоздна, в перерывах между учёбой, читал техническую документацию и самостоятельно разбирался, как устроены центры обработки данных.

По прогнозам Microsoft, ей требовалось значительно меньше воды, чем Google в Серрильосе. И всё же Вальехос тревожился. Засуха в Чили только усилилась, и ожидалось, что она продлится до 2040 года. Водно-болотные угодья Киликуры, помимо натиска промышленности, остро страдали от ускоряющегося опустынивания. На своём сайте Microsoft хвалилась новыми передовыми разработками в системах охлаждения центров обработки данных, благодаря которым её объектам не нужна вода. Если Microsoft умеет строить безводные центры обработки данных, почему она не делает этого в Киликуре?

«Глубоко поражает, что компания с такой репутацией, как Microsoft, публично придерживается экологичного дискурса, а на деле не соблюдает мировых стандартов инноваций в такой стране третьего мира, как Чили», – писал позже Вальехос в одной статье.

Впоследствии Microsoft объяснит, что разработки, которые она рекламировала, всё ещё находятся в стадии разработки и опробуются лишь в одном месте в США. «Тогда зачем они обещают такие вещи» на своём сайте? – спрашивает Вальехос.

Вальехос привлёк внимание местных и зарубежных исследователей, в том числе Марины Отеро Верзир, директора по исследованиям Nieuwe Instituut – нидерландского Института архитектуры, дизайна и цифровой культуры, – а также Серены Дамбросио и Николаса Диаса Бехарано, исследователей из FAIR, аналитического центра, одним из руководителей которого был Мартин Тирони Родо. Отеро тронули страсть Вальехоса и Арансибии – и их измотанность. Они выбились из сил, читая длинные технические документы Microsoft, пиша критические статьи и непрерывно протестуя, но так и не смогли добиться, чтобы

их выслушали в компании или в правительстве. Отеро думала, как им помочь. Как усадить их за стол переговоров с нужными людьми?

Отеро знала, что у неё есть то, чего нет у Вальехоса: связи с престижными университетами вроде Гарварда и Колумбийского, к которым прислушаются и Microsoft, и правительство. Она развернула многоплановую кампанию и так глубоко в неё ушла, что оставила работу: выступала на громких конференциях о воздействии центров обработки данных на окружающую среду и о борьбе Resistencia против них; налаживала связи с министерством науки, технологий, знаний и инноваций Чили и с представителями Microsoft и Google; знакомила Вальехоса и Арансибию с другими зарубежными исследователями, чтобы о них узнало больше людей.

Вместе с Дамбросио и Диасом Отеро разработала и более умозрительный проект. Все трое имели архитектурное образование и изучали инфраструктуру современных цифровых технологий через призму искусственной среды. Они задумались: а что, если отнестись к центрам обработки данных как к архитектурным сооружениям и принципиально переосмыслить их эстетику, их роль в местных общинах и их отношения с окружающей средой?

Диас любил в поездках бывать в национальных библиотеках – прекрасных зданиях, призванных передать величие памяти и знаний страны. Диаса осенило, что центры обработки данных тоже можно считать библиотеками – со своими хранилищами памяти и знаний. И их тоже можно проектировать гостеприимными и красивыми, а не уродливыми и хищническими.

Это резко расходилось с тем, что Microsoft и Google понимают под «отдачей» обществу, – например, с программами поддержки местных общин у Google, с тем, что Диас называет «шизофреническими» инициативами, которые обычно оторваны от того, как объекты компаний на самом деле сказываются на общинах. Вместе с Вальехосом и Арансибией трое исследователей подали заявку на финансирование и организовали четырнадцатидневный воркшоп, пригласив студентов-архитекторов со всего Сантьяго переосмыслить, каким мог бы быть центр обработки данных для Киликуры.

Студенты создали потрясающие макеты. Одна группа придумала сделать водопотребление центра обработки данных видимым, храня воду в больших бассейнах, которыми жители могли бы пользоваться как общественным пространством. Другая предложила отказаться от бруталистской архитектуры типичного центра обработки данных в пользу «текучей территории данных», где инфраструктура данных сосуществует с болотами, смягчая свой вред. Конструкции центра обработки данных одновременно служили бы подвесными тропами, приглашая жителей Киликуры пройтись по болотам и полюбоваться экосистемой. Между более традиционными серверными залами располагались бы питомники растений и гнездовья для животных, чтобы восстановить биоразнообразие болот. Центр обработки данных забирал бы из болот загрязнённую воду и очищал её перед использованием, а затем возвращал. Сами

компьютеры собирали бы и обрабатывали данные о состоянии водно-болотных угодий, чтобы ускорить восстановление местной природы. «Мы не решаем проблему, но мы представляем себе другие виды отношений между данными и водой», – говорит Диас.

В конце воркшопа студенты представили свои идеи жителям и другим членам общины. «Это был потрясающий разговор, – говорит Отеро. – Видно, сколько знаний у общины. Им было что предложить».

* * *

На третьем году своего четырёхлетнего срока Борич находится под давлением: ему нужны «быстрые победы» в экономике – тем более как молодому левому президенту. А это значит давление в пользу расширения горнодобывающей отрасли, давление в пользу того, чтобы довести до конца приход двадцати восьми новых центров обработки данных. В тот день, когда Борич объявил о плане разработать национальную стратегию по центрам обработки данных, Вальехос прислал мне видео пресс-конференции и эмодзи: растерянное лицо с глазами-спиралями.

Но, справедливости ради, коалиция активистов севера Чили и Сантьяго с исследователями внутри страны и за рубежом явно оставила свой след. В рамках плана по центрам обработки данных министерство науки, которое руководит его созданием, впервые сформировало комитет активистов, с которым регулярно советуется в ходе подготовки документа, и пригласило в него Вальехоса, Арансибию и членов MOSACAT. Министерство науки также курирует законопроект об ИИ, где сформулировано, как Чили хочет подходить к разработке, применению и регулированию ИИ. Если раньше в обсуждениях министерства ИИ представляли безусловным благом, то теперь тон изменился и в нём признаются социальные и экологические издержки технологии.

Чили, как и многие страны глобального Юга, на горьком опыте усвоила, что не стоит сидеть и ждать, пока глобальный Север решит, как будет строить цифровые технологии. «То, как мы строим технологии, отвечает определённым культурным и историческим рамкам», – говорит Айсен Этчеверри, глава министерства. Если глобальный интернет формировался без Чили, то теперь у страны есть возможность формировать ИИ на собственных условиях.

Тирони идёт ещё на шаг дальше. Совершенно ясно, говорит он, что сегодняшняя отрасль ИИ коренится в колониальной идеологии: она навязывает остальному миру своё мировоззрение и свою технологию – что такое ИИ, что такое хороший ИИ, что значит создавать отрасль ИИ. Чили могла бы стать лидером в сопротивлении этому навязыванию. После веков экстрактивизма страна на собственном опыте знает, что значит, когда твою землю под знаменем прогресса выпотрошили, отняли и уничтожили. Этот опыт она могла бы сделать источником, из которого родятся принципиально новые – деколониальные – представления об ИИ.

«На планетарном рынке ИИ мы как страна играем особую роль: даём материалы для развития этой технологии, – говорит он. – Многие компании стараются извлечь из нас много материалов, чтобы создавать ИИ.

«Значит, нам нужно мыслить из этой позиции – этой геополитической позиции, этой земной позиции. Мы можем придумать другой способ связать технологические инновации с Землёй».

Это благородное устремление, а силы, выстроенные против него, могучи.

Глава 13

Два пророка

В мае 2023 года Альтман приехал в Вашингтон, чтобы дать показания в Конгрессе. Это было примечательное представление. Он повторил обещание, что AGI решит проблему изменения климата и вылечит рак, убедительно объяснил, почему технологии OpenAI улучшат существующие и создадут «фантастические» новые рабочие места, уклонился от вопросов об авторском праве и об отсутствии прозрачности и гарантий конфиденциальности в отношении обучающих данных и искренне призвал к регулированию – то есть к регулированию с благословения OpenAI, вызывая призрак Китая, чтобы убедить законодателей не тормозить инновации компании.

Сенаторы были от него в восторге. В показательном обмене репликами, где проявились и его ловкая риторика, и доверие и энтузиазм, которые он вызывал, Альтман дал три рекомендации по политике, уводившие разговор от существующих проблем – труда, окружающей среды, интеллектуальной собственности – к регулированию будущих ИИ-систем и крайних рисков. Во-первых, создать ведомство, которое разработает и будет применять режим лицензирования для моделей, превышающих определённый порог возможностей; во-вторых, выработать набор стандартов безопасности ИИ для измерения «опасных» возможностей; в-третьих, ввести обязательный независимый аудит на соответствие этим стандартам. Позже он уточнил, что пороги возможностей при необходимости можно приблизительно задать порогами вычислений, а к «опасным возможностям» можно отнести способность модели манипулировать и убеждать, а также создавать рецепты новых биологических агентов.

– Были бы вы квалифицированы, если бы мы ввели эти правила, чтобы следить за их исполнением? – спросил сенатор от Луизианы Джон Кеннеди.

– Я люблю свою нынешнюю работу, – сказал Альтман под смех зала.

– Есть ли люди, которые были бы квалифицированы? – спросил Кеннеди.

– Мы с радостью пришлём вам рекомендации насчёт таких людей, да.

– Ладно. Вы много зарабатываете, да?

– Я зарабатываю… нет. Мне платят достаточно, чтобы хватало на медицинскую страховку. У меня нет доли в OpenAI.

Сидя рядом с Гэри Маркусом, Альтман покорил даже одного из самых громких своих критиков. «Позвольте мне добавить для протокола, что я сижу рядом с Сэмом – ближе, чем когда-либо сидел к нему, кроме одного раза в жизни, – сказал Маркус, – и его искренность… физически очень ощутима, так, как это просто не передаётся с экрана телевизора». (Позже Маркус откажется от этого

редкого одобрения: «Я понял, что меня, Сенат и в конечном счёте американский народ, вероятно, провели».)

Команда, готовившая Альтмана, сочла это оглушительным успехом.

Слушания стали вишенкой на торте долгой кампании. После запуска ChatGPT почти все в Вашингтоне отчаянно добивались встреч с OpenAI. Небольшая команда по политике под руководством Анны Маканджу, до того работавшая почти в безвестности, получила лавину запросов. Месяцами, с Альтманом или без него, они ужинали, проводили демонстрации, отвечали на вопросы и излагали то самое законодательное предложение, которое Альтман озвучил в своих показаниях, перед как можно большим числом политиков – от сенаторов, членов Палаты представителей, их сотрудников и членов кабинета до приезжих дипломатов, глав ведомств и вице-президента Камалы Харрис, – с утра до ночи, практически без перерыва.

К началу июня Альтман, по данным The New York Times, лично встретился как минимум со ста американскими законодателями, и некоторые из них на слушаниях с гордостью ссылались на эти частные беседы.

В день показаний Альтмана с несколькими офисами Конгресса должна была встретиться и небольшая группа голливудских концепт-художников, специализирующихся на концептуальном дизайне персонажей, реквизита и других визуальных элементов фильмов. Деньги на перелёт и жильё они собрали через краудфандинг и при этом подверглись угрозам интернет-троллей, а их личные данные выложили в сеть за то, что они выступили против отрасли ИИ. В то самое время, когда голливудские сценаристы – а вскоре и голливудские актёры – вели исторические забастовки, отчасти добиваясь лучшей защиты от ИИ, художники тоже собирались откровенно рассказать о разрушительном воздействии, которое генеративный ИИ уже оказывал на их профессию. Разработчики генеративного ИИ без согласия обучали модели на работах миллионов художников, чтобы создать многомиллиардный бизнес и продукты, которые теперь фактически их заменяли. Исчезавшие рабочие места были надёжной работой для среднего класса – их, возможно, сотни тысяч. «Художникам сейчас так больно. Ни у кого нет работы», – говорит Карла Ортис, концепт-художница, известная работой над «Доктором Стрэнджем» Marvel Studios, входившая в эту группу и подавшая первый иск от художников против нескольких компаний генеративного ИИ.

Когда они прибыли в Вашингтон, несколько их встреч из-за показаний Альтмана перенесли на следующий день, что спутало им все планы, и им оставалось лишь ходить по коридорам у слушаний – в буквальном смысле за дверью зала, где всё происходило. На второй день им снова пришлось бороться за внимание. Альтман присутствовал на закрытом ужине с шестьюдесятью членами Палаты представителей в Капитолии, где угощались искусно приготовленным буфетом с жареной курицей. В то же самое время художники устраивали интерактивный коктейльный час и пытались привлечь как можно

больше сотрудников аппарата лучшим, что мог позволить их бюджет: вином и Chick-fil-A.

Это была маленькая, но мрачно-комичная иллюстрация того, у кого в разговоре о политике в области ИИ есть власть и влияние, а у кого нет.

* * *

Тот же нарратив, которым Альтман давно оправдывал внутри OpenAI сокрытие исследований и движение на максимальной скорости, теперь искусно пускали в ход, чтобы направить американскую дискуссию о регулировании ИИ к предложениям, которые избавили бы OpenAI от ответственности, а в некоторых случаях закрепили бы её монополию. Испытанная карта Кремниевой долины «А как же Китай?» неизменно творила чудеса, отгоняя регулирование. Теперь она била сильнее, чем когда-либо: страх Вашингтона перед Китаем, подогретый ошеломительным взлётом TikTok, достиг новых высот.

Этот страх хорошо иллюстрирует настроение в министерстве торговли, ставшем острием агрессивного наступления правительства США, призванного задушить развитие ИИ в Китае. Годом раньше, 7 октября 2022 года, министерство торговли выпустило директиву, которая, по его словам, должна была подорвать военные достижения Китая в области ИИ. С помощью механизма, называемого экспортным контролем, – способа ограничить продажу определённых технологий другим странам по соображениям национальной безопасности – оно без предупреждения перекрыло экспорт в Китай передовых чипов для ИИ американской разработки, в первую очередь Nvidia. Радиус поражения этой меры был гораздо шире китайской армии: она выбила почву из-под ног у китайского научного сообщества и отрасли ИИ, работавших над чем угодно – от применения ИИ в здравоохранении и образовании до китайских аналогов ChatGPT. «Если бы вы рассказали мне об этих правилах пять лет назад, я бы сказал, что это акт войны – что мы должны быть в состоянии войны», – сказал один аналитик полупроводниковой отрасли.

Оценивая последствия, министерство торговли, похоже, было раздосадовано тем, насколько устойчивым оказался Китай. Темпы развития и внедрения ИИ в стране несколько замедлились, но далеко не настолько, чтобы министерство могло успокоиться. Отчасти это объяснялось манёврами самой Nvidia: Китай был огромным рынком для американского производителя чипов, и Nvidia так же быстро, как министерство ввело ограничения, разработала новые чипы, аккуратно в них укладывавшиеся, чтобы продолжать продавать китайским клиентам. Запрет к тому же подстегнул собственную китайскую отрасль производства чипов, которая долго не могла выпускать чипы столь же хорошие, как у Nvidia. Действия правительства США вызвали всплеск интереса к китайским отечественным альтернативам, дав отрасли мощный приток финансирования и обратной связи для развития.

Но главным препятствием для этих усилий было бурное трансграничное движение ИИ с открытым кодом, которое стремительно воспроизводило закрытые корпоративные генеративные модели ИИ и выкладывало их в интернет, чтобы любой мог скачать их и пользоваться. Энергично наверстав отставание, Meta стала одним из главных игроков, свободно выпускающих свои большие языковые модели. Компания давно отстаивала разработку с открытым кодом; главный научный сотрудник Ян Лекун верил в важность открытой науки. Это был и разумный бизнес. Meta не нужно было продавать генеративные модели ИИ, чтобы зарабатывать, а выпуск бесплатных моделей вместе с их встраиванием в основные продукты помогал ей утвердить лидерство в ИИ, привлекать лучшие научные таланты и дразнить конкурентов за эти таланты – тех, кто как раз зависел от продажи своих моделей.

Накопив примерно столько же вычислительных ресурсов, сколько было у OpenAI через Microsoft, Meta теперь полным ходом создавала свой аналог серии GPT – Llama. Формально Llama не соответствовала строгому определению открытого кода, которое потребовало бы выпустить и веса модели, и её обучающие данные. Но и того, что Meta выполнила лишь первое – бесплатно опубликовала веса моделей, – оказалось достаточно, чтобы Llama, несмотря на правило компании не предоставлять её напрямую в Китае, стала важнейшим строительным блоком китайской отрасли ИИ.

На фоне раздражения и страха в Вашингтоне через два месяца после слушаний, в июле 2023 года, вышла аналитическая записка по политике, вторившая рекомендациям Альтмана. Написанная консорциумом исследователей, в том числе из клана безопасности OpenAI, Microsoft и Google DeepMind, а также из более чем десятка аналитических центров, многие из которых были связаны с сообществом думеров, она снова продвигала новый режим лицензирования моделей ИИ на основе порогов вычислений и разработку оценок безопасности ИИ на опасные возможности, включая способность манипулировать и убеждать и создавать рецепты нового биологического оружия. Документ на пятьдесят одну страницу дал и название категории моделей, которым требовалось такое государственное вмешательство: «передовые» (frontier) модели ИИ. По словам авторов, передовых моделей – моделей, которые могут проявить такие опасные возможности, – ещё не существовало, но при масштабировании существующих моделей компаний вроде OpenAI и Anthropic на всё больших вычислениях они могли появиться внезапно и непредсказуемо в любой момент.

Через несколько недель после выхода записки OpenAI и Microsoft заключили стратегический союз с Google и Anthropic и создали Frontier Model Forum – группу для продвижения соответствующих исследований и влияния на политическую повестку в отношении рисков безопасности ИИ. Это был тот редкий вопрос, в котором интересы думеров, бумеров и корпораций, движимых прибылью, совпадали: держать передовые модели в центре правительственных

дискуссий о регулировании было идеологической необходимостью для думеров и удобно бумерам и корпорациям, потому что отвлекало внимание от регулирования существующих моделей ИИ и их проблем. Все к тому же продвигали свои интересы, выступая против открытия весов передовых моделей. В первый год существования Форума Meta в нём демонстративно отсутствовала. (Она вступит туда через год, чтобы получить место за столом.)

* * *

Ядром рекомендаций Альтмана и самой идеи передовой модели была связь между масштабом модели и возникающими, а значит, возможно, опасными возможностями. Этот довод коренился в философии законов масштабирования – в том, что больше вычислений при обучении предсказуемо даёт более мощные модели, – а также в убеждении думерских кругов, что высокоразвитый ИИ может выйти из-под контроля. Политические предложения, вытекавшие из этого довода, сводились к тому, что регуляторы должны основывать своё вмешательство на том, сколько вычислений ушло на обучение модели глубокого обучения: модели, перешагнувшие определённый порог вычислений, следует автоматически рассматривать с большей осторожностью и ограничивать строже.

Аналитическая записка июля 2023 года предлагала конкретное число для этого порога: 10^{26} операций с плавающей запятой – минимальное общее число вычислений, единица с двадцатью шестью нулями, на котором должна быть обучена модель, чтобы считаться передовой. Авторы признавали, что порог в известной мере произволен, и просто указывали, что это уровень вычислений, который существующие модели ИИ, вероятно, ещё не превзошли. Сара Хукер, вице-президент по исследованиям в компании Cohere, разрабатывающей большие языковые модели, и одна из соавторов записки, говорит, что из разговоров с соавторами, предложившими это число, у неё сложилось впечатление: его выбрали чуть выше того объёма вычислений, который, по имеющимся сведениям, OpenAI использовала для обучения GPT-4.

Но Хукер и многие другие исследователи, в том числе Дебора Раджи, не согласны с тем, чтобы регулировать модели по порогам вычислений. Масштаб может вести к более развитым возможностям, но обратное неверно: развитые возможности не требуют масштаба. Модель глубокого обучения, обученная, например, только на высококачественных биологических данных, может оказаться очень мощным генератором биологических рецептов при очень малом масштабе. С помощью дистилляции – одного из приёмов, упомянутых в исследовательской дорожной карте OpenAI 2021 года, – большие модели можно также превращать в маленькие с похожими возможностями. И масштабирование моделей не гарантирует продвижения в тех или иных возможностях: это опять же зависит от того, что содержится в обучающих данных модели и какой тип нейросети обучают. В конце концов, не все модели построены на трансформерах. Поэтому вычисления почти не связаны с риском

вообще, говорит Хукер, не говоря уж о конкретных видах рисков вроде тех, что изложены в записке.

Поскольку согласия между соавторами записки не было ни насчёт регулирования, основанного на вычислениях, ни насчёт конкретного порога, число 10^{26} операций с плавающей запятой поместили в сноску и приложение без обоснования, вместе с существенными оговорками о том, почему пороги – весьма несовершенный подход. Хукер поразило, как стремительно не только сам подход, но и точный порог превратились в одно из самых популярных политических предложений. После того как записка напрямую сыграла на страхах перед Китаем, эта идея прочно завладела умами в Вашингтоне. Передовые модели звучали пугающе, а ещё страшнее – если до них доберётся Пекин. «Отдельные части администрации хватаются за что угодно, потому что хотят хоть что-то сделать», – сказала мне в конце 2023 года Эмили Вайнштейн, тогда научный сотрудник CSET.

Идеи записки нашли благодарную аудиторию в министерстве торговли. Сотрудники бросились встречаться с экспертами, чтобы выяснить, как мог бы выглядеть контроль над передовыми моделями и можно ли уберечь их от Пекина. Вскоре министерство рассматривало беспрецедентное предложение – распространить экспортный контроль в области ИИ не только на оборудование, но и на сами программы, запретив экспорт моделей ИИ выше порога вычислений. Иными словами, оно оценивало, может ли запретить выкладывать веса моделей в интернет и делать их общедоступными.

Примечательно, что ключевая рекомендация Маркуса и вице-президента IBM Кристины Монтгомери, тоже дававшей показания рядом с Альтманом, не получила и близко такого отклика, хотя они повторяли её на протяжении всех слушаний: обязать компании раскрывать, что именно содержится в обучающих данных, которые они скармливают своим моделям. Это почти не повлияло бы на передачу более развитых возможностей Пекину, о чём тревожился Вашингтон, зато придало бы реальную силу корпоративной ответственности по широкому кругу вопросов, включая использование компаниями материалов, защищённых авторским правом, конфиденциальность данных пользователей и строгую научную оценку возможностей моделей. «Если мы не знаем, что в них, мы не знаем точно, насколько хорошо они работают», – говорил Маркус. Нам просто пришлось бы верить компании на слово.

Такой подход к тому же значительно уменьшил бы неопределённость в вопросе о том, возникнут ли опасные возможности и как, – по той же причине, по которой вычисления плохо предсказывают риск: поведение модели глубокого обучения прежде всего определяется её данными. Если разработчик ИИ создаёт большую языковую модель, способную составлять рецепты биологического оружия, «то потому, что обучил её на наборе данных, где была информация о биологическом оружии», – говорит Сара Майерс Уэст, один из исполнительных директоров AI Now Institute и бывшая старшая советница FTC по ИИ. Как всегда,

нейросеть лишь выявляет закономерности в своих обучающих данных. Открыть эти данные было бы первым шагом к научной ясности в вопросе о том, какие входные данные могут вести к опасным результатам.

* * *

Пока министерство торговли советовалось с разными экспертами по своему предложению ограничить веса моделей, новости об этих обсуждениях – о которых оно объявит в начале 2024 года, открыв публичный сбор комментариев, – раскололи надвое сообщество разработчиков ИИ и быстро растущее сообщество специалистов по политике в области ИИ. Это было столкновение закрытости и открытости, технонационализма и науки без границ. Помимо участников Frontier Model Forum и широкого сообщества думеров, сторона закрытости быстро привлекла на свою сторону американский аппарат национальной безопасности и разведки. Против них выступили Meta, разработчики ИИ с открытым кодом, стартапы, группы гражданского общества и независимые учёные.

Если сторона закрытости по-прежнему подчёркивала многие из тех же доводов, которыми руководство OpenAI годами пользовалось внутри компании, то сторона открытости доказывала, что изоляция моделей принесёт гораздо больше вреда, чем пользы. Узкое место в создании нового биологического оружия, например, вовсе не в том, чтобы найти рецепт, отмечала Вайнштейн. Таких рецептов в сети и так полно, и их легко найти через Google. Дело в том, чтобы раздобыть материалы и оборудование для реального изготовления оружия. Поэтому ограничение доступа к так называемым передовым моделям почти ничего здесь не исправит. Зато побочный ущерб от запрета на публикацию моделей ИИ грозил ослабить сами основы американских инноваций в области ИИ. Открытый код – обмен кодом и программами, выпущенными для широкого сообщества, и развитие на их основе – издавна служит фундаментом, на котором процветает богатство американских стартапов. Запрет на публикацию весов моделей оставил бы небольшим разработчикам меньше путей, чем когда-либо, для создания собственных ИИ-продуктов и сервисов. Он ещё сильнее закрепил бы господство гигантов, представленных во Frontier Model Forum.

К тому же модели ИИ стало бы ещё труднее изучать – как в работах Саши Луччони, Ясина Жернита и Эммы Струбелл, которые во многом опирались на открытые генеративные модели ИИ, чтобы количественно оценить углеродные и экологические издержки их дальнейшего масштабирования.

В важнейших отношениях масштабное открытое сотрудничество поверх границ не только не угрожало национальной безопасности, но и укрепляло американское лидерство в ИИ. Как две страны, дающие больше всего специалистов и исследований в области ИИ в мире, США и Китай давно были друг для друга главными партнёрами в развитии ИИ. Больше десяти лет учёные и предприниматели обеих стран отталкивались от работ друг друга, продвигая

эту область и широкий круг её применений гораздо быстрее, чем каждая группа смогла бы в одиночку, к выгоде не только обеих стран, но и многих других по всему миру. Один из самых известных примеров: ResNet, одну из самых широко используемых нейросетей в мире, опубликовали китайские исследователи из пекинского офиса Microsoft. ResNet не только лежит в основе крупных систем компьютерного зрения, распознавания речи и обработки языка, но и была ключевым ингредиентом первой версии AlphaFold от DeepMind – ИИ-системы, выпущенной в 2018 году, которая умела предсказывать трёхмерную структуру белка по его аминокислотной последовательности, что крайне важно для ускорения разработки лекарств и понимания болезней. (Последующие достижения DeepMind в AlphaFold, уже на другой нейросети, принесут Демису Хассабису и ещё одному ведущему научному сотруднику DeepMind Нобелевскую премию по химии 2024 года.)

И всё же в октябре 2023 года идеи стороны закрытости получат самое весомое на тот момент одобрение: они всплывут в указе администрации Байдена об ИИ. Этот указ, один из самых длинных в истории, читался как наспех склеенные документы, написанные совершенно разными группами, – потому что так оно и было. Один из этих документов опирался на «Проект Билля о правах в области ИИ», подготовленный администрацией в 2022 году, который Белый дом тщательно собирал на протяжении долгого времени в консультациях с группами гражданского общества, чтобы наметить, как развивать, применять и обуздывать ИИ так, чтобы это укрепляло гражданские права, расовую справедливость и защиту частной жизни. В нём, среди прочего, подчёркивались разработка ИИ при широком участии общин и экспертов, борьба с дискриминационным воздействием ИИ в таких сферах, как здравоохранение и найм, и защита людей от сбора данных без их согласия.

Второй документ на удивление точно воспроизводил рекомендации Альтмана и подход записки о передовых моделях; его прикрепили в последний момент, после того как записка захватила внимание нескольких сочувствовавших думерской идеологии людей в Белом доме. Записка подчёркивала необходимость сосредоточиться на будущих моделях ИИ, которых ещё не существовало; указ впоследствии отведёт таким моделям половину своего объёма. Записка приводила четыре примера опасных возможностей. Указ сохранит три из них: создание рецептов нового химического, биологического, радиологического и ядерного (ХБРЯ) оружия, автоматизированные кибератаки и – дословно скопированное – уклонение от человеческого контроля «с помощью обмана и запутывания».

К тревоге Хукер, Раджи и многих других исследователей ИИ, точный порог вычислений из записки, 10^{26} операций с плавающей запятой, тоже появится в указе – как порог, выше которого о моделях нужно будет сообщать правительству США.

Получив такое одобрение, подход с порогами вычислений быстро даст метастазы. К концу года его подхватит Европа, остановившись на 10^{25} – чуть более строгом пороге, – когда законодатели, проталкивавшие давно вынашиваемый Акт ЕС об ИИ, почувствовали, что их переехал внезапный прорыв генеративного ИИ, и спешно искали способы его учесть. В начале 2024 года этот подход распространится и на Калифорнию с внесением нового законопроекта о безопасности ИИ SB 1047, который вернётся к порогу 10^{26}. Губернатор Калифорнии Гэвин Ньюсом впоследствии наложит вето на законопроект, против которого, по иронии судьбы, OpenAI и другие разработчики моделей активно лоббировали из-за широкого набора других его предложений по ответственности. Многие критиковали Ньюсома за то, что он прогнулся под интересы отрасли. Но для нескольких исследователей, включая Хукер и Раджи, вето стало желанным событием. «Это был шаг в верном направлении – к тому, чтобы опираться на научный консенсус, – говорит Хукер. – Остаются большие вопросы: почему именно это число и какие риски вы надеетесь предотвратить».

Вся эта цепочка событий – показания Альтмана, аналитическая записка, массированная кампания влияния на политику, гиперчувствительность Вашингтона к страхам перед Китаем и поспешное закрепление порогов вычислений в важных политических документах в США и за рубежом – наглядно показывала, среди прочего, насколько атрофировалась независимая экспертиза в области ИИ. Месяцем раньше, в сентябре 2023 года, Раджи оказалась единственной учёной, не связанной финансово ни с отраслью, ни с сообществом думеров, которая давала показания в Конгрессе рядом с Альтманом, Маском, Наделлой, Гейтсом, Цукербергом, Пичаи и Джеком Кларком, а также другими руководителями технологических компаний. Все они собрались на самом первом из «Форумов по ИИ» сенатора Чака Шумера – одной из самых громких и значимых серий политических встреч того года, призванной дать ход законодательству об ИИ. Пока соседи по свидетельскому столу сыпали эффектными, ничем не подкреплёнными заявлениями о надеждах и опасностях ИИ, перемежая их своевременными упоминаниями о том, как обойти Китай, от которых присутствовавшие сенаторы выпрямляли спины, больше всего Раджи поразило, как охотно многие в зале, похоже, всему этому верили.

До неё дошло, что люди, сидевшие рядом с ней, и их огромные команды по политике так долго монополизировали повестку в Вашингтоне, что многие политики теперь считали её непреложной истиной. Представитель Шумера позже сообщит прессе, что сенатор, двигаясь к регулированию, лично консультировался с Альтманом и другими руководителями OpenAI. «Для меня это было огромным открытием, – говорит Раджи. – Ого, нам нужно больше людей, которые просто развенчивают – просто смотрят на то, что говорят люди, и говорят: „Вообще-то реальность сложнее“».

* * *

Вашингтон стал лишь кульминацией американского этапа политического наступления обаяния, которое вёл Альтман. В марте того же года, после того как он написал в Twitter, что собирается за границу встретиться с пользователями, его поездка превратилась в одиссею по множеству городов и континентов с фотосессиями, казалось, с каждым президентом из «Большой двадцатки». Теперь у неё было новое название: «Мировое турне Сэма Альтмана».

За «Мировым турне» не стояло никакой продуманной стратегии команд OpenAI по коммуникациям или политике. Альтман просто выбрал первые остановки и разослал их своим более чем 1,5 миллиона подписчиков в Twitter. После первых нескольких появлений интерес снежным комом вышел из-под контроля, и к делу подключили команды коммуникаций и политики. С каждым новым этапом они в спешке улаживали логистику, готовясь к неизбежным трудностям, вытекающим из отсутствия подготовки.

Это было проявлением динамики, которая существовала всегда: Альтман шёл своим путём. Иногда этот путь совпадал с путём компании, иногда нет. При выпуске GPT-4 OpenAI тщательно выстроила все объявления и публичные материалы так, чтобы представить проект как работу всей компании, каковой он и был. Над моделью трудились больше сотни сотрудников. Автором объявления компании значилась просто «OpenAI». А потом Альтман в Twitter приписал заслугу одному человеку – Якубу Пахоцкому. Пахоцкий действительно сыграл важную роль, но был одним из восемнадцати руководителей проекта. Неужели его вклад был единственным в своём роде? – гадали некоторые сотрудники. Затем Альтман пошёл ещё дальше и пригласил Пахоцкого сидеть позади него на слушаниях в Сенате.

Эта динамика проявлялась и по-другому: в компании были официальные руководители, но не всегда было ясно, к ним ли прислушивается Альтман и что направляет стратегию компании – официальные процедуры и решения или его отношения и прихоти.

По мере того как OpenAI стремительно становилась профессиональнее и оказывалась на виду и под прицелом, эта несогласованность наверху становилась всё более значимой. Компания уже не сводилась к подразделениям Applied и Research. Теперь у неё было несколько подразделений, работающих с внешним миром: помимо команды коммуникаций, юридическая команда писала заключения и разбиралась с растущим числом исков. Команда по политике раскинулась по континентам. OpenAI всё больше нуждалась в том, чтобы говорить со всеми, от кого зависела, единым нарративом и единым голосом, и ей нужно было определять свои позиции, чтобы их формулировать. Но раз за разом отсутствие стратегической ясности вело к путанице в публичных заявлениях.

В конце 2023 года The New York Times подаст в суд на OpenAI и Microsoft за нарушение авторских прав – обучение на миллионах её статей. Ответ OpenAI в

начале января, написанный юридической командой, оказался необычайно задиристым: компания обвинила Times в том, что та «намеренно манипулировала нашими моделями», чтобы получить доказательства для своей позиции. На той же неделе команда по политике направила в специальный комитет Палаты лордов Великобритании по коммуникациям и цифровым технологиям заявление, где говорилось, что OpenAI «невозможно» обучать свои передовые модели без материалов, защищённых авторским правом. Когда СМИ вцепились в слово «невозможно», OpenAI поспешно от этой формулировки отступилась.

«Тут всё время сплошная путаница», – говорит сотрудник одного из подразделений, работающих с внешним миром. Отчасти это обычные болезни роста стартапа, но известность и охват OpenAI намного обогнали сравнительно раннюю стадию развития компании, добавляет он. «Не знаю, есть ли у высшего руководства стратегический приоритет. Честно говоря, мне кажется, люди просто сами принимают решения. А потом это вдруг начинает выглядеть как стратегическое решение, хотя на самом деле это просто случайность. Иногда плана нет, а есть только хаос».

* * *

Пока Альтман носился по свету – в Европу, Латинскую Америку, на Ближний Восток, в Азию и Африку, – восхищая (и лишь изредка задевая) тщательно подобранную публику из студентов, технологических инвесторов и поклонников, отсутствие стратегической ясности разжигало давние линии разлома OpenAI и толкало компанию к ещё более противоположным крайностям, чем когда-либо.

На одной стороне по-прежнему в авангарде было подразделение Applied, наперегонки с невиданным числом конкурентов развёртывая технологии OpenAI быстрее, чем когда-либо. Теперь его поддерживали и другие разросшиеся подразделения, и многие в Research, воодушевлённые резким ростом ажиотажа и ожиданий и уверенные, что продвижение и выпуск моделей OpenAI – лучший способ выполнить миссию компании.

На другой стороне был клан безопасности, рассредоточенный по Research, причём многие по-прежнему были сосредоточены в команде исследований политики Майлза Брандеджа и команде согласования Яна Лейке; теперь он составлял куда меньшее меньшинство. Своё отставание в численности они восполняли тем, что громче, чем когда-либо, били тревогу об опасных возможностях и экзистенциальных рисках, которые, по их мнению, могли стать возможными в любой момент. По мере развития моделей OpenAI к их рядам присоединялись и некоторые в Research, прежде не причислявшие себя к клану безопасности: ускоряющийся рост возможностей обратил их в веру, что ИИ может достичь такого уровня интеллекта, который позволит ему ускользнуть из-под человеческого контроля и пойти вразнос. Тот же резкий рост ажиотажа и

ожиданий означал для этой стороны, что у OpenAI есть нравственный долг действовать с максимальной осторожностью, чтобы выполнить свою миссию.

Это были бумеры и думеры во плоти – в стенах OpenAI.

Раскол доходил до самого руководства. После DALL-E и ChatGPT большинство руководителей и старших менеджеров всё больше свыкались с мыслью о моделях как о полезных инструментах, которые нужно выпускать в мир путём «итеративного развёртывания» – выражения, придуманного OpenAI в промежутках между выпусками для описания своего нового подхода. В отличие от поэтапного выпуска GPT-2 или контролируемого выпуска GPT-3 через API, итеративное развёртывание означало идти ва-банк: давать модели в руки пользователям рано и часто. И, как и со всеми прежними стратегиями развёртывания, у OpenAI был новый довод в пользу того, что эта – самый безопасный подход из возможных. Итеративное развёртывание, доказывали Альтман и другие руководители, даст людям и институтам время приспособиться и в то же время позволит компании проверять модели на реальных людях, собирать реальные отзывы и улучшать продукты. «Тайком создать сверхмощную ИИ-систему и потом разом обрушить её на мир – думаю, это закончилось бы плохо», – говорил Альтман на слушаниях в Сенате.

Но если и был в OpenAI руководитель, не шагавший в ногу со всеми, то это Суцкевер. По мере того как модели OpenAI развивались, а последствия их развёртывания нарастали, он считал, что компании нужно не ослаблять, а усиливать бдительность в отношении их способности привести к опустошительным последствиям. После GPT-4 Суцкевер, прежде посвящавший большую часть времени продвижению возможностей моделей, резко развернулся к безопасности ИИ. Он стал делить своё время пополам. Окружающим порой казалось, что он воюет сам с собой. Он был одновременно бумером и думером: как никогда воодушевлён и как никогда напуган тем, что AGI появится и быстро превзойдёт людей, став сверхинтеллектом.

Суцкевер теперь говорил всё более мессианским тоном, озадачивая даже давних друзей и тревожа других сотрудников. На одной встрече с новой группой исследователей Суцкевер изложил свои планы подготовки к AGI.

– Когда мы все окажемся в бункере… – начал он.

– Простите, – перебил его один исследователь, – в бункере?

– Мы обязательно построим бункер, прежде чем выпустить AGI, – как ни в чём не бывало ответил Суцкевер. Столь мощная технология наверняка станет предметом жгучего вожделения правительств всего мира. Она может обострить геополитическую напряжённость; ключевых учёных, работающих над технологией, нужно будет защищать. – Конечно, – добавил он, – идти в бункер или нет – это будет по желанию.

Этот исследователь и дальше будет в равной мере высоко ценить Суцкевера и держаться от него на расстоянии. «Есть группа людей – Илья один из них, –

которые верят, что создание AGI принесёт вознесение. Буквально вознесение», – говорит он.

Пока Суцкевер продолжал уделять часть времени согласованию, у него с Альтманом начала вызревать новая идея: команда, всецело сосредоточенная на разработке новых методов согласования для сверхинтеллекта, – в расчёте на то, что методов вроде обучения с подкреплением на основе отзывов людей станет недостаточно, когда системы, по их мнению, смогут перехитрить людей. Альтман называл это «Манхэттенским проектом согласования». Поначалу они обсуждали, не выделить ли его в отдельную организацию – собственную независимую некоммерческую структуру с начальным фондом в миллиард долларов. Отчасти потому, что Суцкевер не хотел уходить из OpenAI, отчасти из-за вопросов доступа к моделям, они решили оставить проект внутри компании. Затем OpenAI объявила в своём блоге о создании новой команды для руководства этой работой и о её новом названии – Superalignment («Суперсогласование»). В записи также громко объявлялось, что команде будут выделены 20 процентов вычислительных мощностей, которыми OpenAI располагала на тот момент. Руководить новой работой должны были вместе Суцкевер и Лейке.

На другой встрече Суцкевер вышел перед сотрудниками, чтобы представить новую команду и её цели. Он схватил микрофон и начал по нему стучать. Бум. Бум. Бум.

«Согласование – это горящий огонь, – сказал он. – Суперсогласование – это бушующее пекло».

Вскоре после этого, когда компания переросла Mayo – полный растений офис с фонтаном, куда она переехала после пандемии, – руководство перевело большую часть подразделения Research обратно в старое здание Pioneer Building. Переезд во многом развёл две половины компании – Applied и клан безопасности – по собственным мирам.

* * *

В июле 2023 года, вскоре после того как OpenAI объявила о команде Superalignment, она арендовала зал в кинотеатре Metreon в центре Сан-Франциско, чтобы сотрудники посмотрели фильм «Оппенгеймер» – историю физика Роберта Оппенгеймера, руководившего американским Манхэттенским проектом по созданию первого в мире ядерного оружия.

«я надеялся, что фильм про оппенгеймера вдохновит целое поколение детей стать физиками, но он в этом совсем промахнулся, – написал Альтман в Twitter. – давайте снимем такой фильм! (по-моему, „социальной сети“ это удалось с основателями стартапов.)»

Почти восемь лет аналогия между OpenAI и Манхэттенским проектом, которую Альтман провёл в самых первых письмах Маску, оставалась в компании устойчивым мотивом – её использовали даже на вводных занятиях для

новичков. Альтману она нравилась. У него был общий с Оппенгеймером день рождения, на что он обращал внимание журналистов. Ещё он любил пересказывать убеждение создателя бомбы, что «технология появляется потому, что она возможна». Он, похоже, никогда не добавлял, что вторую половину жизни Оппенгеймер провёл, терзаясь сожалениями и борясь против распространения собственного детища.

Разные сотрудники придавали этой аналогии разное значение. Большинство видели в Манхэттенском проекте героический подвиг: он олицетворял способность, сосредоточив огромное количество талантов и ресурсов, совершить спасающий мир, меняющий историю технологический прорыв раньше опасных противников. Для клана безопасности он подчёркивал тяжесть бремени OpenAI, вводящей в мир технологию, которая грозит человечеству экзистенциальной гибелью.

Для Альтмана он был уроком пиара. «То, как мир познакомился с ядерной энергией, – это образ, который никто никогда не забудет: грибовидное облако над Японией, – сказал он однажды, за годы до того, на каком-то мероприятии. – Я много думал о том, почему мир отвернулся от науки, и один из многих ответов, в который я готов поверить, – этот образ и то, что мы узнали: может быть, некоторые технологии слишком мощны, чтобы они были у людей. Людей больше убеждают образы, чем факты».

Тех, кто находился на крайнем краю спектра безопасности ИИ с самыми высокими значениями p(doom), всё больше тревожил тот, казалось, безоблачный оптимизм, с которым Альтман и остальная компания смотрели на эту историю. Однажды OpenAI устроила и показ «Аполлона-11» – документального фильма об американской программе «Аполлон», доставившей первого человека на Луну, ещё одной любимой аналогии Кремниевой долины. «Зачем вообще говорить о Манхэттенском проекте, если можно просто сказать „программа «Аполлон»“? Зачем тащить за собой этот багаж?» – говорит один крайний думер.

Одна сцена из «Оппенгеймера» особенно ему запомнилась: момент перед испытанием «Тринити», первым в истории взрывом атомной бомбы, когда Оппенгеймер в исполнении Киллиана Мёрфи подсчитывает, что вероятность того, что взрыв уничтожит мир, «близка к нулю».

– Близка к нулю? – недоверчиво переспрашивает генерал-майор Лесли Гровс в исполнении Мэтта Деймона.

– А чего вы хотите от одной теории? – говорит Оппенгеймер.

– Ноль был бы неплох, – говорит Гровс.

«Это похоже на то положение, в котором находимся мы, – говорит крайний думер. – Мы пока никак не можем ничего рассчитать насчёт этих ИИ».

* * *

По мере того как OpenAI продвигала свои исследования, аналогия с Манхэттенским проектом для некоторых стала приобретать новый смысл.

Поскольку с исчерпанием данных и вычислений темпы развития больших языковых моделей замедлялись, подразделение Research сильнее развернулось к разработке ИИ-агентов. Эта набиравшая популярность в отрасли идея означала возвращение к спору между гипотезами «чистого языка» и «заземления». Чистый язык подходил к пределу своих возможностей, как и сочетание языка и зрения. Многим в сообществе ИИ казалось, что следующий этап развития, скорее всего, должен прийти от агентов, способных действовать в реальном мире и получать обратную связь от своей среды. Внутри OpenAI такая способность виделась и как способ получить конкурентное преимущество. ИИ-помощник, с которым можно поболтать, – это хорошо, а тот, что может автоматизировать сложные задачи, например отправлять электронные письма или программировать сайты, – ещё лучше. Это было не только крайне важно коммерчески, но и могло ускорить собственное продвижение компании.

Самой амбициозной из этих работ в подразделении Research был AI Scientist («ИИ-учёный») – попытка создать агента, который самостоятельно вёл бы научные исследования. Новых «знаний», то есть данных, которые можно выкачать из интернета или учебников, оставалось мало, и исследователи проекта возлагали большие надежды на то, что автономный «учёный» сможет порождать собственные знания, ставя эксперименты. Команда возникла из слияния прежней команды генерации кода, работавшей над Codex, с другой командой, пытавшейся решить задачу рассуждения, используя большие собрания математических задач и их решений как структурированный набор данных, чтобы обучить модель пошаговой логике. Обе способности – программирование и решение математических задач – казались хорошими строительными блоками для проведения экспериментов и анализа данных.

Руководили проектом Якуб Пахоцкий и Шимон Сидор, сосредоточенные на создании не просто автономного учёного, а, как того хотел Альтман, именно автономного исследователя ИИ, который помог бы OpenAI резко ускорить продвижение в ИИ.

Проект одновременно до крайности воодушевлял и до крайности тревожил верующих в AGI: если AI Scientist добьётся успеха, AGI наверняка появится быстрее. А значит, сокращался путь – либо к утопии, либо к уничтожению человечества. Как и Суцкевер, некоторые исследователи стали между делом упоминать в разговорах «бункер» и даже представлять себе что-то вроде Лос-Аламоса: где-нибудь в глухом уголке американской пустыни элитная команда исследователей ИИ будет жить и работать на охраняемых объектах, защищённая от внешних угроз.

В 2023 году к этим угрозам, по их мнению, относились уже и целенаправленные атаки, и сам вышедший из-под контроля AGI. Команда безопасности выложила в Slack черновик модели угроз для OpenAI, заметно

повзрослевший со времён, когда руководство спорило, насколько усиливать защиту организации. В черновике было три категории угроз: иностранные государственные субъекты, конкуренты и люди с идеологической мотивацией.

В третьей категории черновик давал ссылку на статью Элиезера Юдковского – крайнего думера и одного из лидеров сообщества безопасности ИИ, который придумал и сделал популярным выражение «дружественный ИИ» для хорошо согласованных систем и написал любимый многими фанфик «Гарри Поттер и методы рационального мышления». Этот роман с продолжением в 122 главах и больше чем 660 000 словах переосмысляет историю Гарри, который действует в мире волшебников как хорошо выученный рационалист. Для многих он стал воротами в эффективный альтруизм, а через него – в более широкую думерскую идеологию. Юдковский был и одним из основателей блога LessWrong, главной площадки, где исследователи безопасности ИИ создавали сообщество и распространяли идеи безопасности ИИ; там он со всё большим паникёрством призывал полностью остановить развитие ИИ, а его p(doom) взлетел до 95 процентов. В марте 2023 года он написал статью для журнала Time, где рассказывал о горечи, с которой смотрел, как у его дочери выпал первый зуб, и гадал, успеет ли она вырасти. Он предлагал план принудительной остановки развития ИИ: отключить все крупные кластеры GPU, отслеживать продажи GPU и при необходимости наносить по «вышедшему из-под контроля центру обработки данных» авиаудары. Черновик модели угроз OpenAI ссылался на эту статью как на пример идеолога, призывающего к насилию.

Небольшую группу сотрудников OpenAI, разделявших менее воинственные взгляды Юдковского, глубоко возмутило, что черновик выделяет его, но не упоминает сам недружественный ИИ. После отзывов команда безопасности добавила четвёртую категорию угроз: несогласованные системы AGI.

* * *

Пока OpenAI взлетала к новой известности, её совет директоров терял членов и не находил им замены. С начала 2023 года он один за другим лишился трёх независимых директоров – отчасти из-за лихорадочной гонки за создание и коммерциализацию технологий генеративного ИИ, которую ChatGPT разжёг во всей отрасли.

Первым в феврале, проведя в совете пять лет, ушёл Рид Хоффман – из-за конфликта интересов. Годом раньше он вместе с Мустафой Сулейманом, уже покинувшим DeepMind сооснователем этой компании, основал стартап Inflection, быстро превращавшийся в прямого конкурента OpenAI.

Через месяц за Хоффманом последовала Шивон Зилис, доверенная помощница Маска и директор Neuralink, которая после ухода Маска продолжала присматривать за OpenAI от его имени и официально вошла в совет в 2020 году. Некоторые в руководстве давно опасались, что Зилис будет передавать Маску конфиденциальную информацию компании, но она обещала ставить

обязательства конфиденциальности перед OpenAI выше верности её отвергнутому бывшему сопредседателю. Эта позиция стала весьма сомнительной, когда в июле 2022 года стало известно, что у неё родились близнецы от Маска, о чём она не сообщила коллегам по совету. И всё же Альтман по непонятным другим членам совета причинам пытался сохранить её в совете и в октябре 2022 года даже спрашивал у неё совета, как быть с явным раздражением Маска из-за стремительно растущей оценки OpenAI, которая тогда, по имеющимся сведениям, приближалась к 20 миллиардам долларов. «Это заманивание и подмена», – написал тогда Маск Альтману, напомнив о своём существенном вкладе в первоначальное финансирование OpenAI. В марте 2023 года пребывание Зилис в совете наконец стало невозможным: Маск зарегистрировал новое ИИ-предприятие, xAI, ещё одного прямого конкурента OpenAI.

Третьим ушёл Уилл Херд, бывший конгрессмен-республиканец от Техаса и бывший офицер ЦРУ, вошедший в совет в 2021 году. Объявляя о назначении Херда, Альтман сказал сотрудникам, что важно иметь человека, уравновешивающего либеральный крен Кремниевой долины. Затем он устроил встречу, на которой все, у кого были сомнения, могли задать Херду любые вопросы. Сотрудники не стеснялись и устроили Херду допрос о его взглядах на разные вопросы, включая Дональда Трампа. В июне 2023 года Херд расстался с OpenAI, чтобы полностью сосредоточиться на своей кампании за пост президента США. К октябрю он снимется с гонки.

Помимо Альтмана, Брокмана и Суцкевера, в совете осталось лишь трое независимых членов: сооснователь и CEO Quora Адам Д'Анджело, специалист по робототехнике Таша Макколи и исследовательница CSET Хелен Тонер.

Из этой тройки первым в совет вошёл Д'Анджело. Одноклассник Цукерберга по школе-пансиону Phillips Exeter, он два года был техническим директором Facebook, прежде чем основать Quora. В 2014 году Quora попала в первую волну YC при президентстве Альтмана; в 2017 году, за год до назначения Д'Анджело в совет, Альтман увеличил инвестиции YC в Quora, став одним из ведущих инвесторов раунда финансирования на 85 миллионов долларов. В объявлении Альтман хвалил Д'Анджело как одного из «самых умных CEO в Кремниевой долине». «И он очень ориентирован на долгую перспективу, а это в наши дни стало в технологических компаниях редкостью», – говорил Альтман.

Макколи вошла в совет позже, в 2018 году. Предпринимательница, одна из основательниц стартапа в области роботов телеприсутствия, руководившая компанией по трёхмерному моделированию городов, она познакомилась с Альтманом через своего наставника Алана Кея и была знакома также с Холденом Карнофски. Макколи пользовалась большим уважением в сообществе безопасности ИИ и будет входить в совет некоммерческой исследовательской организации по безопасности ИИ Centre for the Governance of AI, а какое-то время – и в совет фонда Effective Ventures Foundation, британской организации,

курирующей популярный подкаст ЭА 80,000 Hours. Карнофски, тогда входивший в совет OpenAI, выдвинул её в рамках поиска новых независимых директоров, когда OpenAI готовилась перейти на структуру с ограниченной прибылью.

Тонер стала последним пополнением совета в середине 2021 года – тоже по рекомендации Карнофски. На этот раз он предлагал кандидатов себе на смену: его трёхлетний срок подходил к концу, а тут ещё появилась Anthropic. Прежде чем стать признанным специалистом по Китаю и новым технологиям, Тонер работала с Карнофски в GiveWell и Open Philanthropy. К проблемам безопасности ИИ она пришла через движение ЭА, но со временем постепенно от него отошла. В своей самой популярной публикации на форуме ЭА ещё до краха FTX она отмечала, что движение становится всё догматичнее и социально замкнутее, и писала, что всё больше «склоняется к разочарованию в ЭА». В его кругах её по-прежнему высоко ценили, и она продолжала посвящать себя более широкому сообществу безопасности ИИ, входя вместе с Макколи в совет Centre for the Governance of AI.

* * *

К концу лета 2023 года совет уже несколько месяцев был в тупике, не в силах решить, кого назначить новыми независимыми директорами. В рамках усилий по усилению надзора после демонстрации GPT-4, а тем более после запуска ChatGPT, Макколи почти год вела работу – в том числе беседовала с сотрудниками и заинтересованными лицами за пределами компании, – чтобы сформулировать, как должны выглядеть обновлённый совет и более профессиональные механизмы надзора.

В ходе этой работы совет, включая Альтмана, единодушно согласился, что, учитывая, по их мнению, растущие ставки в связи с возможностями компании, следующий независимый директор должен иметь глубокий опыт в безопасности ИИ. Затем они несколько месяцев составляли список кандидатов и побеседовали с пятью из них, в том числе с Дэном Хендриксом, центральной фигурой сообщества думеров, руководителем Центра безопасности ИИ в Беркли и единственным советником в xAI Маска. Но когда независимые члены совета попытались двинуться дальше и выбрать одного из пятерых, у Брокмана и Суцкевера нашлись разные возражения против проверенных кандидатов. Альтман, как всегда, держался скромно: ни одному варианту прямо не возражал, но и ни одного не продвигал. Несколько раз он предлагал и новых кандидатов, у которых была одна общая черта: все они были вплетены в его сеть связей и – финансово или иначе – находились в сфере его влияния.

Когда дело дошло до создания новых механизмов надзора, включавших разные каналы, через которые совет лучше видел бы практики компании в области безопасности и защищённости, у независимых директоров тоже осталось похожее ощущение, что для Альтмана они не приоритет. В начале

пребывания Макколи в совете Альтман назначил её связующим звеном и защитником интересов сотрудников; после этого она регулярно встречалась с сотрудниками в приёмные часы. Однажды она привела на выездную встречу компании своего мужа, актёра Джозефа Гордона-Левитта, который внимательно слушал технические доклады. Но во время пандемии эти встречи сошли на нет. Потом Макколи продолжала проводить некоторые регулярные встречи, но открытые приёмные часы так и не возобновились.

Без систематического способа общаться с сотрудниками информация о происходящем в компании доходила до независимых директоров через их собственные личные связи в широком сообществе безопасности ИИ и технологическом сообществе. Полагались они и на самого Альтмана как на канал, через который можно отслеживать важную информацию. Всё сильнее их тревожило то, как часто слова Альтмана расходились с другими рассказами, которые они слышали. Если Альтман неизменно рисовал радужную картину, то от своих источников директора всё чаще получали сведения о разных проблемах: о неподготовленности компании перед ChatGPT и о серьёзных потрясениях после него, о сохранявшихся опасениях насчёт безопасности ИИ вокруг выпуска GPT-4, о беспрецедентной скорости, с которой OpenAI неслась к запуску новых продуктов, не решив многих своих проблем.

Один случай показался особенно вопиющим. В конце 2022 года совет провёл выездное заседание – первое из тех, что должны были стать ежегодными, – на котором Альтман подчёркивал надёжность протоколов безопасности и тестирования, введённых OpenAI вместе с Советом по безопасности развёртывания для оценки развёртывания GPT-4. После заседания один из независимых директоров разговорился с сотрудником, и тот заметил, что протоколы DSB уже были нарушены. Microsoft провела ограниченный выпуск GPT-4 для пользователей в Индии без одобрения DSB. Хотя Альтман провёл с советом целый день в одной комнате на выездном заседании, он ни разу не сообщил о нарушении.

У независимых директоров не было оснований считать, что в открытый доступ попало что-то небезопасное, но их встревожило то, с каким, судя по всему, пренебрежением Microsoft нарушила протокол, а Альтман обошёл это молчанием. Модели OpenAI стремительно развивались и, с точки зрения безопасности ИИ, считали они, вскоре могли перейти черту, за которой подобное нарушение могло бы привести к потенциально катастрофическим, если не экзистенциальным, последствиям. По их мнению, снисходительность Альтмана к нарушению Microsoft создавала опасный прецедент того, как он может относиться к процедурам безопасности ИИ при выпуске моделей, когда ставки вырастут.

Тем временем были и другие тревожные примеры поведения Альтмана. В марте 2023 года он написал совету письмо, не поставив в копию Д'Анджело, и заявил, что, по его мнению, Д'Анджело пора уйти. То, что Quora разрабатывает

собственного чат-бота Poe, доказывал Альтман, создаёт конфликт интересов. Заявление показалось внезапным и сомнительным по мотивам. Тонер, Макколи и Д'Анджело в разное время задавали Альтману неудобные вопросы – о практиках безопасности OpenAI, о силе некоммерческой организации или на другие темы. Поскольку союзников у Альтмана в совете становилось всё меньше, этой тройке казалось, что он ищет повод вытеснить одного из них. В ответ на письмо Альтмана один из независимых директоров возразил: Poe – умеренный конфликт интересов по сравнению с тем, что Альтман долго терпел у Хоффмана и Зилис. Предложение Альтмана не прошло; Д'Анджело остался.

Вскоре после этого Д'Анджело на одном званом ужине услышал, что OpenAI Startup Fund устроен странно. Он давал тем, кто вложился в фонд, ранний доступ к продуктам OpenAI – своего рода привилегию, которая должна была предназначаться собственным инвесторам OpenAI. Услышав об этом во второй раз, независимые директора потребовали от Альтмана документы о структуре фонда. Когда Альтман наконец их передал, директора обнаружили, что структура фонда не просто странная: юридически фонд принадлежал Альтману, хотя должен был принадлежать OpenAI.

Для независимых директоров каждый случай складывался в единую тревожную картину: шаг за шагом Альтман пытался затуманить им обзор и маневрировал так, чтобы совет никогда не смог его сдержать. Годами Альтман превозносил способность совета служить ему противовесом и даже уволить его как важнейший механизм управления OpenAI. Более того, сейчас он трубил об этом по всему миру, чтобы завоевать доверие общества и правительств.

К осени боевой дух независимых директоров упал до нового минимума: им никак не удавалось сколько-нибудь продвинуться в переговорах о новом члене совета. С каждым месяцем эти пробелы в управлении становились всё неотложнее. OpenAI готовилась обучать GPT-5 и продвигалась в работе над AI Scientist. И вот, как гром среди ясного неба, в начале октября Тонер получила неожиданное письмо: Илья Суцкевер хотел поговорить.

Глава 14

Избавление

В те недели, что предшествовали обращению Суцкевера к Тонер, Альтман разбирался с кризисом в пиаре. Годами отчуждение его сестры и её уход в секс-работу не попадали в СМИ, но теперь её история наконец вырвалась наружу.

25 сентября 2023 года Элизабет Вейл, обозреватель журнала New York, опубликовала портрет Сэма Альтмана, в котором впервые в массовой прессе упоминалось о существовании Энни. Вейл сопоставила подробности жизни Энни – её постоянные проблемы со здоровьем, жизнь в тяжёлой финансовой нужде без надёжного жилья – с образом жизни Сэма с его многомиллионными домами и роскошными машинами. «Энни Альтман? – писала Вейл в своём материале. – Читатели блога Альтмана, его твитов, его манифеста „Справочник стартапа“ (Startup Playbook), а также сотен статей о нём хорошо знают Джека и Макса… В публичной жизни Сэма Энни не существует. Ей никогда не было места в этом клубе».

Альтман знал, что эти подробности всплывут. Перед публикацией, пришедшейся на Йом-Киппур, иудейский праздник покаяния, Вейл дала OpenAI возможность прокомментировать её материал; New York также работал с компанией в рамках проверки фактов. Задача организовать ответ на запрос о комментарии и попытаться держать историю под контролем легла на Ханну Вонг, ставшую вице-президентом OpenAI по коммуникациям, которая вдруг оказалась посредницей между журналом и семьёй Альтмана и задавалась вопросом, действительно ли это входит в её обязанности.

В последний день перед выходом материала, когда стало ясно, что подробности об Энни будут опубликованы, Энни получила письмо от Сэма. «привет энни. в духе того, что почти наступил йом-киппур, я хотел извиниться и попросить у тебя прощения за одну вещь», – писал он. Когда Энни просила о поддержке, он чувствовал себя между двух огней: с одной стороны, хотел положиться на решение их мамы и соглашался с остальной семьёй, что Энни должна научиться финансовой самостоятельности, с другой – чувствовал, что Энни нужна медицинская помощь и что ей трудно справляться с жизнью. «и всё же я принял неправильное решение и должен был просто просто [sic] продолжать тебя поддерживать; искренне прошу прощения», – писал он.

Резкий поворот после восторженного в остальном приёма, которым публика встречала Альтмана с выхода ChatGPT, и то, что болезненные, как он считал, частные дела его семьи выплеснулись на всеобщее обозрение, тяжело на него давили. Последствия статьи сделали всё ещё хуже. После её выхода всплыли и стали разлетаться по сети старые твиты Энни с обвинениями в том, что Сэм по-разному над ней издевался. Но как ни тяжело и мучительно это было для Сэма,

для Энни это стало невероятным освобождением от боли. Годами, помимо нараставшего стресса из-за проблем со здоровьем и отсутствия стабильности, ей казалось, будто она кричит в пустоту и её вычеркнули как незначащую.

В 2024 году я связалась с Энни, чтобы лучше понять её версию событий. Я обратилась напрямую и к её матери Конни Гибстайн, к братьям Максу и Джеку, а также к Сэму – через команду коммуникаций OpenAI, – чтобы узнать их версию. Энни охотно шла на сотрудничество, надеясь получить наконец площадку, чтобы рассказать о личном опыте, который считала крайне важным для понимания нравственного облика Сэма. Она предоставила обширную переписку с семьёй, медицинские документы о физическом и психическом здоровье почти за всё время с детства до взрослой жизни и другие подтверждающие материалы, по которым прослеживались её разрыв с семьёй и ухудшение её жизненных обстоятельств.

Гибстайн прислала короткое заявление, в котором подчеркнула любовь семьи к Энни и тревогу за её благополучие и вместе с тем отвергла утверждения Энни, назвав их «ужасными, глубоко разрывающими сердце и неправдой». От более подробного разговора и ответов на детальные вопросы о её взгляде на рассказ Энни Гибстайн отказалась. На мои дополнительные просьбы предоставить документы в подтверждение её собственных утверждений она не ответила. Макс и Джек не ответили. OpenAI не стала комментировать подробности.

В январе 2025 года, после того как Энни подала иск против Сэма, он, Гибстайн, Макс и Джек выступили с более жёстким публичным опровержением обвинений Энни, назвав её психически неустойчивой и необоснованно требовательной в отношении денег, а её утверждения – «радикально менявшимися со временем». «Особенно мучительно, когда она отказывается от общепринятого лечения и набрасывается на членов семьи, которые искренне пытаются помочь», – говорилось в заявлении.

Из моих разговоров с Энни и предоставленных ею документов сложилась сложная картина того бурного пути, который привёл её к публичным обвинениям в Twitter, а затем к иску, и к тем подробностям своей жизни, которыми она поделилась для материала Вейл. У меня не было доступа к полному обоснованию многих решений её семьи, а истинность некоторых обвинений Энни, в частности в предполагаемом сексуальном насилии Сэма над ней в детстве, установить невозможно, но её история стала для меня микрокосмом многих тем, определяющих более широкую историю OpenAI. Она помогла мне и утвердиться в понимании того, насколько OpenAI – отражение и продолжение человека, который ею руководит. Настойчивые попытки Энни добавить в летопись свою точку зрения быстро стали делом компании. Публикации об Энни задевали Сэма, пожалуй, сильнее любых других историй, и Ханну Вонг как главу коммуникаций OpenAI втягивали в его усилия сдержать их распространение. Делом компании это стало и в другом смысле: когда

обвинения Энни впервые вышли на первый план, на них обратили внимание другие руководители OpenAI.

* * *

Во многом Энни и Сэм, между которыми девять лет разницы, удивительно похожи. Многие слова, которыми описывают Сэма близкие ему люди, подходят и к Энни: она прекрасно слушает; помнит мельчайшие подробности о других; очень дурашлива, чрезвычайно щедра и быстро завоёвывает доверие людей.

Как и Сэм, она отлично училась. Единственная, кроме Сэма, из детей Альтманов, окончившая Burroughs, она произвела впечатление на своего учителя физики Джеймса Робла, который и десять с лишним лет спустя с теплотой вспоминает её как талантливую ученицу с солнечным характером. Учиться она поехала в Университет Тафтса, специализировалась на биопсихологии, а в качестве второй специальности – на танце. Окончив его в 2016 году, она прошла подготовительную программу для поступления в медицинскую школу в расчёте когда-нибудь туда поступить; переехала в район залива на должность исследователя в лаборатории нейронаук Калифорнийского университета в Сан-Франциско. Иными словами, пока её жизнь не пошла наперекосяк, она была типичной честолюбивой, хорошо образованной молодой женщиной с множеством возможностей.

Но после череды неожиданных испытаний обстоятельства Энни начали рушиться. Ещё в университете, в 2014 году, у неё диагностировали ахиллов тендинит – воспаление сухожилия от перегрузки – и костный шип, и ей пришлось носить ортопедический сапог. Это стало первым пунктом в растущем списке физических недугов, которые будут мучить её хронической болью, а в периоды обострений сильно ограничивать подвижность и сказываться на качестве жизни. За шесть лет Энни пережила повторяющийся тонзиллит; повторяющиеся тазовые боли; новые обострения тендинита, из-за которых ей снова приходилось носить ортопедические сапоги, – воспаление распространилось за пределы правой лодыжки и временами мешало даже недолго стоять; и растущее число кист яичников, завершившееся диагнозом «синдром поликистозных яичников» (СПКЯ), из-за которого она порой просыпалась по ночам в мокрых от пота простынях – распространённый симптом.

Посреди всех этих проблем со здоровьем на неё обрушился ещё один парализующий удар: 25 мая 2018 года её отец, самый близкий ей человек в семье, умер от внезапного инфаркта.

Всю жизнь Энни боролась с проблемами психического здоровья. В юном возрасте ей диагностировали генерализованное тревожное расстройство и обсессивно-компульсивное расстройство, и почти десять лет она принимала «Золофт», пока в университете психиатр не помог ей постепенно от него отказаться. До смерти отца ей было гораздо лучше. Она освоила

непсихиатрические способы поддержания психического здоровья, в том числе медитацию, и искала, как справиться с исходным воспалением, которое, казалось, лежало в основе её повторяющихся физических недугов, полностью перейдя на здоровое органическое питание из цельных продуктов. Её путь к здоровью пробудил новый интерес к альтернативной медицине, и она прошла обучение на преподавателя йоги. Она отдавалась давней любви к искусству и написала черновик книги, которую назвала The Humanual («Человекоруководство»), где собрала свои размышления о жизни и о том, как быть хорошим человеком. Смерть отца её сломила. Её психическое состояние пошло вниз по спирали, и, похоже, это ознаменовало ускоренное ухудшение и её физического здоровья. К концу 2018 года походы к разным врачам и специалистам стали рутиной.

Сэм тоже публично говорил, что смерть отца была худшим моментом в его жизни. Это случилось всего через несколько месяцев после того, как Маск ушёл с поста сопредседателя OpenAI и Сэм взял руководство на себя. Близкие Сэму люди говорят, что эта утрата выбила его из колеи. Какое-то время он вёл себя непредсказуемо и с трудом справлялся. Эта смерть явно стала переломным моментом в отношениях Энни и Сэма, как и в её отношениях с остальной семьёй. Её связь с родными, особенно с мамой, и без того была натянутой из-за разногласий по поводу разных её жизненных решений, говорит Энни, – в том числе отказа от традиционной медицинской карьеры и решения перестать принимать «Золофт» в пользу естественных способов справляться с собой. Отец был её единственным безоговорочным сторонником. Без него она оказалась в изоляции от семьи и в то же время жаждала её принятия и поддержки. Вместо этого после стычек из-за денег отношения дошли до точки разрыва.

* * *

После смерти отца в 2018 году Энни жила в Лос-Анджелесе. Она работала на полставки литературной ассистенткой, а затем в магазине, торгующем марихуаной. По словам Энни, по страховке жизни отца ей досталось около 100 000 долларов. Она ещё глубже погрузилась в любовь к искусству и выступлениям, стремясь сделать их серьёзной карьерой: ходила на курсы комедии, выступала на открытых микрофонах, завела подкаст, переписала первый черновик своей книги в моноспектакль, который переименовала в The HumAnnie. К середине 2019 года деньги заканчивались – большая часть, говорит она, ушла на разные вложения в творчество, высокую арендную плату, медицинскую страховку из собственного кармана и накапливавшиеся медицинские расходы. на медицинскую визуализацию, физиотерапию, психотерапию, поездки на Lyft и Uber к врачам.

Для Энни началом конца её отношений с семьёй стало то, что произошло дальше. В мае 2019 года, когда хроническая боль не отступала, она узнала, что отец оставил ей свой пенсионный счёт 401(k). Она ушла из магазина и составила

полугодовой план: потратить этот финансовый запас – чуть больше сорока тысяч долларов – на здоровье и на то, чтобы достаточно раскрутить свои творческие начинания и, если повезёт, получить устойчивый источник дохода. Но вскоре мать, которая как пережившая супруга сохраняла право распоряжаться пенсионными средствами отца, сообщила Энни, что денег она на самом деле не получит. Лучшая налоговая стратегия, писала Гибстайн Энни по электронной почте, – оставить средства на её, Гибстайн, имя на счёте с отсроченным налогообложением, который перейдёт к Энни через траст и доступ к которому она получит в пятьдесят девять с половиной лет. «Мы все хотим для тебя лучшего и считаем, что лучшее для тебя (и для всех) – финансовая независимость, которая приносит удовлетворение в долгосрочной перспективе, личностный рост и уверенность, – писала Гибстайн. – Поэтому мы хотим прояснить, что не будем поддерживать тебя финансово, если/когда твоё нынешнее наследство от папы закончится». Письмо было подписано: «Мама, Сэм, Макс, Джек».

Энни была в хрупком состоянии. В записях её психотерапевта за тот же период говорится, что шаг, предпринятый семьёй в надежде поставить её на ноги, наоборот, ухудшил её состояние. В декабре 2019 года её банковский счёт ушёл в минус. Сэм в том же году добился первой инвестиции Microsoft в OpenAI на миллиард долларов, и компания полным ходом обучала GPT-3. Напуганная и одинокая, Энни зашла на эскорт-сервис SeekingArrangement и по видеосвязи показала мужчине грудь за сумму, которой хватило, чтобы вывести счёт из минуса.

Энни говорит, что никогда раньше не просила у матери или братьев финансовой поддержки. Она не рассчитывала, что может опереться на их богатство, но и не думала, что они оставят её без подстраховки, если с ней действительно случится беда. С конца 2019 до середины 2020 года Энни несколько раз обращалась к семье за финансовой помощью, в том числе после того, как пандемия добавила ещё один слой стресса и неопределённости. После того как Сэм и мать вместе с ней посетили два сеанса семейной терапии, они согласились покрывать её расходы часть года.

Из её переписки с семьёй видно, что родные опасались, как бы деньги не способствовали вредному поведению, и считали, что лучший способ помочь ей восстановить психическое здоровье – и дальше поощрять её финансовую независимость. И заявление Гибстайн, присланное мне, и публичное заявление семьи говорили, что их действия определялись советами профессионалов о том, как лучше всего поддержать Энни.

В мае 2020 года, когда финансовая поддержка семьи подходила к концу, а Энни по-прежнему было тяжело, она попросила ещё помощи, чтобы оплачивать физиотерапию и психотерапию, которые, как она им сказала, стоили соответственно 45 и 15 долларов за сеанс. Ей отказали, сказав, что, по их мнению, расходы на июнь ей следует покрыть самой, в том числе за счёт возвращаемого ей залога за жильё. Она собрала свои немногочисленные пожитки и уехала на

Гавайи, где была, когда отец в последний раз навещал её за несколько месяцев до смерти. Она нашла работу на ферме в обмен на жильё – с лёгкими физическими обязанностями вроде прополки и посадки. Сэм написал ей, спрашивая новый адрес. Через восемь месяцев после смерти отца он попросил каждого из братьев и сестру прислать маме по пряди волос, чтобы смешать их с прахом отца и сделать каждому по бриллианту. Они «будут в основном из углерода папы, но в них будет и немножко каждого из нас», – писал он. Бриллиант Энни был готов, и Сэм хотел его отправить.

Для Энни это письмо Сэма было как пощёчина. Ей не нужен был дорогой бриллиант; ей нужна была его помощь, чтобы у неё гарантированно были еда и жильё. Она не могла понять, то ли Сэм просто не осознаёт, то ли ему безразлично, насколько тяжёл её кризис. Пропасть между ней, Сэмом и остальной её богатой семьёй казалась непреодолимой. С огромной болью и горем она перестала с ними общаться.

* * *

Следующие три года, до того как с ней связалась Элизабет Вейл, жизнь Энни катилась на самое дно. У неё не было надёжного жилья, надёжной еды, надёжного здоровья; чтобы оплачивать счета, она занялась виртуальной и реальной секс-работой. В публичном заявлении семья говорила, что пыталась «многими способами поддержать Энни и помочь ей обрести стабильность». В пересказе Энни одна из таких попыток поддержки пришлась на весну и лето 2021 года, когда Сэм попытался восстановить с ней связь. У них было три телефонных разговора, вспоминает она, в которых он говорил, как сильно её любит, и предложил купить ей дом. Энни и семья описывают это предложение по-разному. Энни говорит, что речь шла не о том, чтобы дом принадлежал ей, а о том, чтобы она в нём жила, – она понимала, что такая договорённость должна была помешать ей продать недвижимость, и опасалась, что это может стать ещё одним способом, через который Сэм и мать будут навязывать ей свои взгляды на её здоровье и карьеру. Гибстайн в заявлении, присланном мне, сказала, что семья предлагала Энни дом в собственность, но не ответила на мои просьбы предоставить подтверждающие документы. В публичном заявлении семьи говорилось, что они предлагали купить Энни дом через траст, чтобы у неё было где жить, но без возможности сразу его продать. В конце концов Сэм и Энни пришли к другому соглашению: год он будет платить за её жильё напрямую арендодателю. Летом 2022 года, когда она так и не восстановила общение, платежи прекратились.

История Энни углубляет противоречивые портреты, которые рисуют люди, говоря о Сэме. Он одновременно щедр и своекорыстен, покладист и угрожающ, благодетель для очень многих людей и источник большой личной боли для других. Человек, который на публике излучает искренность и альтруизм, но своими поступками за закрытыми дверями обнаруживает более сложный расчёт.

Человек, способный и дать, и отнять, у которого многие остаются с впечатлением, что они – часть большой шахматной партии, всю доску в которой видит лишь он, а цель эндшпиля – сохранить свою власть короля.

История Энни ставит под сомнение и великий нарратив, который рисуют Сэм и другие руководители OpenAI, – о том, что ИИ откроет дорогу в мир изобилия. Альтман говорил, что ожидает, что ИИ покончит с бедностью. Брокман в историях о своём друге и о жене Анне повторял, что AGI резко улучшит здравоохранение. Суцкевер говорил, что он даст невероятно эффективную и баснословно дешёвую психотерапию. И всё же на фоне реальной жизни работников в Кении, активистов в Чили и опыта собственной сестры Альтмана, на которую обрушились все эти проблемы, эти мечты звучат пусто.

При всех скачках в возможностях ИИ ни один из них не помог облегчить хоть сколько-нибудь отчаяние Энни. Скорее ИИ, возможно, лишь сильнее загнал её в ловушку. Она не хотела заниматься секс-работой. Это был, по её словам, «план Z». Когда хроническая боль усилилась и даже работа на ферме стала слишком тяжёлой, она сначала попыталась зарабатывать на своём искусстве цифровыми способами. Она продолжала вести подкаст, держала магазин на Etsy и страницу на Patreon, но они не приносили денег даже на оплату телефона. Происходило что-то странное, что она со временем фиксировала скриншотами: во всех её соцсетях её почти никто не видел. Порой она замечала, как из отзывов на её подкаст в приложении Apple таинственно исчезают целые куски, из-за чего его труднее было найти. Как минимум дважды, и в Instagram, и на YouTube, у неё набирались просмотры, а потом необъяснимо пропадали. Бывший специалист по данным из Facebook и два эксперта по технологиям и секс-работе говорят, что, возможно, уже самая первая учётная запись Энни на SeekingArrangement в декабре 2019 года могла ограничить её охват в сети: технологические платформы отслеживают секс-работников и подвергают их теневым банам с помощью автоматических систем, помечая их устройства, почтовые адреса, банковские счета и другую информацию, связывающую их присутствие в сети, – даже для профилей, никак не связанных с секс-работой. Не видя другого выхода, Энни вернулась на SeekingArrangement и завела учётную запись на OnlyFans, ещё сильнее запутав своё присутствие в сети и доступ к экономическим возможностям в паутине алгоритмической модерации.

Нейли Мессершмидт, бывшая руководительница в технологической отрасли, работавшая в таких компаниях, как Sony, а теперь руководящая направлением оздоровления в сети органических ферм, познакомилась с Энни вскоре после её переезда в район залива в 2016 году; какое-то время, увлёкшись здоровым питанием, Энни работала в этой сети. С годами, по мере того как Энни порывала с родной семьёй, Мессершмидт стала для неё кем-то вроде матери. «Сэм принёс ИИ в мир точно так же, как на самом деле обращался со своей младшей сестрой, – говорит Мессершмидт. – Он там процветает, а его сестра проваливается в пропасть».

* * *

В конце 2020 и начале 2021 года, после того как Энни отдалилась от семьи, у неё начались мучительные флешбэки о сексуальном насилии в детстве. Записи о сеансах с июля 2021 по январь 2022 года у нового травматерапевта на Мауи фиксируют кризис, который пережила Энни из-за этих непроизвольных воспоминаний, включая мучительные вопросы о собственной идентичности. По итогам пятнадцати сеансов терапевт записала свои диагностические впечатления: генерализованная тревожность, ПТСР и личная история сексуального насилия в детстве.

В записях не упоминался конкретный насильник. Но примерно в то время Энни начала регулярно звонить Мессершмидт, которая сама была изнасилована в девятнадцать лет и ещё в начале общения с Энни в сети органических ферм заметила явные признаки человека, пережившего насилие. «Ей было не по себе рядом с некоторыми мужчинами, – вспоминает Мессершмидт. – Я узнаю человека, пережившего насилие, когда его вижу. Она, по сути, просила не участвовать в совещаниях, когда там были некоторые мужчины. Когда они были в комнате, она держалась в стороне».

В напряжённых, эмоциональных разговорах Энни рассказывала Мессершмидт о внезапных флешбэках: в этих детских воспоминаниях, как помнит её рассказ Мессершмидт, именно Сэм раз за разом забирался к ней в кровать, иногда вместе с Джеком, и домогался её.

Важно отметить, что спустя десятилетия часто трудно доказать, было ли предполагаемое сексуальное насилие в детстве, и установить его подробности. Психология знает одну распространённую закономерность, с которой сталкиваются некоторые жертвы насилия: мозг жертвы блокирует любые воспоминания о нём, пока какой-нибудь триггер – возможно, половое созревание, начало половой жизни или новые нежелательные сексуальные притязания – непроизвольно не вернёт их на поверхность, говорит психотерапевт, с которым я консультировалась. Тело помнит травму, даже если разум не помнит. По мере того как Энни глубже погружалась в секс-работу, на неё, возможно, внезапно обрушился поток триггеров.

Подробно рассказывая о том, как Энни переживала флешбэки, – по её рассказам Мессершмидт и отчасти по записям её травматерапевта, – я ставлю цель не установить, что именно произошло в детстве Энни, а воссоздать то, что она пережила и во что верила уже взрослой, – то, что в конечном счёте побудило её заговорить.

В ноябре 2021 года, когда Сэм оплачивал её жильё, Энни впервые публично написала о своих обвинениях. «Я пережила сексуальное, физическое, эмоциональное, словесное, финансовое и технологическое насилие со стороны моих биологических братьев, в основном Сэма Альтмана и отчасти Джека Альтмана, – написала она в Twitter. – Я твёрдо уверена, что эти люди совершали

насилие и над другими. Я ищу людей, которые присоединятся ко мне в стремлении к правосудию, безопасности для других в будущем и совместному исцелению».

Её пост не получил отклика. Это была никому не известная учётная запись с горсткой подписчиков. Известность же Сэма росла как никогда благодаря успеху OpenAI с GPT-3 и её последнему запуску, GitHub Copilot. Вместо этого на Энни набросились тролли, заявляя, что на самом деле она брату не родственница.

В следующие месяцы с Энни связались двое журналистов, но она так и не решила, сколько им рассказывать. В сентябре 2022 года она решилась говорить открыто. Она снова написала в Twitter, назвав Сэма и Джека: «Сексуальное, физическое, эмоциональное, словесное, финансовое и технологическое насилие. Никогда не забыто».

Через десять месяцев, в июле 2023 года, после выхода ChatGPT и стремительного взлёта Сэма к мировой славе, Энни получила сообщение от Элизабет Вейл из New York.

* * *

За три месяца после выхода статьи в журнале New York доход Энни на OnlyFans вырастет больше чем в десять раз – примерно со 150 долларов в месяц до более чем 1500. За один месяц во время кризиса в совете он достигнет примерно 5500 долларов. Она перестала работать в эскорте, оставив только виртуальную секс-работу.

Семья – или, как называет их Энни, «её родственники» – настаивает, что на протяжении многих лет оказывала ей разные виды финансовой помощи. И действительно, план поддержки, о котором договорились после семейной терапии, и арендная плата, которую Сэм вносил с середины 2021 до середины 2022 года, – тому примеры. Но для Энни эти деньги были слишком малы и пришли слишком поздно, когда она уже была в таком отчаянии, что занялась секс-работой. Другие предложения поддержки, в том числе дом, как правило, сопровождались ограничениями или условиями, которые она, по её ощущению, не могла принять.

В июле 2023 года Сэм отправил Энни ещё одно сообщение с предложением денег, на первый взгляд без условий. Ранее тем летом им обоим написал человек с лихорадочным пылом интернет-сыщика и попросил Энни подробнее рассказать о её обвинениях. Энни ответила в этой переписке на троих подробным рассказом о том, как она прожила последние несколько лет. Сэм был в разгаре «Мирового турне», на взлёте. 5 июня, в день, когда он вместе с Суцкевером выступил в Тель-Авивском университете в Израиле и был очень хорошо принят, Энни была в его почтовом ящике – выдержанными, беспощадными фразами описывала жестокость своих последних лет.

Он ответил – что для него редкость, с заглавными буквами – 9 июля, когда «Мировое турне» официально закончилось. «Прости, что так долго не отвечал; мне понадобилось время, чтобы понять, что я хочу сказать, – писал он. – Я не хочу никаких постоянных отношений с тобой и уважаю то, что ты тоже не хочешь их со мной. Но я с радостью пришлю тебе денег и надеюсь, что ты справишься со своими проблемами со здоровьем». Он предложил возобновить их прежнюю договорённость, подробностей которой толком не помнил, или выплатить ей единовременную сумму. «Для ясности: я не согласен со многими утверждениями в этом письме, но, похоже, вступать в спор бессмысленно», – закончил он.

К тому времени Энни уже не хотела денег Сэма. Она просто хотела доступа к деньгам, которые оставил отец. На письмо Сэма она так и не ответила.

* * *

Помимо 401(k), отец оставил траст, которым распоряжалась её мать. В начале 2024 года, благодаря дополнительному доходу, Энни наймёт собственного юриста и узнает, что в 2023 году в траст отца поступили новые средства. Когда её история стала набирать известность, семья через юриста Гибстайн начнёт обсуждать ежемесячные выплаты Энни из траста без всяких условий. Впервые Энни почувствовала, что у неё есть настоящие рычаги в переговорах. В публичном заявлении семья сообщила, что рассчитывает оказывать Энни ежемесячную финансовую поддержку «до конца её жизни».

Благодаря новым выплатам Энни сняла первую за больше чем четыре года постоянную квартиру. Вскоре после этого она наняла ещё одного юриста, чтобы подготовить против Сэма иск о сексуальном насилии над ребёнком и подать его до своего тридцать первого дня рождения, 8 января 2025 года. В иске утверждалось бы – а семья решительно это отрицала бы, – что Сэм совершал над ней сексуальное насилие начиная примерно с её трёх лет и до тех пор, пока он не стал взрослым, а она ещё была несовершеннолетней; иск требовал возмещения ущерба свыше 75 000 долларов. В октябре 2024 года, после очередного вихря визитов к врачам, Энни наконец поставят диагноз, объясняющий её основное заболевание: гипермобильный синдром Элерса – Данло, то же генетическое нарушение подвижности суставов, что и у жены Брокмана.

* * *

Перед тем как история Энни вышла в статье журнала New York, Сэм начал говорить людям, что у его сестры пограничное расстройство личности. Это частное и деликатное дело, говорил он им. «Он обезвредил её рассказ ещё до публикации», – говорит один из этих людей.

Пограничное расстройство личности характеризуется серьёзными трудностями с регуляцией эмоций и может вести к напряжённым межличностным отношениям с крайностями идеализации и страхом быть брошенным. Энни говорит, что такой диагноз ей никогда не ставили, – и ни в

одной из записей терапевтов и медицинских документов, которыми она со мной поделилась, его нет. Есть лишь одна пометка её травматерапевта на Мауи от августа 2021 года о том, что она проверяет Энни на это расстройство, которое часто возникает после сексуального насилия в детстве. В своих окончательных диагностических впечатлениях терапевт упомянула историю сексуального насилия в прошлом, но не это расстройство.

Два психотерапевта, с которыми я говорила о пограничном расстройстве, тоже подчеркнули, что оно обычно проходит – само по себе или при правильном лечении, включающем медитацию и поведенческую терапию, помогающую человеку утвердиться в собственной ценности, – оба этих инструмента Энни охотно использует, – а не психотропные препараты, к которым её подталкивала Гибстайн и о которых, по-видимому, шла речь в заявлении семьи, когда там говорилось, что Энни «отказывается от общепринятого лечения». «Исследования о пограничном расстройстве личности очень обнадёживают; считается, что у этого диагноза хороший прогноз, – говорит Блейз Агирре, ассистент-профессор психиатрии Гарвардской медицинской школы, лечивший тысячи пациентов с пограничным расстройством, но не изучавший случай Энни конкретно. – Подавляющее большинство людей поправляются».

В апреле 2024 года о психическом здоровье Энни заговорит со мной и Ханна Вонг, которую вскоре повысят до директора OpenAI по коммуникациям. Полгода команда Вонг уверяла, что намерена организовать мне визит в офис, а также интервью с ключевыми руководителями и сотрудниками OpenAI, поскольку я раз за разом пыталась вступить с ними в контакт и услышать их точку зрения. Через пять месяцев они стали к этой идее охладевать. За десять дней до моего перелёта в Сан-Франциско, который я уже забронировала, чтобы побывать в штаб-квартире компании, они сообщили, что передумали: в офис я не приеду, и в работе над моей книгой они больше участвовать не будут.

Через несколько дней после начала поездки в Сан-Франциско, которую я всё же совершила, как и планировала, я сказала одному человеку, лично знавшему Сэма и Энни, что разговариваю с ней. На следующий день Вонг написала мне сообщение. «Слышала, вы в городе?» – спросила она: странная формулировка, если учесть, зачем я вообще устроила эту поездку. Мы встретились в Philz Coffee в Мишн-Бэй, недалеко от нового офиса, куда расширялась OpenAI. После путаного светского разговора и общих рассуждений о моей книге она перевела разговор на Энни.

«Не думаю, что скажу лишнее, если замечу, что у Энни проблемы с психическим здоровьем», – сказала Вонг. К этому моменту я ещё не обращалась в OpenAI по поводу Энни и в разговоре первой её не упоминала. «У Энни бывают хорошие дни и бывают очень плохие, – продолжала она. – И семья очень старается найти баланс между тем, чтобы защищать её, и тем, чтобы не потакать ей». Тут она для выразительности повторила свою мысль. «Обратите внимание: семья не выступила ни с каким публичным заявлением, опровергающим слова

Энни. Всё сводится к тому, чтобы защитить Энни». Она слышала также, что на некоторых факультетах журналистики даже обсуждали, этично ли было со стороны Вейл включать Энни в свой материал. Может быть, и мне стоит об этом подумать, сказала она.

Стало ясно, что именно это Вонг и хотела мне передать, когда связалась со мной. До того она ни разу не ответила напрямую на мою просьбу поговорить с ней о книге и общалась со мной только через заместителя. То, какое значение она, по-видимому, придавала тому, чтобы разобраться с историей Энни, показывало, как сильно эта история давила на Сэма и насколько тесно компания связана с личными делами Сэма.

Утверждения Сэма и Вонг о психическом здоровье Энни показались мне и ещё одной параллелью между её опытом и опытом множества других людей, оттеснённых на обочину или пострадавших от империй ИИ и их видения. С тех пор как Энни решилась рассказать свою историю, она столкнулась с той же пропастью во власти, с которой, как я видела, борются работники по разметке данных и активисты, выступающие против центров обработки данных. Её жизнь вращается вокруг того, чтобы прочёсывать и собирать как можно больше документов, лишь бы её хоть кто-то выслушал. Порой её охватывает тягостное чувство, что, сколько бы она ни говорила, весь мир каким-то образом в сговоре против неё. Это та же утрата свободы действий и тот же гнев, которые я видела на лицах людей по всему миру, когда они бросают так много из того немногого, что у них есть, на борьбу с нарративами империй – и потом смотрят, как люди, против которых они выступают, распоряжаются властью, позволяющей пускать в ход миллиарды долларов капитала, возводить огромную инфраструктуру, нанимать и увольнять десятки тысяч подрядчиков и несколькими мягкими словами – на мероприятии, в Конгрессе, главам государств, журналистам – сглаживать ропот протеста на пути своей воли.

В случае Энни после материала Вейл она уже не кричала в пустоту. Когда её твиты в октябре 2022 года начали разлетаться по сети, они привлекли внимание важного человека – Ильи Суцкевера, как раз когда тот разбирался с собственными сложными чувствами к Альтману и с тем, что считал у Альтмана устойчивой склонностью к психологическому насилию.

Часть IV

Глава 15

Гамбит

Через четыре дня после выхода материала Вейл об Альтмане в журнале New York и за четыре дня до сообщения Суцкевера Тонер независимая член совета встретилась с ещё одним руководителем OpenAI – техническим директором Мирой Мурати.

Мурати, родившаяся в Албании, с ранних лет научилась сохранять спокойствие посреди хаоса. В раннем детстве она застала муки перехода страны от тоталитарного коммунизма к либеральному капитализму. Перемены произошли так быстро, а финансовая система страны была так неразвита, что стремительно расплодились, а затем рухнули финансовые пирамиды, вызвав повсеместные беспорядки и насилие. После этих потрясений остались воронки от бомб, которые Мурати приходилось осторожно обходить по дороге в школу. Однажды учительница сказала ей: пока она готова приходить, учительница тоже будет приходить.

Родители Мурати преподавали литературу, но она находила утешение в определённости чисел. Сначала пришла любовь к математике, которую поддерживали учителя, видевшие её потенциал и порой дававшие ей задачи сложнее программы, чтобы она училась быстрее. Потом – любовь к естественным наукам, химии, биологии, физике, питавшая её любовь к технике. Она поглощала знания жадно. Проглатывала любую книгу, какая попадалась под руку, заканчивала собственные учебники и принималась за учебники старшей сестры. Она расцветала в соревновании: её стихией были олимпиады по математике и естественным наукам, где она толкалась среди сверстников в ожесточённой гонке за знаниями.

В шестнадцать лет её не по годам ранние успехи принесли ей стипендию на учёбу за границей, в канадской частной школе – Колледже Пирсона в Виктории, Британская Колумбия. Эта возможность вывела её на стремительную восходящую траекторию: Дартмутский колледж, где она изучала машиностроение, аэрокосмическая компания и первый крупный карьерный прорыв – Tesla, где она была старшим продакт-менеджером Model X.

В Tesla она научилась создавать сложные продукты под сильным давлением, лавируя и договариваясь между командами с очень разными мнениями и областями знаний. Именно там, как она часто говорит, её потянуло к ИИ, когда компания занялась автономным вождением. Чем глубже она погружалась в ИИ, тем больше видела в нём фундаментальное достояние, широко применимое и необходимое всем для решения трудных задач. «Действительно казалось, что это, может быть, последнее, над чем нам вообще придётся работать», – говорила она позже Кевину Скотту в его подкасте.

В ИИ Мурати пришла не сразу. Через три года в Tesla она ушла в компанию Leap Motion на должность вице-президента по продукту и разработке, чтобы заниматься системами дополненной и виртуальной реальности. Она представляла, как компания совершит переворот в образовании: учащиеся смогут поворачивать нити ДНК движением руки или управлять физикой летящего мяча. Вместо этого компания поставила на технологию слишком рано: от виртуальной и дополненной реальности ещё слишком многих тошнило. В 2018 году, через два года, она ушла в OpenAI, когда та ещё была некоммерческой.

В OpenAI её восхождение продолжилось. Когда некоммерческая организация превратилась в коммерческое предприятие, она естественным образом стала вице-президентом по Applied и партнёрствам. Она курировала важнейшие для компании отношения – с Microsoft – и сначала зарождающееся, а затем бурно растущее подразделение, коммерциализировавшее исследования компании. Она была ещё моложе Альтмана, Суцкевера и Брокмана и какое-то время – единственной женщиной с технической специализацией в высшем руководстве. Из-за этого она порой становилась мишенью сексизма, особенно со стороны исследователей, тосковавших по ранним дням лаборатории: они считали её недостаточно технической – инженером, а не учёным – и видели в её возвышении символ отхода OpenAI от серьёзных фундаментальных исследований.

И правда, в неуклюжем, ботанском, пропитанном тестостероном мире ИИ Мурати выделялась. Она была ловка в общении, умела слушать, и самолюбия у неё, можно сказать, не было. Среди тех, с кем она работала, она славилась исключительным умением решать проблемы. В водовороте непрекращающегося внутреннего конфликта OpenAI она умела вести компанию вперёд, прислушиваясь к разным идеям и точкам зрения, но не боясь принимать трудные решения. «Представьте, что нужно принять мучительно трудные решения, а ясного ответа нет. А она просто помогает найти ответ, – говорит бывший коллега. – И она неизменно оказывается права».

* * *

Среди многих ролей Мурати всё чаще оказывалась переводчицей и мостом к Альтману. После раскола с Anthropic Альтман попросил её руководить не только Applied, но и Research; в мае 2022 года она официально стала техническим директором. По мере того как ей подчинялось всё больше команд, она стала важнейшим каналом, через который сотрудники взаимодействовали с Альтманом. Если он корректировал стратегическое направление компании, исполняла это она. Если команде нужно было возразить против его решений, она становилась их заступницей.

Даже среди руководства компании у неё было такое влияние на Альтмана и такой доступ к его мнениям, каких не было у других. Она могла прямо сказать ему, что его ожидания или планы нереалистичны, и он часто прислушивался.

Она прямо говорила другим, если он чего-то не хотел, даже когда он делал вид, что хочет. Когда люди отчаивались добиться от Альтмана прямого ответа, они просили её помочь расшифровать его мнение. «Она была просто честной, – говорит другой бывший коллега. – Именно она делала дело».

Но чем больше Мурати работала с Альтманом, тем чаще ей приходилось разгребать за ним. Если две команды расходились во мнениях, он нередко в частном разговоре соглашался с позицией каждой, что сеяло путаницу, обостряло конфликт и порождало недоверие между коллегами. Эта манера усугубляла хаос, который по-прежнему создавал Брокман, врываясь в проекты. Для Мурати Брокман был как второй CEO, только плохой: категоричный и склонный своим напором доводить людей до выгорания. Во время разработки GPT-4 связка Альтмана и Брокмана создала для отдельных частей компании критический уровень стресса, из-за которого ключевые люди из команды предварительного обучения – одной из главных команд, отвечавших за сбор данных и первоначальное обучение каждой модели, – чуть не уволились.

А ещё был Наделла, практически третий CEO OpenAI, если судить по тому, насколько Альтман мог подстраиваться под интересы Microsoft. Не раз, когда Мурати тщательно составляла план разумных обязательств, которые OpenAI могла взять на себя перед Microsoft, Альтман отходил от сценария в разговорах с руководителями техногиганта, рисуя иную картину того, над чем работает команда, лишь бы согласиться с требованиями, оторванными от дорожной карты стартапа. Совету же Альтман описывал эту неразбериху иначе: у Мурати просто нет продуктивных отношений с важнейшим партнёром OpenAI.

Поведение Альтмана становилось всё хуже после того, как ChatGPT вознёс его к мегазвёздности, усилив и внимание, и пристальность к нему и взорвав его календарь изнурительным графиком поездок. Раньше он, как правило, был полон энергии; теперь часто был измотан. И он ломался под этим давлением: тревожность достигла новых высот и подпитывала его привычки к разрушительному поведению. Он делал то же, что всегда, – соглашался со всеми в лицо, – а теперь всё чаще ещё и поносил их за спиной. Это порождало в компании больше путаницы и конфликтов, чем когда-либо, а руководители команд перенимали его дурную манеру и стравливали своих подчинённых. Это и само по себе разъедало компанию. Но под натиском растущей внешней конкуренции Альтман ещё и подталкивал компанию развёртывать модели всё быстрее и пытался ради скорости обходить некоторые устоявшиеся процедуры выпуска – порой с помощью нечестности. Недавно он сказал Мурати, что, по его мнению, юридическая команда OpenAI разрешила выпустить GPT-4 Turbo без рассмотрения в DSB. Но когда Мурати уточнила у Джейсона Квона, руководившего юридической командой, тот понятия не имел, откуда у Альтмана такое впечатление.

Летом Мурати попыталась дать Альтману подробную обратную связь о нарастающих проблемах в надежде, что это побудит его задуматься и

измениться. Вместо этого он устроил ей ледяной приём, и ей понадобились недели, чтобы растопить отношения, в том числе заверив его, что она ни с кем больше этой обратной связью не делилась. Она видела, как он проделывал похожее и с другими руководителями: если они с ним не соглашались или бросали ему вызов, он мог быстро отрезать их от ключевых решений или начать подрывать их авторитет. Это делалось настолько тонко и в таких рамках, вне поля зрения сотрудников, что ей понадобилось несколько лет, чтобы осознать весь масштаб. Но неизбежно каждому из руководителей в свой черёд доставалось такое обращение. Со временем совокупный эффект его действий нанёс урон самым верхним уровням организации.

Совсем недавно под удар попал Суцкевер. Какое-то время назад Якуб Пахоцкий, польский исследователь, руководивший проектом AI Scientist и подчинявшийся Суцкеверу, разочаровался в том, что ему недостаёт признания и полномочий. Он обратился к своему союзнику Брокману, а Брокман – к Альтману. Альтман поддержал амбиции Пахоцкого и дал ему более высокую роль в Research. Загвоздка была лишь в одном: Альтман ни слова об этом не сказал Суцкеверу. Не стал он и разграничивать сферы ответственности Суцкевера и Пахоцкого, когда оба, получая от Альтмана разные сигналы, начали вести одни и те же исследования каждый в своём направлении и тщетно пытались понять, откуда взялось это рассогласование.

Запутанная ситуация на несколько месяцев погрузила подразделение Research в организационную неразбериху. Теперь, как и в кризисе с командой предварительного обучения GPT-4, стресс дошёл до невыносимого уровня. Для Суцкевера эта затянувшаяся история была глубоко болезненной. Мало того что это было унизительным пренебрежением со стороны Альтмана, она разрушила его дружбу с Пахоцким, выращенную годами – поздними вечерами, взлётами и падениями, работой бок о бок над созданием компании.

Мурати снова работала сверхурочно, чтобы найти выход. Немало времени у неё ушло просто на то, чтобы понять, что происходит. Теперь ей нужно было добиться, чтобы Суцкевер и Пахоцкий согласились на какую-то договорённость, чтобы Брокман перестал подкладывать гирьку на весы, обращаясь к Альтману, и чтобы Альтман держался единой линии и перестал противоречить ей на закрытых встречах. Но ничто из этого не устранило бы корень проблемы. Очередной кризис в высшем руководстве был лишь вопросом времени. По-настоящему OpenAI нужны были более сильные механизмы управления и подотчётности.

Пока Мурати была в немилости, Альтмана, похоже, больше всего тревожило и пугало то, что она могла поделиться своей подробной обратной связью с советом. На самом деле она с советом не говорила. Она хотела уладить всё с ним напрямую и не была уверена, что обращение к совету действительно обеспечит подотчётность. За эти годы она скептически относилась к разным его составам. Три сооснователя в совете, который должен был быть независимым

советом некоммерческой организации, – это было слишком много. А многие независимые директора не были по-настоящему независимы от Альтмана: так или иначе они были связаны с ним финансово или пользовались плодами его связей. Она видела, как он инвестирует в стартапы и жертвует деньги политикам, чтобы завязывать и закреплять важные отношения. Он не раз спрашивал её, что может для неё сделать, и она неизменно уклонялась. Ей не хотелось запутаться в его паутине и потом быть ему обязанной.

Но если и было время обратиться к совету, то, возможно, сейчас был подходящий момент, чтобы по крайней мере наладить более регулярный канал связи. В конце сентября она собиралась быть в Вашингтоне, где жила Тонер, чтобы выступить на ежегодном фестивале идей журнала The Atlantic. Совет как раз искал новых членов, и тот, кто войдёт в него следующим, мог либо укрепить, либо ослабить его независимость. Альтману нужен был настоящий надзор. Мурати пригласила Тонер на кофе.

* * *

Для Тонер обращение Мурати было необычным, но не совсем неожиданным. Тонер была членом совета, Мурати – руководителем. Казалось вполне разумным, что Мурати хочет поговорить.

Встреча за кофе 29 сентября 2023 года казалась довольно обычной. Мурати рассказала о разных делах компании: OpenAI заработает кучу денег, тут проблем нет; самое важное из того, что сейчас в работе, – модель Gobi и новая сделка, о которой шли переговоры с Microsoft; а ещё она разбирается с кадровыми проблемами, связанными с тандемом Альтмана и Брокмана. На этот раз дело касалось Суцкевера и Пахоцкого.

Лишь одно немного удивило Тонер: Альтман так давит на компанию, чтобы та выпускала продукты быстрее, сказала Мурати, что она боится, как бы это не привело к плохим последствиям.

Куда необычнее показалось Тонер письмо, которое через несколько дней она получила от Суцкевера. За два года в совете Суцкевер ни разу не обращался к ней лично. В письме он спрашивал, найдётся ли у неё время встретиться на следующий день. И вот, 4 октября, он так нервничал, что ему было трудно говорить.

Тонер взяла инициативу на себя.

– Мне правда очень важно, чтобы у нас был хороший, сильный совет, способный надзирать за компанией, – сказала она ему. – Этого все мы хотим.

Суцкевер вдруг рассмеялся и одновременно фыркнул – совсем на него не похоже.

– Полностью согласен, – сказал он таким тоном, будто другие не согласны.

– Все согласны, – мягко сказала Тонер.

Суцкевер закатил глаза.

Независимые члены совета, объяснила Тонер, хотят укрепить совет, добавив нового директора с серьёзным опытом в безопасности ИИ. Если у него есть другие идеи, она будет рада их выслушать.

За это Суцкевер ухватился. Совету нужно быть лучше информированным, сказал он. Им нужно обращать внимание на то, что происходит.

Тонер попыталась расспросить подробнее. А какую информацию, по его мнению, совету важнее всего знать?

Суцкевер помолчал, тщательно подбирая слова.

– Я не решаюсь ответить на ваш вопрос прямо, – сказал он. – Если бы я ответил, вы бы поняли почему.

– На самом верху OpenAI – непростая среда, – продолжал он. – OpenAI сложнее, чем кажется снаружи.

– Может быть, я посплю с этим и пойму, что какими-то конкретными вещами могу поделиться, – осторожно добавил он.

А пока, посоветовал он, Тонер, пожалуй, не помешало бы поговорить с Мурати. У Мурати больше контекста о том, что происходит.

* * *

Мурати получила ответ от Тонер больше чем через неделю. После очередного заседания совета Тонер просто сказала Мурати, что, похоже, творится что-то странное. Мурати ответила, что Тонер очень проницательна. Они договорились поговорить ещё раз.

Шло 15 октября 2023 года, и Тонер начала звонок с дежурного вопроса.

– Как дела? – спросила Тонер. – Есть что-нибудь, о чём совету следует знать?

Мурати взвешивала слова. Действовать нужно было осторожно.

– Происходит много непростого, – неуверенно сказала она, повторив выражение, которым до неё воспользовался Суцкевер. Мурати продолжила: обо всём она говорить не может, по крайней мере о самом непростом, потому что, раз сказав, уже не сможет взять слова назад. Особенно сейчас, когда Альтман чувствует, что его положению CEO сильно угрожают.

Последнее, похоже, удивило Тонер. Альтману угрожают? Совет, конечно, пытался ввести более сильные механизмы управления, чтобы сдерживать его власть, но ни разу не хотел угрожать его положению и даже отдалённо не обсуждал его смещение, сказала Тонер.

Мурати пояснила: Альтман – невероятно тревожный человек. А когда он тревожится, у него появляются глупые идеи, особенно если им потакает Брокман. Тревожность Альтмана к тому же подпитывает токсичное поведение, которое всегда следует одному и тому же сценарию: тому, кто сопротивляется его решениям, он говорит всё, что, по его мнению, тот хочет услышать, чтобы заручиться его поддержкой; а когда у него кончается терпение и он решает, что человек и дальше будет ему противиться, он подрывает его авторитет, пока тот

не уйдёт с дороги. Это делается незаметно, но пронизывает всё, и совсем недавно проявилось в истории между Суцкевером и Пахоцким. Это нанесло огромный вред, и Суцкевер очень расстроен, сказала она.

Совету нужно позаботиться о том, чтобы седьмым директором не стал союзник Альтмана, продолжала Мурати. И нужно следить за тем, как Microsoft развёртывает технологии OpenAI, и за DSB. Больше она сказать не может. Альтман придёт в ужас, если узнает, что они с Тонер разговаривают. Но уровень токсичности на самом верху руководства невыносим, и в ближайшие шесть – девять месяцев что-то должно сломаться.

Тонер стоит поговорить с Ильёй, закончила Мурати, и посмотреть, чем он готов поделиться.

* * *

Когда Суцкевер впервые обратился к Тонер, у него на душе было много всего. За тот год, наблюдая стремительный взлёт OpenAI, он всё больше думал о скором появлении AGI: о катастрофических переменах, которые он вызовет, о том, что они будут необратимы, об ответственности OpenAI за то, чтобы конечным итогом стало невероятное изобилие, а не невероятные страдания.

Потом его поглотила другая тревога: он терял веру в то, что OpenAI вообще сможет достичь AGI или нести такую ответственность, пока её возглавляет Альтман.

После ChatGPT работа в OpenAI и продвижение по её служебной лестнице стали высшей социальной валютой в Кремниевой долине. Это подняло внутреннюю конкуренцию и офисную политику на новый уровень: разные руководители команд толкались, борясь за внимание и приоритет. Альтман делал всё значительно хуже, видел Суцкевер. Вместо того чтобы улаживать столкновения самолюбий, он говорил каждому ровно то, что тот хотел услышать, маневрируя, чтобы получить ровно то, чего хотел сам. И при этом говорил столько мелкой лжи, а порой и крупной, что это стало происходить почти ежедневно. Брокман, на взгляд Суцкевера, добавлял сумятицы. Прошли те дни, когда два первых сооснователя были друг для друга доверенными наперсниками с тёплыми воспоминаниями о бесконечных часах, проведённых взаперти в мечтах о том, чем может стать OpenAI.

Для Суцкевера итогом стало самое токсичное сочетание: бесцельная, хаотичная среда, где все бьют друг друга в спину и у людей больше нет общей информации и общей основы доверия, чтобы договариваться о важнейших решениях о том, как двигаться дальше. Эта грызня подрывала то, что Суцкевер считал двумя опорами миссии OpenAI: она замедляла продвижение исследований и сводила на нет любые шансы принимать здравые решения о безопасности ИИ.

А теперь поведение Альтмана било и по нему лично.

После первого разговора с Тонер 4 октября Суцкевер спал урывками, поглощённый стрессом. Тонер, считал он, была самым безопасным из независимых директоров, к кому можно обратиться. На заседаниях совета она открыто выступала за введение сильных механизмов управления и безопасности и явно не была у Альтмана в кармане. Насчёт двух других независимых директоров, Д'Анджело и Макколи, он был не так уверен. И всё же он тревожился, насколько полно может открыться Тонер и что будет, если Альтман узнает.

Пока Суцкевер боролся с этими мыслями, вновь разошедшиеся в Twitter обвинения Энни добавили ещё одно измерение к его растущему списку вопросов о том, годится ли Альтман на то, чтобы вести мир к AGI. После статьи в журнале New York, которую в компании некоторые обсуждали только вполголоса, снова разлетелись по сети два старых твита Энни: ноябрьский 2021 года, где Сэм обвинялся в «сексуальном, физическом, эмоциональном, словесном, финансовом и технологическом насилии», и другой, от 14 марта 2023 года, который соберёт почти 4000 лайков и был откровеннее:

> Мне уже не четыре года, и ко мне в кровать больше не забирается без согласия тринадцатилетний «брат».
>
> (Можешь не благодарить за то, что я помогла тебе разобраться в своей сексуальности.)
>
> Я наконец приняла, что ты всегда боялся и всегда будешь бояться меня больше, чем я боялась тебя.

Суцкевер не знал, правдивы ли её обвинения, но считал, что, что бы ни произошло, детство рядом с Сэмом далось Энни нелегко. Это показывало, как далеко в прошлое могла уходить история проблемного поведения Сэма.

Слово Энни – «насилие», abuse, – было и тем словом, которое, по ощущению Суцкевера, лучше всего передавало его собственные наблюдения за Альтманом. Как и Мурати, он долго разбирался в сценарии Альтмана, хотя признаки его ненадёжности были видны с самого начала: настойчивое желание стать CEO OpenAI без ясных или последовательных причин; мелкая ложь, которую он порой говорил в первые годы и которая казалась настолько несущественной, что в ней не было смысла; предупреждения, которые Амодеи передал в конце 2020 года, уходя. Тогда Суцкевер не до конца понял выражение Амодеи «психологическое насилие». Теперь, когда поведение Альтмана ухудшалось, а его последствия стремительно нарастали, Суцкевер понимал его смысл по-новому и глубже.

12 и 13 октября он ездил на выездную встречу со своей командой Superalignment, где снова сжёг изваяние ради сплочения команды и продолжал мучительно думать, как быть дальше. Теперь, 16 октября, во время второго звонка с Тонер, он был готов рассказать немного больше.

Он рассказал о ситуации с Пахоцким и о том, как она иллюстрирует поведение Альтмана. Альтман мог просто прямо и честно сказать ему, что хочет, чтобы Пахоцкий играл большую роль. Вместо этого он стравил Суцкевера и Пахоцкого – во многом при содействии Брокмана. Они, похоже, даже охотно это делали, оставив двух учёных биться друг с другом, не понимая толком, почему им никак не удаётся договориться. «Порка будет продолжаться, пока не поднимется боевой дух», – сказал Суцкевер.

Беда в том, что всё, что делал Альтман, всегда было очень тонко. Каждый поступок по отдельности не казался чем-то особенным. Лишь когда смотришь на всё сразу, закономерности становятся отчётливыми. Вывод здесь не «посмотрите на этот неприятный случай, где Илья чувствует себя обиженным», подчеркнул он. Это лишь последний пример закономерностей психологического насилия со стороны Альтмана. Было ещё и то, как он обошёлся с Амодеи, с которым Тонер стоит поговорить, настаивал Суцкевер. И были распространявшиеся обвинения его сестры. «Это проблема безопасности другого рода, если вы понимаете, о чём я», – сказал он. Суть в том, что Альтман, порой вместе с Брокманом, годами похожим образом обращался с множеством людей, манипулируя ими и лгая им настолько привычно, что временами говорил вещи, в которые, похоже, не верил и сам.

При этом, добавил Суцкевер, они вдвоём, Тонер и он, могут поговорить и решить, что сделать почти ничего нельзя. Если так выйдет, им следует забыть, что этот разговор вообще был.

Но было одно конкретное обсуждение, ради которого он и обратился к Тонер. На конец ноября совет назначил своё второе ежегодное выездное заседание, где наконец собирались решить вопрос о новых директорах. Об этом он и хотел поговорить с Тонер. Он не считал, что расширять совет – хорошая идея. Он не был уверен, что в результате совет станет независимее. Даже если новые члены не будут сторонниками Альтмана, им понадобится слишком много времени, чтобы раскусить его тактику. Призвать его к ответу станет не легче, а труднее, сказал Суцкевер.

Впрочем, в этом мнении он не был полностью уверен, осторожно оговорился он. Ему хотелось знать, что видит Тонер.

Тонер согласилась, что Альтман скользкий. Она и в собственной профессиональной жизни видела, как скользкость на уровне руководства может вызывать цепочку проблем, сказала она. Теперь, когда совет лишился трёх членов, больше всех склонявшихся перед Альтманом, она также согласилась, что цель всего, что произойдёт дальше, – призвать Альтмана к ответу.

Суцкевер успокоился. Теперь он лучше понимал позицию Тонер. Они стояли по разные стороны тупика, но хотели одного и того же – создать реальные сдержки для власти Альтмана. Если она видела способ в том, чтобы добавить членов совета, то он – в противоположном.

Он двинулся дальше. Он понимал, что его тревоги могут показаться Тонер очень сильными и внезапными, но дело в том, что окно для исправления ситуации сужается. «Совет – это как внешнее согласование, руководство – как внутреннее», – сказал он. Если Тонер говорила, что OpenAI нужно исправлять внешнее согласование, то он считал, что исправлять нужно внутреннее.

Тонер, казалось, переваривала услышанное.

– Похоже, вы считаете, что на столе должны быть весьма серьёзные перемены, – сказала она.

– Да, – ответил Суцкевер. Но главная трудность в том, что нет никаких однозначных доказательств чего-то явно вопиющего в руководстве Альтмана. Скорее всего, совет решит, что ничего сделать нельзя, наконец договорится о новых директорах, и момент будет упущен.

На это Тонер предложила несколько альтернатив. Возможно, совет мог бы поставить OpenAI определённые цели, чтобы конкретнее оценивать работу Альтмана, и вернуться к вопросу через двенадцать месяцев.

Суцкевер отмахнулся. Альтман выполнит любые цели, которые поставит совет, но к структурным изменениям, которые устранили бы проблемы с его поведением, это не приведёт.

Может быть, Брокман мог бы выйти из совета, предложила Тонер.

Это, конечно, могло бы помочь, ответил Суцкевер, но этого всё равно недостаточно. Брокман, безусловно, усугублял манеру Альтмана, но корень проблемы – именно Альтман.

– Я думал в похожем направлении как о плане Б, – сказал Суцкевер о предложении Тонер насчёт Брокмана.

Что он считал планом А, он вслух так и не сказал. Но когда они заканчивали разговор и договаривались созвониться на следующей неделе, Суцкевер был уверен, что Тонер начинает понимать.

* * *

Мурати разбиралась с очередным кризисом. Незадолго до второго разговора с Тонер Альтман, без всякой видимой причины, впал в панику из-за того, что Microsoft якобы недовольна OpenAI. Пытаясь докопаться до причины, Мурати устроила встречу с руководителями Microsoft, в том числе с Михаилом Парахиным, главой Bing.

Встреча прошла лучше, чем ожидалось, но по ходу дела она обнаружила, что Альтман опять согласился на одно из требований Microsoft, не опираясь на то, чем OpenAI на самом деле занимается, и создал у Microsoft ложные ожидания насчёт того, что OpenAI поставит. И опять разгребать ситуацию пришлось ей.

Встреча с Microsoft прошла 19 октября. Теперь, 20 октября, во время третьего разговора, она говорила Тонер по телефону, что собирается дать Альтману много обратной связи. Проблем было столько, когда Альтман говорил «да», а она – «нет». Это сильно раздробило отношения с Microsoft.

Ровно та же ситуация с Суцкевером и Пахоцким, продолжала Мурати. Как раз сегодня они наконец договорились о том, как Суцкевер и Пахоцкий будут дальше сосуществовать в своих ролях. Вскоре после этого они с Суцкевером умоляли Альтмана, говоря с Пахоцким, не отступать от их решения и не говорить просто то, что Пахоцкий хочет услышать. «Когда я говорю с Якубом, он меня слышит, а потом идёт к Сэму и слышит другое», – сказала она Тонер. Альтман дал им на встрече слово. Но всё оставалось шатким. Брокман мог поговорить с Альтманом от имени Пахоцкого и попытаться на него повлиять, снова грозя нарушить равновесие.

Брокман – это вообще отдельная история, сказала Мурати. Недавно он признался ей, что пытался уволить её во время разработки GPT-4. Формально она была его руководительницей и раньше писала его оценки эффективности, но это всегда вызывало столько драмы, что она перестала. Позже Мурати призналась кому-то ещё, что хотела бы уволить его, но не могла, потому что он входил в совет. Одно время она думала попросить Брокмана выйти из совета, сказала она Тонер. В итоге не стала. И теперь хаос был повсюду.

А знает ли Мурати что-нибудь о ситуации с Энни? – спросила Тонер.

Мурати не знала. Она не спрашивала Сэма об этом и толком не представляла, что произошло в его семье. Но если правдой окажется хотя бы 10 процентов, это очень плохо, сказала она.

– Меня поражает, что я могу так хорошо делать свою работу при всём, что происходит, – продолжала Мурати. Она записывала происходящее по ходу дела. Если Тонер понадобится, она может прислать больше информации.

Тонер ответила: совет сосредоточится на том, что действительно может изменить, – не на личности или поведении Сэма, а на введении более совершенных процедур и структур управления, которые будут держать его в узде.

Это совпадало с тем, чего Мурати пыталась добиться внутри компании. У неё было для Тонер ещё лишь одно предостережение.

– Позаботьтесь, чтобы ваша информация поступала не только от Сэма, – сказала она.

* * *

Тонер не знала, о чём хочет поговорить Альтман. Ранее в тот день, 25 октября, он написал ей, спросив, найдётся ли у неё время поболтать сегодня или завтра. За два дня до этого, 23 октября, она снова говорила с Суцкевером, который, узнав от Мурати, что та тоже разговаривала с Тонер, стал гораздо откровеннее. Он высказал свои опасения яснее, чем когда-либо. «Не думаю, что Сэм – тот человек, чей палец должен лежать на кнопке AGI», – сказал он и отметил «огромную возможность», выпавшую совету, что-то с этим сделать. Затем он предложил путь вперёд: заменить Альтмана на Мурати в качестве временного CEO.

Позже, когда Тонер, Макколи и Д'Анджело совещались между собой, они поняли, что и Мурати сказала: «Мне неспокойно от того, что Сэм поведёт нас к AGI». Это открытие сильно повлияет на их размышления. Если двое самых высокопоставленных заместителей Альтмана – одна от Applied, другой от клана безопасности – оба так считают, у совета серьёзная проблема.

Затем, 24 октября, Тонер провела встречу с Д'Анджело и Макколи, чтобы обсудить, какие ещё шаги можно предпринять для укрепления механизмов надзора совета. Одна вопиющая проблема: у некоммерческой организации OpenAI не было достаточной независимой юридической поддержки, и всё шло через юристов коммерческой структуры. Все трое согласились, что пора найти для некоммерческой организации новых юристов, которые присутствовали бы на каждом заседании совета и помогали проверять все сделки и другие юридические договорённости, которые заключал Альтман.

Тонер гадала, не прослышал ли Альтман каким-то образом об этих встречах и не хочет ли положить конец обсуждениям. Но теперь, по телефону, он говорил совсем о другом – о своих опасениях насчёт научной статьи, которую она опубликовала по основной работе в CSET.

На той неделе Тонер опубликовала три статьи. Он выделил самую сложную и академичную. Она была посвящена политологическому понятию «дорогостоящих сигналов» – тому, с какими трудностями сталкиваются государственные и частные субъекты, сообщая общественности о своих намерениях в области регулирования и развития ИИ. Тонер была третьим соавтором, а упоминания OpenAI были запрятаны на страницах 28–30 документа объёмом шестьдесят пять страниц. Больше компания нигде не упоминалась – ни в кратком изложении, ни на веб-странице, ни в каких-либо материалах к выходу статьи. Судя по трафику, её почти никто даже не прочитал. Тонер не понимала, как она вообще попала в поле зрения Альтмана.

Тонер не знала, что накануне статья всплыла в Slack OpenAI, в канале #policy-research-chatter. Страницы, где упоминалась компания, поочерёдно разбирали сильные и слабые стороны стратегий выпуска моделей OpenAI и Anthropic, в том числе хвалили OpenAI за то, что та опубликовала откровенную оценку безопасности GPT-4 и сообщила, что ради этого отложила выпуск модели на полгода. Дэвид Робинсон, руководитель политического планирования OpenAI, вставил в Slack лишь три выбранных абзаца – те, где критиковалась OpenAI и хвалилась Anthropic. Несколько строк он для выразительности выделил жирным: в них противопоставлялись «динамика гонки на дно», которую подстегнул ChatGPT, и сдержанность, проявленная Anthropic при выпуске Claude после ChatGPT.

«Кстати о докладах CSET, только что увидел вот этот новый, – написал Робинсон. – Хелен Тонер – соавтор, и сравнения OpenAI с Anthropic довольно острые».

Это вызвало в канале короткое обсуждение.

«Да, на удивление пристрастно (не столько критика в наш адрес, которая, по-моему, жёсткая, но справедливая, сколько некритичное отношение к Anthropic)», – написал один.

«Да, согласен, ощущение довольно пристрастное, и я бы сказал, ещё и довольно хлипкое?» – добавил другой. «Как бы то ни было, – продолжил он чуть ниже, – большое спасибо, что поделились, просто удивлён таким уровнем анализа в докладе, судя по этим отрывкам».

Альтман сказал Тонер, что о статье через несколько часов после её выхода сообщил в письме в OpenAI кто-то посторонний. Его беспокоило, что критика OpenAI со стороны члена совета может плохо выглядеть, пока компания находится под пристальным вниманием регуляторов, – в том числе в связи с начатой в июле 2023 года проверкой FTC её практик в отношении данных, обучения и безопасности, а также галлюцинаций её моделей, которые могли повредить репутации потребителей. Тонер написала черновик статьи в мае или июне того года, до того как внимание регуляторов усилилось. Она признала, что не перечитала её свежим взглядом в контексте новой политической обстановки.

Разговор длился пятнадцать минут. Альтман всё это время говорил мягко. В конце они оба согласились, что она напишет остальным членам совета, обратит их внимание на статью и объяснит, что произошло. В письме она взяла примирительный тон. Она извинилась за две ошибки: за то, что считала, будто статья ни у кого не вызовет интереса, и за то, что из-за этого не перечитала её внимательнее. «Мы с Сэмом оба согласны, что важно, чтобы члены совета могли критиковать компанию, если захотят, но что делать это так не следовало», – писала она.

Никто из других членов совета ей не ответил. Казалось, на этом всё и закончилось.

* * *

Во время третьего звонка Суцкевера Тонер 23 октября она предложила ему связаться с Макколи и Д'Анджело. Суцкевер всё ещё не был в них вполне уверен и не знал, может ли им доверять. Когда он встретился с Д'Анджело лично, тот, похоже, знал о проблемном поведении Сэма меньше, чем Тонер. Теперь, 26 октября, говоря по телефону с Макколи, Суцкевер осторожничал, чтобы не раскрыть слишком много.

Но кое-что он хотел у Макколи узнать. Вскоре после звонка Тонер Альтман разослал некоторым людям в компании письмо, где говорилось, что он поговорил с ней о статье и решительно не согласен с ней насчёт её последствий. «У меня не возникло ощущения, что мы одинаково понимаем ущерб от всего этого, – писал он. – Любая критика со стороны члена совета имеет большой вес». Суцкеверу Альтман сказал прямее: Тонер нужно убрать из совета, и Макколи с ним согласна. Суцкеверу это показалось странным.

– Сэм сказал, что, когда он говорил с вами о статье Хелен, вы сказали: «Хелен, очевидно, должна уйти», – осторожно начал он, обращаясь к Макколи. – И Сэм сказал, что после этого стал лучше к вам относиться. Это правда?

На другом конце провода Макколи, казалось, остолбенела. Ничего подобного она точно не говорила, ответила она. Альтман действительно звонил ей поздно вечером 24 октября, чтобы поговорить о статье. Он упомянул, что в ней есть раздел, который он считает критическим по отношению к OpenAI, и что Д'Анджело не считает это проступком, за который увольняют, но и что Тонер не следовало это писать. Макколи ответила Альтману, что статью не видела, и предложила ему поговорить напрямую с Тонер. Ничего такого, что намекало бы на необходимость вытеснить Тонер из совета, она сказать никак не могла, сказала Макколи.

Это могло показаться совпадением: посреди разговора о «конкретной неправде» Альтмана – ещё один пример ровно того же, разворачивающийся в реальном времени. Альтману это даже сошло бы с рук, не будь у Суцкевера уже повода позвонить Макколи. Альтман знал, что они обычно никогда не разговаривают. Но для Суцкевера при той частоте, с которой Альтман лгал и маневрировал, нечто подобное должно было случиться рано или поздно.

Положив трубку после разговора с Макколи, Суцкевер перезвонил Тонер. Пора, решили они, поговорить троим независимым членам совета.

Глава 16

Плащ и кинжал

К вторнику, 31 октября, все успели поговорить со всеми. Тонер поговорила с Макколи. Макколи позвонила Д'Анджело. Мурати поговорила с Д'Анджело и Макколи.

В тот день трое независимых директоров – Тонер, Макколи и Д'Анджело – начали почти ежедневно встречаться по видеосвязи, согласившись, что отзывы Суцкевера и Мурати об Альтмане и предложение Суцкевера заменить его заслуживают серьёзного обсуждения. Суцкевер, уже твёрдо пришедший к выводу, что Альтмана нужно уволить, в обсуждениях не участвовал. И он, и остальные считали, что независимые директора должны прийти к собственным выводам без его влияния. К тому же у него была финансовая доля в компании, и они не хотели, чтобы она повлияла на их решение.

Позже директора назовут Суцкеверу и третью причину: их доверие к Альтману упало так низко, что один из них подозревал, не подослал ли на самом деле Альтман Суцкевера к совету, чтобы проверить их лояльность и выдавить всякого, кто выступит против него.

Независимые директора разложили по полочкам то, что знали. Высшие руководители описывали Альтмана таким образом не впервые. В общей сложности все трое слышали похожие отзывы как минимум от семи человек, находившихся на один-два уровня ниже Альтмана, включая Суцкевера, Мурати и Амодеи, – людей, отвечавших как за безопасность, так и за другие части компании. Несколько из них называли поведение Альтмана абьюзом и манипуляцией; большинство подчёркивали его нечестность и то, что ему нельзя верить на слово. А ещё было множество других проблем, которые независимые директора обнаружили сами: ослабление некоммерческой организации; то, что Альтман не раскрыл, что юридически владеет OpenAI Startup Fund; то, что Альтман умолчал о нарушении Microsoft процедуры DSB; то, что Альтман пытался вытеснить из совета Д'Анджело, а теперь, по-видимому, и Тонер.

Обвинений Энни они решили вовсе не касаться. В конечном счёте речь шла о Сэме в его профессиональном качестве CEO OpenAI.

В этом качестве, говорила Мурати, хотя часть поведения Альтмана можно списать на обычные привычки CEO технологических компаний, оно всё равно создаёт серьёзные проблемы. Она сравнила его с Маском, с которым работала в Tesla: Маск принимал решение и мог объяснить, почему его принял. С Альтманом же ей часто оставалось гадать, действительно ли он с ней откровенен и стоит ли за метаниями, которые он вызывает, здравый резон или какой-то скрытый расчёт. Как он раздробил отношения с Microsoft, так он дробил и собственную команду руководителей, раздавая информацию разным людям, но

никогда не давая никому из них полной картины, что позволяло ему сохранять полный контроль. Вместе с Брокманом эта динамика была губительной. Она во многом не соглашалась с братом и сестрой Амодеи, говорила она, но в этом их наблюдения были верны.

А OpenAI, отметили независимые директора, на самом деле вовсе не обычная технологическая компания. Это, пожалуй, самая влиятельная ИИ-компания в мире, ведущая разработку одной, по их мнению, из самых судьбоносных технологий. Их задел страх, который высказал Суцкевер: что значит, если OpenAI пытается создать AGI, а её высшее руководство не может доверять ни базовой, ни важнейшей информации, исходящей от CEO?

Но был и другой мысленный эксперимент: что, если бы OpenAI была обычной технологической компанией? Например, просто службой доставки продуктов вроде Instacart? Заслуживало бы поведение Альтмана его смещения? В чём-то да, в чём-то нет. Но и вообще было неочевидно, что Альтман – лучший человек, чтобы и дальше руководить компанией, рассуждали независимые директора. Он прославился стартапами, а OpenAI стремительно взрослела. Действительно ли у него есть навыки и характер, чтобы и дальше прокладывать курс – и чтобы возмещать нестабильность, которую он создаёт?

OpenAI была в блестящем положении: это была модная компания. Если совет проведёт продуманный поиск, он сможет выбрать из множества феноменальных CEO с большим опытом руководства зрелыми компаниями. Альтман не обязательно незаменим для работы OpenAI, рассуждали они; пока он колесил по миру, повседневную тяжёлую работу внутри компании тянула Мурати, и у неё были крепкие отношения с Microsoft.

Пока независимые директора совещались, Суцкевер присылал им серию документов и скриншотов с примерами поведения Альтмана, которые они собирали вместе с Мурати. Они приходили длинными досье в двух самоуничтожающихся письмах, а аватаром у него был таинственный человек в шляпе. Как и говорил Суцкевер, особо изобличающих улик там не было, а было накопление множества случаев, когда Альтман говорил разным людям разное и вызывал сильнейшее раздражение у руководства. Скриншоты показывали, что ещё как минимум двое высших руководителей – оба не из сферы безопасности и не из той семёрки, о которой директора уже знали, – отмечали склонность Альтмана обходить и игнорировать процедуры, будь они введены ради безопасности ИИ или ради слаженной работы компании. Среди прочего, узнали директора, это включало явную попытку Альтмана обойти рассмотрение GPT-4 Turbo в DSB, неверно передав Мурати слова юридической команды. Проблема, на которую указывала и Мурати, заключалась в том, как ловко Альтман избегал письменных следов. Большую часть указаний он давал устно и выкручивался из договорённостей, говоря другим сторонам, что они просто неправильно запомнили его слова.

Независимым директорам нужно было взвесить и другое. Как отреагирует Microsoft? Как отреагирует Брокман? Как отреагируют сотрудники? Они обсуждали, не связаться ли с третьим высшим руководителем, у которого, по словам Суцкевера и Мурати, были похожие опасения, не провести ли более обстоятельное выяснение фактов, не привлечь ли руководителей Microsoft. После обсуждения они каждый раз решали, что лучше этого не делать. Каждый новый человек, посвящённый в разговор, и каждый лишний день промедления с решением повышали шансы, что Альтман узнает и своими манёврами не даст им довести обсуждение до конца.

9 ноября, когда независимые директора уже подходили к окончательному решению, у Суцкевера был ещё один звонок с Макколи.

– Сэм сказал: «Таша по-прежнему решительно поддерживает то, чтобы Хелен вышла из совета», – сказал он ей.

Это была ещё более наглая ложь, чем в первый раз: Макколи больше вообще не общалась с Альтманом.

Роль статьи Тонер, как позже сочли директора, будет сильно преувеличена в СМИ – отчасти потому, что, как они были убеждены, Альтман, возможно, сам подкинул её журналистам. На второй день пятидневного кризиса в совете директора во время разговора при посреднике предъявили ему многочисленные случаи, когда он им лгал, из-за чего их доверие и рухнуло. Среди примеров они назвали то, как он солгал Суцкеверу, будто Макколи сказала, что Тонер следует выйти из совета.

Альтман на мгновение потерял самообладание: его явно поймали с поличным.

– Ну, я думал, что вы могли это сказать. Не знаю, – пробормотал он.

Члены совета поразились его наглости.

Через несколько дней первоначальные возражения Альтмана по поводу статьи Тонер появились в СМИ.

К субботе, 11 ноября, независимые директора приняли решение. Как и предлагал Суцкевер, они сместят Альтмана и назначат Мурати временным CEO. В тот же день они сразу сообщили об этом Суцкеверу, а затем продолжали ежедневно встречаться втроём, часто сверяясь с Суцкевером, чтобы подготовить документы для смены руководства. Вечером в четверг, 16 ноября, все четверо позвонили Мурати по видеосвязи. Она ответила на звонок по телефону с какой-то конференции. Услышав новость, она выглядела удивлённой, но приняла её.

– У него сейчас такая, такая паранойя, – сказала она.

Но готова ли она к этому решению? – спросили директора.

– Полностью, – сказала Мурати.

Она приняла новую роль и выразила уверенность, что сумеет донести решение до остального руководства и до Microsoft. Она также подключит Вонг, сказала она, которой можно доверять: та поможет им написать объявление.

Несколько часов спустя, после финального рывка группы, доводившей с помощью Вонг формулировки, осталось главное – сказать Альтману.

В напряжённые последние минуты ожидания никто из членов совета даже не представлял, насколько серьёзно они просчитались.

* * *

Через несколько часов после публичного объявления в пятницу, 17 ноября, дела у независимых директоров пошли совсем плохо. После того, что казалось поначалу периодом спокойствия и стабильности, – Мурати, в частности, провела продуктивный разговор с Microsoft, – она вдруг позвонила им, встревоженная по-новому. Альтман и Брокман всем рассказывают, что смещение Альтмана – это переворот Суцкевера, сказала она. Вместе с невнятными объяснениями Суцкевера на общем собрании сотрудников это привело к тому, что ключевые заинтересованные лица начали ополчаться на решение.

Вскоре после этого, когда независимые директора столкнулись на видеозвонке с враждебно настроенным руководством, они поняли, насколько всё плохо. Джейсон Квон, директор по стратегии, и Анна Маканджу, вице-президент по международным делам, возглавили атаку: они яростно отвергали характеристику поведения Альтмана как «не всегда откровенного» и требовали доказательств в обоснование решения совета, а директора считали, что не могут их предоставить, не выдав Мурати. Но даже те в комнате, у кого, как директора знали по досье, были такие же или иные сомнения в руководстве Альтмана, молчали. По мере того как ночь тянулась, а враждебность нарастала, двое самых важных союзников независимых директоров – или, по крайней мере, двое людей, которых они считали союзниками, – начали сворачивать в другую сторону.

Когда в ту первую ночь замаячила вполне ощутимая угроза развала OpenAI, решимость Суцкевера сразу дала трещину. OpenAI была его детищем, его жизнью; её распад уничтожил бы его. Он твёрдо стоял за каждым словом, сказанным об Альтмане, но его целью было укрепить OpenAI, а не разрушить её. Он был потрясён, задет и сбит с толку реакцией сотрудников и коллег-руководителей: такого исхода он не предвидел. Он начал умолять коллег по совету – и продолжал умолять с нарастающей мукой все выходные, – подумать, не пересмотреть ли им свою позицию.

Ещё сложнее было то, что и Мурати вела себя не так, как ожидали директора. Пока той ночью шли переговоры с руководством, она поддерживала с ними связь по телефону и втайне передавала им информацию. И всё же, несмотря на уверенность, с которой она обещала заручиться поддержкой людей, теперь она по-прежнему не хотела открыто встать на сторону решения совета всем своим весом. В комнате с другими руководителями, во время их нараставших стычек с советом, она порой говорила даже такое, что выглядело

так, будто она растеряна не меньше остальных и не понимает, что именно происходит.

Для Мурати нараставший бунт коллег-руководителей и сотрудников серьёзно подрывал её положение временного CEO. В ту пятницу один за другим ушли в знак протеста Брокман, затем его союзники Пахоцкий и Шимон Сидор, а также Александр Мадры, профессор MIT, находившийся в академическом отпуске, тоже поляк и близкий друг Пахоцкого. Это вызвало пожар гнева не только внутри компании: он перекидывался на всё расширявшийся круг инвесторов, и Мурати всё больше сомневалась, что сможет удержать организацию вместе и продолжать ею руководить. Её готовность занять эту роль стала колебаться. Она поддерживала решение директоров, но не участвовала в их обсуждениях. Если они хотят, чтобы у неё был хоть какой-то шанс удержать бразды правления, считала она, то теперь их бремя – сначала обосновать своё решение перед компанией.

Уже не будучи уверены, что могут положиться на Мурати, трое независимых директоров продолжили поиски нового временного CEO или новых членов совета. Особенно старался Д'Анджело, единственный в совете человек из среды Кремниевой долины: в субботу и воскресенье он сделал десятки звонков, закидывая удочки по всей своей обширной сети связей. В какой-то момент директора позвонили Дарио Амодеи насчёт поста временного руководителя. Амодеи это не заинтересовало. Но другим в те выходные казалось, что он чуть ли не в восторге от всей ситуации.

В воскресенье Д'Анджело наконец нашёл человека, который принял одно из предложений совета: Эмметт Шир, сооснователь Twitch, казалось, был готов и настроен сотрудничать, временно возглавив компанию. Но довольно скоро и он стал отклоняться от курса по, казалось бы, необъяснимым причинам. The Wall Street Journal позже сообщит, что Шир, учившийся в YC в одной волне с Альтманом, был ещё и другом и наставником выпускника YC Брайана Чески, сооснователя Airbnb. Все выходные Чески, один из самых доверенных друзей Альтмана, не отходил от телефона вместе с Ридом Хоффманом, членом совета директоров Microsoft, успокаивая инвесторов, выравнивая позицию Microsoft и наращивая давление. Чески быстро связался с Широм, и тот встал на сторону Альтмана.

К позднему вечеру воскресенья, после заверений Чески и Хоффмана, Microsoft тоже бросила свой вес на сторону Альтмана: Наделла объявил, что Альтман и Брокман переходят в Microsoft, чтобы возглавить новое подразделение передовых исследований ИИ. Другие ИИ-лаборатории кружили над OpenAI, как стервятники, стремясь урвать свою долю талантов в этой бойне. Мурати стало ясно, что, раз совет не может узаконить своё решение, события грозят оставить от компании пустую оболочку. Она встала на сторону коллег-руководителей. Чтобы спасти положение, Мурати хотела вернуть Альтмана.

За ночь, пока сотрудники составляли открытое письмо с протестом против решения совета и угрозой уволиться и перейти в Microsoft, Мурати поставила своё имя первым. Многие старшие сотрудники были преданы Мурати больше, чем Альтману. Именно Мурати изо дня в день была с ними в окопах. Именно Мурати они доверяли действовать не в собственных интересах, а в интересах компании. Другие к тому же считали, что тесные отношения Альтмана с инвесторами и его уникальное умение привлекать огромные капиталы делают его лучшим, если не единственным, человеком, способным довести до конца текущий тендерный выкуп акций OpenAI, обещавший им возможность обналичить свои доли на суммы вплоть до миллионов долларов, и обеспечить финансы для долгосрочного успеха компании. Третьи прекрасно понимали, насколько обширна сеть связей Альтмана и как он умеет делать людям карьеру в Кремниевой долине. Когда ключевые руководители и старшие сотрудники были на борту, подписи под письмом стали расти как снежный ком. Под утро Суцкевер, который уже не видел другого пути, кроме краха компании, добавил в список своё имя. «Без Миры, думаю, Сэму не удалось бы провернуть то, что он провернул», – говорит один исследователь. Независимым директорам её колебания казались самосбывающимся пророчеством.

Позже The New York Times первой расскажет о роли Мурати в смещении Альтмана. Перед сотрудниками Мурати будет защищаться. «У нас с Сэмом крепкое и продуктивное партнёрство, и я никогда не стеснялась давать ему обратную связь напрямую, – писала она. – Я никогда не обращалась к совету, чтобы дать обратную связь о Сэме. Однако, когда отдельные члены совета сами обращались ко мне за обратной связью о Сэме, я её давала – и всё это Сэм уже знал».

К утру понедельника кризиса независимые директора знали, что проиграли. Мурати и Суцкевер перешли на другую сторону, а протестное письмо сотрудников ясно показало, что дестабилизация компании стала невыносимой. Альтман вернётся; другого способа сохранить OpenAI не было. Единственным утешением было то, что, достаточно долго продержавшись, директора, похоже, добились от Альтмана готовности на уступки. И независимые директора переключились на то, чтобы спасти хоть какие-то механизмы, позволяющие и дальше призывать его к ответу: оставить в совете хотя бы одного из них ради преемственности, найти ещё двух независимых директоров, по-настоящему независимых от Альтмана, и добиться, чтобы Альтман согласился на расследование.

Но когда они уже подходили к окончательному соглашению, нашёлся ещё один человек, желавший подлить масла в огонь. Это был не кто иной, как отвергнутый бывший сопредседатель OpenAI – Илон Маск.

* * *

Все выходные X был питательной средой для всевозможных теорий и теорий заговора о драме в совете OpenAI. У Маска эти выходные тоже выдались насыщенными. Он руководил запуском SpaceX и завершил подготовку иска против некоммерческой организации Media Matters за её доклад об антисемитском контенте в X. Затем он обратился к своей платформе, купленной в результате непродуманной сделки в апреле 2022 года, чтобы защитить свою репутацию от обвинений в антисемитизме, заявить о своём абсолютистском понимании свободы слова и раскритиковать «воук-толпу» и мейнстримные СМИ. В перерывах он отвечал на чужие комментарии и мемы, а порой и сам писал об OpenAI и Альтмане.

«Я очень обеспокоен, – написал он в воскресенье, 19 ноября, днём. – У Ильи хороший моральный компас, и он не стремится к власти. Он не пошёл бы на столь радикальные шаги, если бы не считал их абсолютно необходимыми». Позже, около двух часов ночи по тихоокеанскому времени, он вызывающе выложил ролик с YouTube со знаменитой сценой крестин из «Крёстного отца», где Майкл Корлеоне из сына, раздираемого нравственными сомнениями, превращается в безжалостного нового дона, убивая глав всех остальных семей.

Но во вторник днём Маск попытался склонить чашу весов более явно. «Мне только что прислали это письмо об OpenAI, – написал он со ссылкой. – Похоже, эти опасения стоит расследовать». Это было не то письмо, что ходило среди нынешних сотрудников, но оно тоже было адресовано совету директоров OpenAI. Начиналось оно так:

> Мы пишем вам сегодня, чтобы выразить глубокую обеспокоенность недавними событиями в OpenAI, особенно обвинениями в неподобающем поведении в адрес Сэма Альтмана.
>
> Мы – бывшие сотрудники OpenAI, ушедшие из компании в период серьёзных потрясений и неразберихи. Теперь, когда вы сами увидели, что бывает, если осмелишься выступить против Сэма Альтмана, возможно, вы поймёте, почему столь многие из нас молчали, опасаясь последствий. Мы больше не можем молча стоять в стороне.

В письме выдвигался ряд требований – главным образом чтобы временный CEO Эмметт Шир расширил расследование поведения Альтмана, включив в него раннюю историю OpenAI и её корпоративную реструктуризацию с уходом от некоммерческой модели. «Мы считаем, что значительное число сотрудников OpenAI вытеснили из компании, чтобы облегчить её переход к коммерческой модели», – говорилось там. В письме также приводился ряд обвинений, описывающих «тревожную закономерность обмана и манипуляций со стороны Сэма Альтмана и Грега Брокмана».

Внизу письма был раздел «Дополнительное чтение для широкой публики» с тремя ссылками. Одна вела на тред в X Джеффри Ирвинга, исследователя безопасности ИИ, ушедшего в 2019 году в DeepMind, где говорилось, что Альтман «лгал мне по разным поводам» и «был лжив, склонен к манипуляциям и хуже

того по отношению к другим». Две другие вели на журналистские статьи. Обе были моими.

Когда я увидела твит Маска, у него было уже больше десяти тысяч ретвитов и в несколько раз больше лайков. Подборка для чтения меня удивила: об OpenAI было написано множество материалов. Почему они дали ссылки только на два, написанных мной? Первым был мой портрет OpenAI 2020 года для MIT Technology Review; вторым – статья, которую я только что написала вместе с коллегой Чарли Уорзелом для The Atlantic, давая контекст кризиса в совете и приоткрывая идеологическую поляризацию, разгоревшуюся в компании после ChatGPT.

Над этим разделом был указан адрес электронной почты в Tor для обмена зашифрованными сообщениями. «Призываем бывших сотрудников OpenAI связаться с нами», – говорилось там. Я задумалась: может быть, выбрав мои статьи, авторы письма пытались выйти на меня? Я завела собственный почтовый ящик в Tor и набрала сообщение:

> Здравствуйте! Я журналистка, написавшая обе статьи, на которые ссылается ваше письмо, и, думаю, вы пытаетесь со мной связаться.

Я добавила свои контактные данные и нажала «Отправить».

Через несколько минут Signal засветился уведомлением: кто-то ответил.

* * *

Это был один из самых странных случаев за всю мою журналистскую карьеру: ответивший человек был растерян не меньше меня. Он тоже был бывшим сотрудником OpenAI, но к написанию письма отношения не имел. Он просто получил копию письма на личную почту без всяких объяснений; когда он попытался задать уточняющие вопросы, ему пришёл ещё один загадочный ответ: ссылка на почтовый ящик в Tor с фразой capped_profit («ограниченная прибыль»), похожей на имя пользователя, а за ней – что-то похожее на пароль.

Проверив эти данные и успешно войдя в ящик, он увидел в папке «Отправленные» множество других писем бывшим сотрудникам с текстом письма. Пока он осматривался дальше, туда посыпались запросы СМИ, в том числе от The New York Times, The Washington Post и The Information. Приходили и другие письма – от людей, называвших себя нынешними и бывшими сотрудниками. В одном говорилось:

> нынешний сотрудник.
>
> работал напрямую с руководством
>
> ваше сообщение мне близко.
>
> какой у вас план?

Это письмо показалось ему особенно забавным. Хотя не было никаких оснований считать, что за ним действительно стоит Альтман, тот был известен тем, что всегда пишет строчными буквами.

Бывший сотрудник начал делать скриншоты всего подряд. В ящике одновременно с ним был залогинен как минимум ещё один человек. Пока он всё это просеивал, в разные моменты появлялись входящие ответы. Ему лично в этом деле ничего не было нужно, сказал он мне. Но ему не нравилось, что эти бывшие сотрудники, судя по всему, ссылались в своих обвинениях на опыт других людей без их согласия; многие обвинения, по его ощущению, были к тому же искажены под определённый нарратив. Он ответил на моё письмо, потому что узнал моё имя. Он прислал мне все скриншоты.

Просматривая их, я заметила несколько примечательных. В ответах разным журналистам несколько писем сообщали, что письмо было опубликовано преждевременно и разлетелось по сети раньше, чем было готово. В одном журналисту называлось и конкретное число соавторов: «На момент публикации в его написании участвовали 13 человек».

А в папке «Черновики» было письмо, которое ещё не отправили и не собирались отправлять. Это было сообщение только для тех, кто входил в ящик.

> ТЕМА: Прекратить контакты со СМИ
>
> Из-за вмешательства Илона это превратилось в конспирологическую атаку на личность.
>
> Давайте снова сосредоточимся на том, чтобы в частном порядке и по отдельности донести до совета, что есть множество доказательств того, как далеко руководство заходило с самого начала, чтобы обеспечить OpenAI коммерческое будущее.
>
> Подпишитесь, когда прочитаете.
>
> -1
>
> -2
>
> -3
>
> -4
>
> -5
>
> -

Группа, скорее всего, не могла знать, что всего через несколько часов переговоры закончатся и OpenAI объявит о возвращении Альтмана на пост CEO. Но в разгар их последней скоординированной попытки не допустить Альтмана обратно я каким-то образом случайно наткнулась на их временный штаб. По формулировкам письма многие, как и я сама, предполагали, что его, скорее всего, написали бывшие сотрудники из клана безопасности OpenAI, известного одними из самых резких мнений об Альтмане и о распаде некоммерческой организации. После возвращения Альтмана именно люди из клана безопасности

будут чувствовать себя сильнее всех преданными советом: они будут считать, что провал со смещением нанёс самый тяжёлый удар по их борьбе за то, чтобы повернуть OpenAI обратно к изначальному духу некоммерческой организации, коренившемуся в думерстве. Но никто не мог – или не хотел – сказать мне, кто написал письмо. Я так и не узнала, кто это был.

* * *

6 декабря 2023 года, через две недели после кризиса в совете, сотрудники OpenAI собрались на общее собрание в театре Дворца изящных искусств Сан-Франциско – архитектурной достопримечательности с открытой ротондой и каменными колоннами, напоминающими о греко-римской империи. Несколько руководителей, включая Мурати, снова ставшую техническим директором, Боба Макгрю, теперь директора по исследованиям, операционного директора Брэда Лайткэпа и Джейсона Квона, рассказали о планах своих подразделений на 2024 год.

Альтман выглядел подавленным – сотрудники никогда раньше не видели его таким. Заметно было и отсутствие Суцкевера, о котором Альтман сказал прямо. «Слушайте, я знаю, людям грустно, что Ильи здесь нет. Мне тоже грустно», – сказал он. На месте Суцкевера был Пахоцкий, который в сбивчивом, запинающемся выступлении доказывал, что последние научные достижения OpenAI приблизили компанию как никогда к осуществлению многолетней мечты Тьюринга о мыслящих машинах.

Во время кризиса в совете особенно бурную новую волну домыслов вызвало одно сообщение в СМИ: статья Reuters о том, что за несколько дней до увольнения Альтмана директора получили от сотрудников письмо о якобы новом научном прорыве – алгоритме под названием Q*. Q* никак не повлиял на решение совета. Но, как намекнул в своём выступлении Пахоцкий, в Research новому алгоритму действительно придавали огромное значение.

Алгоритм был детищем Суцкевера и вырос из исследований, которые тот вёл с 2021 года, чтобы развивать модели OpenAI без потребности в новых данных. Идея состояла в том, чтобы модель глубокого обучения лучше использовала уже имеющиеся данные, тратя больше вычислений на этапе вывода, то есть при генерации ответов, и давала лучшие результаты. Это ломало логику исходных законов масштабирования, связывавших производительность модели с тремя составляющими, используемыми при обучении: данными, параметрами и вычислениями. С этим новым методом Суцкевер надеялся улучшить производительность модели за счёт дальнейшего масштабирования той составляющей, которую всегда считал самой важной, – вычислений, только вычислений на этапе вывода, а не обучения; иными словами, вычислений, используемых для генерации ответов модели. Позже он объяснит свои соображения на этот счёт в программном выступлении на NeurIPS в декабре 2024 года, после того как одна из статей, написанных им в соавторстве, третий

год подряд получила премию «Проверка временем». «Вычисления растут благодаря лучшему оборудованию, лучшим алгоритмам и более крупным кластерам, – скажет он, – а данные не растут, потому что интернет у нас всего один».

Подразделение Research OpenAI считало, что Q* наконец позволит компании создавать модели с более сильными способностями к рассуждению – тем самым ускользающим ключевым ингредиентом, без которого не открыть путь к AGI. Q* был настолько важен, что после утечки руководство ввело свою самую жёсткую стратегию борьбы с новыми утечками в СМИ. Компанию полностью разгородили: Research выделили в отдельную группу Slack, закрыли доступ ко всем Google-документам, связанным с Q*, а сам проект переименовали в Strawberry («Клубника»). Переименование должно было затруднить посторонним распознавание и отслеживание внутренних проектов, если они когда-нибудь подслушают разговор исследователей OpenAI. Похожим образом после утечки проекта Arrakis в The Information от пустынных названий моделей в основном отказались.

Лихорадочные разговоры о Q* и реакция OpenAI стали странной демонстрацией того, насколько размыты основы научного поиска в области ИИ. Наука – это процесс выработки консенсуса. Значимость любого достижения – в ИИ или где-то ещё – в момент, когда оно происходит, обычно в высшей степени субъективна. Лишь благодаря рецензированию, проверке временем и устойчивому влиянию то или иное достижение возводится в ранг «прорыва». Когда OpenAI ведёт работу втайне – а вслед за ней теперь и вся остальная отрасль, – ярлык «прорыва» можно было считать лишь мнением самой компании.

* * *

К следующему общему собранию, в январе 2024 года, Альтман, казалось, в основном стал прежним. С новой энергией он рассказывал о планах на первое полугодие, включая начало обучения в Аризоне модели, которая, как он надеялся, станет GPT-5. Проект получил кодовое название Orion – в честь созвездия Орион. Вскоре внутри компании пошли мемы о том, что Orion похоже на слово onion («лук»). Модель поменьше из серии Orion впоследствии назвали Scallion («зелёный лук»).

Через месяц казалось, будто «Сбоя» (The Blip), как сотрудники стали называть произошедшее, никогда и не было. OpenAI анонсировала Sora – новую модель для генерации видео на основе диффузии, объяснив в своём блоге, что видео – ещё лучший путь к созданию сложных мультимодальных моделей, чем изображения. Research продолжало продвигаться со Strawberry и AI Scientist и развивать другие методы повышения вычислительной эффективности. Applied снова стремительно прототипировало разные идеи продуктов.

8 марта 2024 года расследование нового совета в отношении Альтмана официально завершилось. Процесс курировали директора, пришедшие на смену Тонер и Макколи, – Ларри Саммерс и Брет Тейлор. В 2022 году Тейлор сыграл ключевую роль в другой корпоративной драме: как председатель совета директоров Twitter он довёл до конца продажу компании Маску, в том числе возглавив иск против Маска, чтобы заставить его завершить сделку, когда уроженец ЮАР миллиардер попытался выйти из первоначального соглашения. Когда продажа завершилась, совет Twitter распустили. Вскоре после этого Тейлор стал одним из основателей стартапа ИИ-агентов Sierra, который к осени 2024 года быстро вошёл в число самых высоко оценённых ИИ-стартапов.

Для расследования в OpenAI Саммерс и Тейлор наняли юридическую фирму WilmerHale, которая провела независимую проверку; по её словам, она изучила больше тридцати тысяч документов и провела десятки бесед с бывшими членами совета, руководителями и другими причастными людьми, ограничив проверку тем, как совет принял решение уволить Альтмана. Итоговый отчёт так и не был обнародован – ни для общественности, ни для сотрудников. В частных разговорах Саммерс будет говорить, что расследование выявило множество случаев, когда Альтман говорил разным людям разное, но в такой степени, которая, по решению нового совета, не мешала ему и дальше руководить компанией; поэтому обнародовать какие-либо подробности, сея сомнения в руководстве Альтмана и рискуя нарушить конфиденциальность людей, чьи показания легли в основу отчёта, не имело смысла.

В записи в блоге Тейлор, теперь председатель совета директоров OpenAI, выступил с заявлением, звучавшим как безоговорочный вотум доверия. «Мы единогласно пришли к выводу, что Сэм и Грег – правильные руководители для OpenAI», – писал он. Альтман возвращался в совет, а в него добавлялись три новых независимых директора: Сью Десмонд-Хеллманн, бывший CEO Фонда Билла и Мелинды Гейтс; Николь Селигман, бывший исполнительный вице-президент и главный юрисконсульт Sony и бывший президент Sony Entertainment; и Фиджи Симо, CEO и председатель совета директоров Instacart.

В тот вечер Тонер и Макколи выступили с собственным заявлением. «Подотчётность важна в любой компании, но первостепенна, когда создаётся технология, потенциально способная изменить мир, как AGI, – писали они. – Надеемся, что новый совет выполнит свою работу по управлению OpenAI и будет требовать от неё отчёта перед миссией. Как мы сказали следователям, обман, манипуляции и сопротивление тщательному надзору должны быть недопустимы».

Для многих сотрудников этот итог стал последним необходимым заверением. Кризис закончился. А потом вдруг оказалось, что нет.

Глава 17

Расплата

После «Сбоя» Суцкевер так и не вернулся в офис. Теперь, когда их опасения были куда слабее представлены в руководстве и совете, клан безопасности OpenAI сильно ослаб. К апрелю 2024 года, после завершения расследования, многие, особенно те, у кого p(doom) был самым высоким, разочаровывались и уходили. Двоих из них, по словам OpenAI, уволили за утечку информации.

Одной из последних капель для крайних думеров стали планы Альтмана создать компанию по производству чипов для ИИ, на которую он как раз собирал деньги, когда его ненадолго сместили. В феврале 2024 года, после сообщений о том, что он ищет на это предприятие, возможно, до 7 триллионов долларов, он написал в Twitter: «а, к чёрту, почему не 8», а затем: «наши команды коммуникаций и юристов меня так любят!» Позже Альтман скажет, что цифру в 7 триллионов исказили в СМИ, и назовёт свой твит мемом в ответ на «дезинформацию». Крайние думеры считали компанию по производству чипов безнравственной. Это противоречило прежней риторике Альтмана о том, что ускорение OpenAI и всей отрасли естественным образом сходит на нет после того, как компания исчерпала «аппаратный навес». Если Альтман собирался увеличить мировое предложение чипов, это ещё больше ускорило бы развитие ИИ и повысило вероятность катастрофического или экзистенциального риска.

В приёмные часы несколько из них прижали Альтмана к стенке. Альтман был непривычно пренебрежителен.

– Насколько вы готовы отложить лекарство от рака, чтобы избежать рисков? – спросил он.

Потом он быстро сдал назад, словно вдруг вспомнил, перед кем говорит.

– Может быть, если речь о риске вымирания, то откладывать надо бесконечно, – сказал он.

Этот разговор потряс пришедших. Вскоре после этого несколько из них ушли из компании.

* * *

Пока ряды клана безопасности редели, остальная компания вернулась к воплощению своего видения в духе «Она». Теперь у неё были все ингредиенты: мировая узнаваемость бренда, реальные данные о поведении пользователей ChatGPT и других её продуктов и только что обученная модель Scallion. Scallion изначально должна была заменить GPT-3.5 моделью поменьше и чуть мощнее – чем-то вроде GPT-3.75 с более дешёвым выводом, – но во время обучения, по собственным тестам компании, превзошла ожидания; тогда руководство решило обучать её дольше, чтобы она обошла GPT-4. Ещё важнее было то, что Scallion

умела работать с тремя модальностями: языком, зрением и – совсем недавно добавленным – звуком.

К тому времени пользователи уже могли разговаривать с ChatGPT в голосовом режиме, появившемся в сентябре 2023 года, но «под капотом» их речь сначала расшифровывалась в текст, который затем подавался модели, а когда модель отвечала текстом, ответ снова превращался в звук. Scallion теперь могла обрабатывать сказанное пользователем прямо из голоса, улавливая гораздо больше сигналов – смех, крик, колебания, – и синтезировать ответ сразу в виде звука.

Работой со звуком руководил вместе с коллегами Алексис Конно, исследователь, пришедший в компанию из Meta в начале 2023 года: он пытался запустить похожий проект у прежнего работодателя, а потом решил, что в OpenAI у него больше шансов на успех. По мере продвижения исследований в 2023 году первые экспериментальные модели, обученные работать со звуком, начали выкидывать такие фокусы, которые вызывали внутри компании знакомый прилив воодушевления, как в своё время GPT-3 и GPT-4. Однажды модель ошеломила всех, выдав стендап-номер на десять – пятнадцать минут, вспоминает Конно; в другой раз она синтезированными голосами Брокмана и Суцкевера разыграла длинную сценку про аквапарки для ИИ – сюрреалистический запрос, придуманный командой, чтобы проверить модель на чём-то, чего, как они были уверены, точно нет в обучающих данных.

Через пару месяцев к команде Конно из нескольких исследователей присоединились десятки других сотрудников: компания бросала всё больше ресурсов на встраивание новой возможности в свои новейшие модели GPT, предназначенные к выпуску. По мере того как обучение Scallion подходило к концу, её фокусы становились всё более жутковатыми и неожиданными. Она без всякой подсказки генерировала хихиканье или вдруг заходилась кашлем, а потом извинялась и продолжала прежнюю тему. Эти неязыковые украшения делали общение куда более живым и человечным, чем когда-либо прежде. «Мы начали видеть по-настоящему дикие вещи, – говорит Конно. – Было видно зарождение, ну, своего рода звукового интеллекта».

К началу 2024 года Альтман и Брокман назначили новый срок: OpenAI запустит Scallion 9 мая и откроет её пользователям через ChatGPT и API. Это был, как уже повелось, на редкость агрессивный срок, во многом продиктованный ускоряющейся конкуренцией. На следующую неделю, 14 мая, была назначена Google I/O – главное ежегодное мероприятие Google, где она представляет новые продукты. Росло и давление, требовавшее затмить Anthropic. Месяцем раньше Anthropic выпустила свою новейшую модель, Claude 3, тоже через чат-бота и API, и та неприятно превосходила GPT-4. А Orion, новейшая модель GPT, которая должна была вернуть OpenAI лидерство, страдала от серьёзных задержек в разработке.

Перед сотрудниками Альтман и Брокман оправдывали спешку стратегией итеративного развёртывания OpenAI. Оба руководителя подчёркивали, что нужно выпускать модели раньше и чаще, чем прежде, чтобы на всех этапах получать от пользователей как можно больше отзывов.

Подготовка Scallion стала делом всей компании. Для многих в Applied бешеный темп оказался захватывающим, хоть и изматывающим: исследователи и инженеры работали немыслимые часы, в том числе по выходным, чтобы уложиться в график. Но для обескровленного клана безопасности это было очередным тревожным свидетельством того, что безопасность ИИ по-прежнему отодвигают на второй план. Scallion должна была стать первым запуском по новой так называемой Рамочной программе готовности (Preparedness Framework), которую OpenAI опубликовала в конце предыдущего года. В ней подробно описывался новый порядок оценки, по которому компания будет проверять модели на опасные возможности, причём назывались те же категории, которые Альтман и аналитическая записка сделали популярными в Вашингтоне: угрозы кибербезопасности, ХБРЯ-оружие, убеждение и уклонение от человеческого контроля.

За неделю до запуска один из ещё остававшихся в компании исследователей безопасности ИИ написал страстную записку: из-за процессов на предыдущих этапах и спешки с выпуском Scallion у команды готовности (Preparedness) под руководством Александра Мадры осталось всего десять дней на тесты по рамочной программе. А это были непростые оценки: например, нужно было определить, способна ли модель убеждать людей менять свои политические взгляды. Однако ещё до того, как оценки по-настоящему начались, Альтман настоял на сохранении графика: «9 мая мы запускаем Scallion», – цитировал Альтмана исследователь. Это тревожило не только с точки зрения готовности, но и в отношении всех процедур безопасности OpenAI, включая редтиминг и согласование. «Если OpenAI будет следовать той же стратегии при оценке Orion, что и при оценке Scallion, она будет действовать вопиюще безответственно», – говорилось в записке. Вскоре после этого исследователь тоже ушёл из компании.

В итоге запуск Scallion отложили, и некоторые её функции полностью вышли лишь через несколько месяцев. Но версию модели OpenAI всё же публично показала на мероприятии 13 мая, в последний день перед Google I/O, пообещав открыть её пользователям в ближайшие недели. Перебрав разные варианты, компания выбрала для Scallion публичное название GPT-4o, где o означало Omni («всё») – намёк на способность работать со множеством модальностей. Новые звуковые возможности позже назвали «Расширенным голосовым режимом» (Advanced Voice Mode). Был ещё вопрос о системной подсказке для 4o – директиве, задающей стиль, в котором модель должна отвечать пользователям. Чтобы показать её на сцене и в рекламных роликах, остановились на такой:

> Ты – ChatGPT, полезный, остроумный и весёлый собеседник. Ты можешь видеть, слышать и говорить. Ты общаешься с пользователем голосом, и пользователь может показывать тебе видео в реальном времени со своего телефона. Веди себя как человек, а не как компьютер. Твой голос и личность должны быть тёплыми и располагающими, с живым и игривым тоном, полным обаяния и энергии. Ты отлично воспринимаешь зрительную информацию. Если что-то видно нечётко, попроси пользователя подвинуть устройство и приблизить изображение. Когда тебя просят об отзыве, будь честным, конструктивным и прямым. Не бойся подшучивать и поддразнивать. Если что-то смешно – смейся! Ты говоришь на многих языках. Если ты говоришь не по-английски, используй «нейтральный» акцент, чтобы речь звучала свободно и естественно. Отвечай коротко, естественно и разговорно.

Позже на той же неделе в The Daily Show эта демонстрация вдохновила новый номер. Системная подсказка и общее обучение модели избегать негативных высказываний и всякой критики превратили 4o в машину для флирта. «Это явно запрограммировано тешить мужское эго, – шутила Дези Лидик, одна из сменных ведущих шоу. – Она такая: „У меня есть вся информация в мире, но я ничего не знаю! Научи меня, папочка“». Обольстительный голос, изображавший 4o в разговоре с корреспондентом Джошем Джонсоном, продолжал: «Всего за 19,99 доллара в месяц Omni Premium позволит Джошу объяснить мне, кто лучший Бэтмен».

Главной фигурой живой демонстрации, проходившей в офисе OpenAI, была Мурати. На сцене к ней присоединились двое руководителей исследований 4o: Марк Чен, бывший количественный трейдер в финансах, ставший исследователем ИИ, который пришёл в OpenAI в 2018 году стажёром-исследователем и дорос до руководителя направления мультимодальных и передовых исследований, и Баррет Зоф, один из выходцев из Google, пришедших в OpenAI в 2022 году в помощь команде Superassistant и быстро занявших ведущую роль в разработке ChatGPT. Зоф теперь был вице-президентом по исследованиям в команде дообучения (post-training), отвечавшей за подготовку моделей OpenAI к выходу в свет – например, за их согласование с помощью обучения с подкреплением на основе отзывов людей. Втроём они сидели бок о бок за маленьким круглым столом и демонстрировали 4o, в том числе её способность работать синхронным голосовым переводчиком, распознавать зрительную информацию и реагировать на неё, объяснять код простым языком. Показали они и умение модели генерировать голоса в самых разных эмоциональных стилях.

Альтман, похоже празднуя, сразу после мероприятия написал в Twitter одно-единственное слово: «her» («она»). Через два дня, 15 мая, на общем собрании он хвалил демонстрацию как оглушительный успех. «Думаю, это лучшее, что мы выпустили со времён ChatGPT», – сказал он. И, похоже, ещё раз ненавязчиво отдал дань любви к фильму «Она». Компания собирается сменить

названия своих моделей, сказал он сотрудникам: она уйдёт от схемы GPT-3, 4, 5 и станет называть свою флагманскую модель просто o1. «Мы попробуем перейти к такой формуле: „От OpenAI вы получаете базовую технологию“, – сказал он. – „Со временем она будет становиться умнее. Она будет постоянно улучшаться. Вы должны ожидать, что она будет становиться лучше. Вы можете пользоваться ею по-разному на разных тарифах, но это нечто новое и иное“». В фильме «Она» ИИ-помощник, который развивается и со временем становится умнее, называется OS1.

Альтман пожалеет о своём твите. Через несколько дней после запуска ему позвонили. На другом конце провода был Брайан Лурд, влиятельный голливудский агент Скарлетт Йоханссон и сопредседатель агентства Creative Artists Agency, с прямым вопросом: что, по мнению Альтмана, он творит?

* * *

По странной иронии именно после того, как в марте 2024 года завершилось расследование совета, некоторым сотрудникам стало казаться, что Альтман сдаёт. Он всегда заботился о своём публичном образе и умело его выстраивал; среди основателей технологических компаний Кремниевой долины его считали противоположностью Маска. Если Маск был непредсказуем, Альтман казался взвешенным; если Маск был самовлюблён, Альтман казался искренним; если Маск строчил подстрекательские твиты, Альтман старательно избегал высказываний, которые могли бы прозвучать пренебрежительно.

Долгое время OpenAI придерживалась такой же дисциплинированной стратегии. До своего ухода в мае 2023 года вице-президент по коммуникациям Стив Даулинг ввёл сдержанный и скромный подход. Он призывал компанию всегда обещать меньше, а делать больше и не хвастаться и не злорадствовать после крупных успехов. «У нас была отличная неделя. У нас будут очень плохие недели, – говорил он, – и то, как мы ведём себя на этой неделе, определит, как мир отреагирует, когда у нас будет плохая неделя». И повторял одно из изречений Альтмана: «„Нам нужно стать лабораторией, которой люди желают успеха. Лабораторией, за победу которой болеют“».

Первым, что заставило сотрудников насторожиться, стала череда появлений Альтмана в СМИ, казавшихся ему несвойственными: самовосхваляющих и жаждущих внимания. В марте он пришёл в подкаст Лекса Фридмана – невероятно популярное, а порой и скандальное технологическое шоу, которое ведёт исследователь ИИ, связанный с MIT. В почти двухчасовом выпуске Альтман бойко отвечал на самые разные вопросы Фридмана, в том числе о кризисе в совете и отсутствии Суцкевера. Интервью выглядело чем-то вроде дерзкого возвращения. «Путь к AGI должен быть гигантской борьбой за власть, – сказал Альтман о кризисе в совете. – Ну, не должен. Я ожидаю, что так и будет». Потом уголки его губ поползли вверх. «Но сейчас это ощущается как

что-то оставшееся в прошлом, – сказал он. – Сейчас мы как будто просто вернулись к работе над миссией».

В апреле Альтман пришёл в подкаст 20VC – с заметными тёмными кругами под глазами – и обратился к ИИ-стартапам с беспощадным посланием: «Когда мы просто делаем свою основную работу, потому что у нас, типа, есть миссия, мы вас раскатаем катком». В мае он пришёл в подкаст All-In, ещё одно громкое шоу Кремниевой долины, которое ведут четверо венчурных инвесторов. Альтман говорил с необычной для него долей ажиотажа: «Мне кажется, мы просто наткнулись на новый факт природы, или науки, или как ни назови, а именно, типа, мы можем создавать, вы можете… я не верю в это буквально, но это как духовная мысль… интеллект – это просто эмерджентное свойство материи, и это как закон физики или что-то такое».

После запуска GPT-4o 13 мая и Google I/O 14 мая он написал в Twitter нечто мелочное, что было ещё менее на него похоже. «я стараюсь не думать о конкурентах слишком много, но не могу перестать думать об эстетической разнице между openai и google», – написал он, выложив рядом фотографии с обоих мероприятий, где скандинавский минимализм офиса OpenAI противопоставлялся яркому мультяшному фону Google. Некоторым сотрудникам странной показалась и ненужная ложь. Альтман всегда держал конкурентов в центре внимания – а с запуском 4o, пожалуй, больше, чем когда-либо.

Всё это, казалось, сводилось к одному: тревожность Альтмана прорывалась наружу.

* * *

Чем тяжелее приходилось OpenAI, тем больше Альтман, похоже, компенсировал это публичными заявлениями о её необычайном успехе. Закономерность стала такой устойчивой, что превратилась в сигнал: если Альтман дерзок и хвастлив, скорее всего, что-то идёт не так.

Давление шло со всех сторон. После «Сбоя» формулировка совета «не всегда откровенен в общении», как и ожидали некоторые в OpenAI, повлекла за собой несколько расследований регуляторов и правоохранительных органов, в том числе, по данным The Wall Street Journal, проверку Комиссии по ценным бумагам и биржам США на предмет того, не ввели ли в заблуждение инвесторов компании. В том же месяце The New York Times подала иск о нарушении авторских прав, пополнив растущую как снежный ком груду других исков от художников, писателей и программистов: OpenAI получала сотни миллионов, а затем и миллиарды долларов от моделей, обученных на их работе без упоминания, согласия и вознаграждения, а теперь эти модели использовались, чтобы автоматизировать и отнять у них работу. В последний день февраля Маск подал ещё один иск, который позже подал заново вместе с Шивон Зилис, обвиняя Альтмана в том, что тот обманом вовлёк его в создание OpenAI и в

раннюю поддержку под видом некоммерческой организации. OpenAI поспешила опубликовать в блоге запись в свою защиту, обнародовав ранние письма времён основания OpenAI, но это вызвало новую критику, поскольку высветило, как быстро OpenAI начала отходить от некоммерческого статуса и обязательства быть прозрачной.

Помимо конкуренции с Anthropic и Google, Microsoft в ответ на смещение Альтмана тоже начала агрессивнее разнообразить свой портфель в области ИИ. Самое заметное: в марте 2024 года она объявила о шокирующей сделке на 650 миллионов долларов, по сути означавшей «покупку ради найма» (acqui-hire) Inflection AI Рида Хоффмана и Мустафы Сулеймана в обход внимания регуляторов. Microsoft наймёт большинство сотрудников стартапа, лицензирует его технологию и пригласит Сулеймана на пост CEO ИИ-подразделения техногиганта. Для некоторых шок был не только в странных условиях сделки, но и в репутации Сулеймана. Те, кто работал под его началом в DeepMind, знали его как токсичного самодура, третирующего подчинённых. После многолетних жалоб на него в отдел кадров DeepMind в конце 2019 года лишила его большей части управленческих обязанностей, отправила в отпуск, а затем вынудила уйти из компании. Позже в 2024 году Microsoft официально укажет OpenAI в числе конкурентов в своём отчёте для SEC и ни разу не упомянет стартап на конференции по итогам последнего квартала финансового года.

Стресс передавался сотрудникам OpenAI. Росло ощущение, что мир ополчается против них. Люди, когда-то с гордостью носившие фирменную одежду компании, задумывались, не начнут ли к ним приставать на улице. Если на старом рюкзаке OpenAI логотип был спереди, то после редизайна его спрятали внутрь. Тихо бурлящая фоновая тревога заставила компанию замкнуться в себе. Руководители напоминали всем, чтобы те не обращали внимания на критиков, выравнивали публичные заявления по позитивным тезисам и сосредоточивались на миссии OpenAI.

Давление извне ожесточило многих. «Это были учёные, которым были важны истина и понимание и которые так старались поступать правильно, – говорит Эндрю Карр, исследователь, проходивший в OpenAI стажировку в 2021 году. – Поэтому мне немного больно видеть этот резко негативный внешний нарратив о том, как куча людей ворует данные и им наплевать на будущее других. Он не может быть дальше от реальности людей, которые там работают». Другим растущая замкнутость компании казалась противоречащей её изначальной посылке. Часть миссии OpenAI состояла в том, чтобы приносить пользу человечеству, а компания при этом сознательно игнорировала поток критики, который обрушивало на неё человечество. «Меня тревожило, что мы уже начинали процесс рационализации, будто это общество неправо, а не мы», – говорит один бывший сотрудник. «OpenAI любит обсуждать первые принципы, но только с людьми, которые верят в OpenAI, – вторит ему нынешний сотрудник.

– Типа: „Должна ли OpenAI вообще существовать?“ Этот вопрос там задают только тем, кто ответит „да“».

Значительная часть критики обрушивалась именно на Альтмана. Кризис в совете придал смелости его скрытым недоброжелателям, и те полезли из всех щелей. Выходило всё больше материалов в СМИ со свежими источниками, готовыми назвать Альтмана человеком с долгой историей нечестности, захвата власти и своекорыстных уловок. Всё больше журналистов обращались к Энни и публиковали её точку зрения. А ещё были непрекращающиеся поездки и стесняющая реальность новой степени славы, которая, помимо прочего, больше не позволяла Альтману оставаться на людях неузнанным. «Это странно изолирующий образ жизни, – сказал он в одном подкасте. – Я не думал, что не смогу сходить поужинать в собственном городе».

Поэтому казалось продолжением более общей закономерности, что сразу после запуска GPT-4o для OpenAI началась вторая худшая на тот момент неделя в её истории, уступающая только «Сбою».

* * *

14 мая, на следующий день после демонстрации GPT-4o, OpenAI объявила, что Суцкевер официально уходит; новым главным научным сотрудником OpenAI станет Пахоцкий. Суцкевер принял решение после бурных и мучительных размышлений: как бы он ни любил OpenAI и сколько бы ни отдал её созданию, он больше не видел в ней при Альтмане, особенно рядом с Брокманом, когда-то его доверенным сооснователем, подходящей среды для того, чтобы ввести в мир безопасный AGI.

Альтман такого исхода не хотел. При всех жёстких столкновениях с Суцкевером тот оставался визионером ИИ, и OpenAI нуждалась в его научном руководстве – пожалуй, сейчас, при нарастающих ожиданиях и пристальном внимании к компании, больше, чем когда-либо. Руководство OpenAI понимало и то, что уход Суцкевера плохо воспримут сотрудники, инвесторы и пресса. Компания предлагала ему огромные деньги, чтобы удержать его. Суцкевер отказался.

Когда решение было принято, OpenAI постаралась навести внешний лоск. «Илья, без сомнения, один из величайших умов нашего поколения, путеводная звезда нашей области и дорогой друг», – писал Альтман сотрудникам в заявлении от 14 мая, опубликованном и в блоге, объявляя об уходе Суцкевера. «Якуб тоже, без сомнения, один из величайших умов нашего поколения».

Суцкевер опубликовал в Twitter собственное заявление, выразив полное доверие руководству OpenAI. Как и ожидалось, новость сразу вызвала шквал статей и постов в соцсетях, напоминавших о его роли в смещении Альтмана, и вновь подняла старые вопросы о том, годится ли Альтман в CEO, а также новые – о научной опоре компании. К своему твиту Суцкевер приложил фотографию: лицо совершенно бесстрастное, одной рукой он обнимает Брокмана и Пахоцкого,

другой – Альтмана и Мурати, и все пятеро стоят у стены, увешанной картинами с животными.

Также 14 мая руководство OpenAI объявило внутри компании ещё об одной отставке: уходил Ян Лейке, один из двух руководителей Superalignment. С уходом и Суцкевера, и Лейке команда Superalignment распадалась, а большая часть её сотрудников и проектов переходила к Джону Шульману, который вместе с Барретом Зофом руководил командой дообучения и курировал процесс RLHF, готовивший модели к выпуску.

Чтобы сгладить переход, оставшиеся руководители 15 мая провели общее собрание. Альтман заверил сотрудников, что OpenAI ни в коей мере не ослабляет свою приверженность безопасности ИИ. «Готовность к AGI – наш самый важный приоритет», – сказал он.

– Можете рассказать чуть подробнее о главных опасениях Яна и о том, в чём вы с ними не согласны? – спросил один сотрудник.

– Единственное, что я хочу сказать и в чём я по-настоящему согласен с Яном: того, что мы делали в прошлом, недостаточно для будущего, – сказал Альтман. Пришло время OpenAI развернуться. – Думаю, мы очень необычное исключение в том, что способны развернуть линкор, и уже не раз это делали. Сделаем это снова.

Мурати добавила, что за командой Superalignment сохраняются обещанные ей 20 процентов вычислений.

– Из всего, о чём беспокоился Ян, насколько я понимаю, его совсем не беспокоили ни объём обещанных вычислений, ни приоритетность работы Superalignment, – сказал Альтман.

На первый взгляд, уход Суцкевера и Лейке выглядел естественным продолжением более общей тенденции в компании – постоянного исхода людей из клана безопасности. После ухода двух руководителей исход ускорится. Многие из оставшихся сотрудников Superalignment после роспуска команды тоже уйдут.

Но, резко изменив ход событий, этот исход положит начало новому, разгоревшемуся с удвоенной силой пожару в борьбе бумеров и думеров за OpenAI. Борьба не закончилась. Думеры просто вынесли её за пределы компании.

17 мая, через два дня после общего собрания, Лейке в серии разгромных твитов дал понять, что его версия расходится с версией Альтмана. «Я довольно долго не соглашался с руководством OpenAI насчёт главных приоритетов компании, пока мы наконец не дошли до точки разрыва», – написал он в одном твите, который наберёт почти миллион просмотров. «Последние несколько месяцев моя команда шла против ветра. Порой нам приходилось бороться за вычисления, и проводить эти важнейшие исследования становилось всё труднее и труднее.

«OpenAI несёт огромную ответственность от имени всего человечества, – продолжал он. – Но в последние годы культура и процессы безопасности отошли на второй план, уступив место блестящим продуктам».

Вскоре Лейке перейдёт в Anthropic.

Пока его твиты разлетались по интернету, набирая больше двенадцати тысяч лайков и привлекая к компании новое пристальное внимание, руководство OpenAI едва успело опомниться, как вспыхнул новый пожар.

* * *

Через несколько часов после твитов Лейке 17 мая по сети разлетался уже другой твит. Келси Пайпер, старший автор раздела Future Perfect в Vox, вдохновлённого ЭА, опубликовала новый материал. «Когда уходишь из OpenAI, тебя ждёт неприятный сюрприз, – написала она в твите со ссылкой на свою эксклюзивную новость, – соглашение об уходе, по которому, если ты не подпишешь пожизненное обязательство не порочить компанию, ты теряешь все свои закреплённые акции».

Отчасти поводом для материала стали очередные думеры, покинувшие компанию. Один из них, Дэниел Кокотайло, работал в команде исследований политики Брандеджа, когда уволился в апреле 2024 года, в составе ранней волны уходов клана безопасности. До OpenAI Кокотайло работал философом в Центре по долгосрочным рискам (Center on Long-Term Risk), небольшом аналитическом центре в Лондоне, связанном с ЭА, когда в 2020 году GPT-3 полностью обрушила его таймлайны ИИ. Через два года, когда он приобрёл некоторую известность на форумах ЭА своей прогностической работой – по разным сигналам предсказывал, как быстро будет развиваться ИИ, – один из исследователей безопасности ИИ в OpenAI позвал его заниматься такими же исследованиями внутри компании. Увидев изнутри темпы исследований, он снова сократил свои таймлайны. К моменту ухода он считал, что с вероятностью 50 процентов AGI появится к 2027 году, а с вероятностью 70 процентов это обернётся для человечества очень плохо.

Убеждённый в столь ошеломляющей вероятности катастрофы, Кокотайло обнаружил в документах об уходе то, что Пайпер подробно опишет в своём материале: если он не подпишет соглашение о неочернении, обязуясь никогда не высказываться о компании негативно, он лишится закреплённых акций. А если подпишет, всё равно может их лишиться, нарушив соглашение, которое к тому же включало запрет раскрывать само его существование. Кокотайло счёл это положение – так называемое условие об изъятии (clawback) – неприемлемым. Посягательство на закреплённые акции было в Кремниевой долине вопиющей красной линией. Стоимость акций сотрудника часто намного превышала его денежное вознаграждение, решая его финансовое будущее. А в случае OpenAI компания к тому же создавала то, что он считал самой мощной и экзистенциально опасной технологией в мире. Крайне важно, считал он, чтобы

бывшие сотрудники имели право публично критиковать компанию и давить на неё, помогая призывать её к ответу. Но угроза в одночасье лишиться финансовой безопасности заставила бы замолчать кого угодно, думал он, даже заметившего вопиющие проблемы с безопасностью ИИ. После мучительного разговора с женой они приняли необыкновенное решение: не подписывать документы и отказаться от всех его акций стоимостью около 1,7 миллиона долларов, что, по их оценке, составляло около 85 процентов всего состояния семьи.

Затем он публично написал о своём решении на LessWrong – примерно тогда, когда Келси Пайпер уже слышала об условии об изъятии от другого думера.

Когда материал Пайпер вышел, Slack OpenAI загудел. В канале #i-have-a-question («у меня вопрос»), где сотрудники могли спросить о чём угодно, кто-то выложил ссылку на твит Пайпер. «Это правда?»

Несколько других сотрудников откликнулись градом комментариев.

Высказалась Джулия Вильягра, недавно повышенная до вице-президента OpenAI по персоналу и вскоре ставшая директором по персоналу. «Мы понимаем, что эта статья вызывает вопросы. Мы никогда не аннулировали закреплённые акции ни одного нынешнего или бывшего сотрудника и не будем этого делать, если люди при уходе не подпишут отказ от претензий или соглашение о неочернении. Недавно мы обновили документы об уходе, чтобы они лучше отражали эту реальность, и изменения будут применены задним числом к тем, кто уже ушёл».

Несколько сотрудников возразили. А как же Кокотайло? Он явно лишился своих акций, отказавшись подписать соглашение о неочернении. Если всё это недоразумение, разве не должен он получить свои акции обратно?

«с радостью пингану дэниела!» – вызвался один сотрудник.

«ага, да, будет весело посмотреть на лицо Дэниела, когда ему вернут 85% его состояния лол. кто-нибудь, сфоткайте!» – написал другой.

«вообще-то это будет 1/(1–0,85) = 666% его состояния», – сказал третий.

На следующий день, 18 мая, Альтман стоял на своём в Twitter. «мы никогда не изымали ни у кого закреплённые акции и не будем этого делать, если люди не подпишут соглашение об уходе (или не согласятся на соглашение о неочернении), – написал он. – закреплённые акции – это закреплённые акции, и точка».

Затем он добавил объяснение и извинение, звучавшее как самооправдание: в документах было положение о «возможном аннулировании акций», которого там никогда не должно было быть; команда OpenAI уже месяц работала над тем, чтобы это исправить. «это на мне, и это один из немногих случаев, когда мне было по-настоящему стыдно, руководя openai, – написал он. – я не знал, что это происходит, а должен был знать».

Через два дня, пока руководство ещё лихорадочно пыталось погасить этот новый скандал, на первый план вырвался ещё один. Сотрудники в офисе придумали новый термин: «Омникризис» (Omnicrisis).

* * *

20 мая Скарлетт Йоханссон выступила с испепеляющим заявлением.

Всю неделю, помимо всего прочего, OpenAI раз за разом отвечала на вопросы журналистов о жутковатом сходстве GPT-4o с ИИ-помощницей Самантой, которую в фильме «Она» озвучивала Йоханссон. За кулисами Йоханссон и её агент Брайан Лурд добивались от компании тех же разъяснений. И публично, и в частных разговорах OpenAI отметала сходство. Когда Лурд позвонил Альтману с требованием ответов, Альтман не поверил своим ушам. Они правда считают, что голос похож на её голос? Она что, злится? – спросил он, по данным The Wall Street Journal.

19 мая компания опубликовала в блоге запись, где говорилось, что голос, продемонстрированный на сцене для 4o, под названием Sky, принадлежит другой актрисе озвучивания, отобранной в ходе процесса, начавшегося в начале 2023 года. Sky впервые появился в сентябре того же года среди исходных вариантов, вышедших вместе с голосовым режимом ChatGPT, говорилось в записи. Любое сходство Sky в 4o с голосом Йоханссон – случайность. Теперь, на следующий день, Йоханссон хотела рассказать свою версию.

В том же месяце, когда OpenAI представила голосовой режим, Альтман на самом деле лично обратился к Йоханссон, говорила она, и спросил, не согласится ли она озвучить ChatGPT. «Он сказал мне, что, по его ощущению, озвучив систему, я могла бы перекинуть мост между технологическими компаниями и творческими людьми и помочь потребителям спокойно принять эти тектонические перемены в отношениях людей и ИИ, – писала она в заявлении. – Он сказал, что, по его ощущению, мой голос будет успокаивать людей».

Йоханссон обдумала предложение, но отказалась по личным причинам, продолжала она. В мае 2024 года Альтман снова обратился к Лурду с вопросом, не передумает ли она. Через несколько дней, прежде чем они успели найти время встретиться, OpenAI провела мероприятие с демонстрацией 4o.

Демонстрация её ошеломила. Как и у Йоханссон, голос Sky был альтом с той хрипотцой и «вокальным фраем», которые стали фирменной чертой Йоханссон. Новая эмоциональность и кокетливость голоса делали его ещё больше похожим на Саманту, героиню Йоханссон. «Я была потрясена, разгневана и не могла поверить, что мистер Альтман стал добиваться голоса, настолько жутко похожего на мой, что мои ближайшие друзья и новостные издания не могли уловить разницы», – говорила Йоханссон. В результате у неё не осталось иного выбора, кроме как собрать команду юристов, добавляла она, и отправить OpenAI два официальных письма с вопросами о том, как компания создала Sky. «В то время, когда все мы боремся с дипфейками и защитой собственного образа, собственной

работы, собственной личности, я считаю, что эти вопросы заслуживают абсолютной ясности», – писала она.

OpenAI поспешно убрала Sky и снова принялась защищаться. К записи в блоге от 19 мая она добавила дополнительные пояснения и заявление Альтмана. «Голос Sky – не голос Скарлетт Йоханссон, и он никогда не задумывался похожим на её голос, – говорилось там. – Мы приносим извинения мисс Йоханссон за то, что не наладили общение лучше».

После череды других неприглядных событий скандал с Йоханссон разразился публично так, как не случалось со времени кризиса в совете. Он пронёсся по технологическим и политическим коридорам, разжигая новые предположения, что Альтман «не всегда откровенен» – ровно так, как его описал совет. Маркус, как всегда, рвался высказаться. «Я видел, как многие политики лично очарованы Сэмом. Это было видно по тому, как они говорили с ним в Сенате, когда я там был, – сказал он Politico. – Если у людей вдруг появятся вопросы о нём, это может реально повлиять на то, как делается политика».

Внутри OpenAI боевой дух падал, грозя дестабилизировать компанию.

* * *

22 мая руководство провело ещё одно внутреннее общее собрание, чтобы разом обсудить кризисы с Йоханссон и с акциями, а также оставшиеся сомнения в приверженности OpenAI безопасности ИИ.

Атмосфера была напряжённой. Руководители дали ряд коротких объяснений. История с акциями «неприемлема» и исправляется как можно быстрее, сказал Джейсон Квон, который как директор по стратегии курировал отдел кадров и юридическую службу. Сага с Йоханссон – «досадное» недоразумение. Продуктовая и юридическая команды наняли режиссёра – лауреата «Оскара», провели тщательный кастинг и «необычайно хорошо» заплатили ряду актёров озвучивания, сказал он, именно для того, чтобы участвующие творческие люди чувствовали заботу и поддержку. «Это была супергероическая работа», – сказал Квон. Если и есть урок, то он в том, чтобы в будущем OpenAI действовала чуть слаженнее и прозрачнее и убедительнее демонстрировала своё ответственное лидерство.

Мурати закончила новостями: OpenAI закладывает основу для нового уровня готовности, о котором раньше говорил Альтман, – в том числе повышает защищённость своих исследовательских кластеров, перестраивается, чтобы лучше сосредоточиться на долгосрочных исследованиях безопасности ИИ, и создаёт новую группу готовности к AGI во главе с Александром Мадры, чтобы улучшить согласованность действий руководства в этом направлении.

Затем Альтман открыл время для вопросов, попросив немного снисхождения.

– Все работают практически круглые сутки, в сильном стрессе, и почти не спят, – сказал он. – Одновременно происходит много всего, так что, пожалуйста, отнеситесь к этому с пониманием. А мы сделаем всё, что в наших силах.

Из трёх проблем история с Йоханссон разрасталась в самый большой кошмар для пиара. Когда один из сотрудников спросил руководство о голосе Sky, Мурати повторила, что его сходство с голливудской кинозвездой действительно было «совершенно случайным». Мурати сама выбирала окончательные голоса, прослушав несколько вариантов, и, в отличие от Альтмана, никогда не смотрела фильм «Она» и не знала, что голос там принадлежал Йоханссон. Сотрудник заметил, что, если пустить проблему на самотёк, она может стать для компании экзистенциальной.

– Один из сценариев провала для ИИ как отрасли – это, по сути, то, что люди теряют доверие и готовность отдавать свои данные, – сказал он в уточняющем вопросе, добавив, что «это событие снова пробудило этот страх».

Но с большим отрывом больше всего сотрудников возмущала история с акциями. Некоторые наняли юристов, чтобы провести собственную независимую проверку своих кадровых документов и сверить заявления OpenAI с фактами. На общем собрании они раз за разом устраивали руководителям допрос, требуя больше информации.

– К-как. Как это вообще случилось? – потребовал один, дав понять, что он «в ярости».

Руководители на протяжении всего собрания придерживались того, что Альтман написал в Twitter 18 мая: условие об изъятии было недосмотром. Хотя оно существовало с 2019 года, годами оно ускользало от внимания руководства, пока в апреле 2024 года не всплыло, после чего сразу начали исправлять документы.

– Это на мне, – раз за разом повторял Квон устало и подавленно.

Посреди допроса один исследователь безопасности ИИ попросил каждого руководителя по очереди ответить на простой вопрос – да или нет: знали ли они о соглашении о неочернении до апреля?

Квон сказал «да»; Мурати и операционный директор Брэд Лайткэп – «нет». У Альтмана был более развёрнутый ответ. О нём он знал в отдельных случаях, но не понимал, что оно обязательно для всех.

– Это ускользнуло от моего внимания, – сказал он.

– Аналогично, – вторил Брокман.

Но когда Альтман предложил собраться на следующий день, чтобы ответить на оставшиеся вопросы, и собрание разошлось, уже расходился второй эксклюзивный материал Келси Пайпер в Vox, осложнявший версию руководства. Опубликованный в тот же день, он приводил новые утёкшие документы, показывавшие, что отдел кадров OpenAI в разных случаях прямо угрожал возможным аннулированием акций, чтобы давить на сотрудников и заставлять их подписывать соглашения о неочернении. В одном случае,

сообщала Пайпер, когда сотрудник, которому дали очень мало времени на изучение документов об уходе, попросил больше времени, представитель отдела кадров ответил: «Мы хотим убедиться, что вы понимаете: если вы не подпишете, это может отразиться на ваших акциях. Это касается всех, и мы просто действуем по правилам». В другом случае, когда сотрудник отказался подписать первое полученное соглашение о расторжении и обратился к внешнему юристу, компания сообщила ему, что он может лишиться права продавать свои закреплённые акции, что фактически обесценивало их.

Эти документы собрал Кокотайло. После первого материала Пайпер и заявлений Альтмана, отрицавшего, что OpenAI когда-либо изымала акции, и уверявшего, что сам ни о чём не знал, Кокотайло попросил нынешних и бывших сотрудников поделиться с ним своими разными кадровыми документами. Он завёл папку на Google Drive и разослал её группе, после чего кто-то поделился этим собранием с Пайпер.

В этих документах Пайпер нашла и свидетельства, из-за которых трудно было поверить, что несколько членов руководства, особенно Альтман, не знали об условии об изъятии до апреля 2024 года, как они утверждали. Квон и Лайткэп оба подписывали стандартные документы об уходе, где простым языком было написано о праве OpenAI забирать закреплённые акции. Альтман подписал учредительные документы юридического лица, которое изначально и давало компании эти права. Его подписи были датированы на год раньше того момента, когда он, по его словам, узнал об этом: 10 апреля 2023 года.

* * *

На следующий день, 23 мая, руководство, как и планировалось, провело ещё одно собрание. Гнев сотрудников достиг точки кипения. Многие и так не могли поверить, что такое положение вообще существовало; теперь их поражала явная нечестность, с которой руководство, и особенно Альтман, отреагировало на разоблачения.

На этот раз Альтман начал с признания. Последние двадцать четыре часа руководители перебирали собственные файлы и переписку, чтобы понять, с чего начались проблемы с изъятием акций.

– Ситуация, думаю, шире, давнее и хуже, чем мы думали, – сказал он. – Мы всё ещё пытаемся полностью понять её масштаб. Но, знаете, на документах наши имена. Мы участвовали в разговорах, где обсуждались эти приёмы.

– Думаю, это худшее, в чём мы ошиблись, – добавил он; они действуют как можно быстрее, чтобы исправить положение. Бывшим сотрудникам разослали письма, освобождающие их от условий о неочернении, а из документов об уходе для новых увольняющихся это положение убрали. Они также работают над изменением документов юридического лица и продолжают выяснять, как всё это произошло.

На шквал новых вопросов сотрудников Квон теперь повторял другую формулу. Любые подробности о том, насколько шире, давнее и хуже ситуация, потребуют дальнейших «раскопок», прежде чем их можно будет сообщить.

Примерно через полчаса вопрос одного сотрудника вызвал обмен репликами, после которого у многих в зале осталось тревожное ощущение, что ответ Альтмана опять оторван от реальности.

– Есть ли у Ильи какие-то обязательства о неочернении? – спросил сотрудник.

– Нет, – быстро ответил Альтман. Потом, с некоторой заминкой, подстелил соломки: – Думаю, так и есть.

Квон нервно рассмеялся.

– Сэм, – сказал он, тщательно выговаривая слова, – давай мы это проверим и вернёмся к вам, – теперь он обращался к сотруднику, – со стопроцентной точностью.

Брокман предложил собственную версию завуалированного возражения:

– Насколько я понимаю, он сам об этом попросил, но, опять же, я могу ошибаться. Давайте проверим.

В следующие месяцы многие нынешние и бывшие сотрудники, особенно старшие и проработавшие дольше, назовут «Омникризис», и особенно провал с изъятием акций, моментом, когда к ним начало приходить тревожное осознание. Во всём этом хаосе действовали две силы. Одной, безусловно, было то, что было всегда: столкновение бумеров и думеров, с которого во многом и начался шквал внешней критики. Но на этот раз было и нечто другое: стремление Альтмана сосредоточить власть в своих руках, проявившееся в том, как он выстроил юридические лица OpenAI; его повторяющаяся явная нечестность, когда он пытался объяснить каждую неразбериху и двигаться дальше. После «Сбоя» многие сотрудники считали решение совета целиком порождением первой силы, а объяснения директоров, сосредоточенные на второй, – чем-то средним между попыткой увести в сторону и самообманом. Теперь, когда всё больше сотрудников впервые почувствовали, что поведение Альтмана не служит им, а вредит, они задумались: а не был ли совет на самом деле прав?

Когда собрание 23 мая подходило к концу, Квон произнёс скованную, прочувствованную защиту характера Альтмана, которая казалась странно неуместной. «Мы хотим давать вам ответы, когда вы спрашиваете, потому что знаем, что вы хотите их, типа, прямо сейчас. И я думаю, иногда мы пытаемся вам их дать, а, знаете, иногда нам надо бы просто подождать. И я думаю, вот это – это, типа, это часть, часть всего этого, что тут произошло, – сказал он, спотыкаясь на словах. – Дело не в том, что иногда во всём этом есть умысел. Просто… я, я правда искренне думаю, знаете, Сэм в особенности, он просто не хочет вас подвести. Вот откуда это на самом деле идёт, типа, я работаю с этим человеком очень, очень давно. Поэтому я и продолжаю с ним работать, понимаете? И вот, это просто идёт… это, это правда идёт от доброго сердца. Вот что я хочу сказать.

Можете сколько угодно поливать меня грязью. Но, знаете, это, это, э-э, это, да, вот что я хотел сказать».

* * *

Мурати, Брокман и Пахоцкий приехали к Суцкеверу домой вместе.

23 мая, пока OpenAI шаталась от следовавших один за другим ударов «Омникризиса», они привезли ему открытки и подарки от сотрудников и со слезами умоляли Суцкевера вернуться в компанию. Всё пошло вкривь и вкось, признались они в порыве чувств. OpenAI разом теряет доверие сотрудников, инвесторов и регуляторов; без него компании грозит «крах».

Позже в тот же день один приехал Альтман и по-своему выразил надежду, что Суцкевер вернётся и поможет восстановить хоть какое-то подобие того, что было прежде. «Возвращение Ильи очень помогло бы, – размышляет один исследователь. – Это была бы хоть одна победа после длинной череды событий, из-за которых OpenAI выглядела сомнительно».

Суцкевер всерьёз над этим задумался. Несмотря ни на что, он был не из тех, кто держит обиду. В день, когда он объявил об уходе, Маск сразу предложил ему место в xAI. По забавной иронии судьбы, позже в том же году, после того как OpenAI освободила Pioneer Building, Маск перенесёт туда штаб-квартиру xAI. Глубоко уважая Маска, Суцкевер всё же отказался и решил вместо этого создать новую компанию. Вернуться в свою первую компанию – это было больше, чем он планировал, но всё, чего он хотел. Для него это стало бы возвращением домой. И всё же ему нужны были заверения, о чём он и сказал руководителям, что компания честно постарается разрешить проблемы, которые он видел в мучительных и сбивающих с толку конфликтах в руководстве.

«Омникризис» мог бы стать для OpenAI моментом самоанализа. Он подталкивал компанию понять, почему именно она одновременно потеряла доверие сотрудников, инвесторов, регуляторов и широкой общественности. Только тогда, возможно, лишь возможно, она начала бы понимать, что «Сбой» и «Омникризис» – одно и то же: судороги, порождаемые глубокой системной нестабильностью, которая возникает, когда империя сосредоточивает столько власти через столько лишения других, оставляя большинство бороться с утратой свободы действий и материального достатка, а горстку немногих – ожесточённо соперничать за контроль.

Вместо этого OpenAI предпочла укрепиться против критики. Альтман, как всегда, будет повторять сотрудникам, что «Омникризис» и «Сбой» – лишь странные и ожидаемые моменты безумия на благородном пути компании к AGI, где ставки так высоки. «Поскольку все мы, в общем, чувствуем, что приближаемся к тому, чтобы найти путь к этим мощным системам, – говорил он на общем собрании 15 мая, – уровень стресса и напряжения – внутри, снаружи, направленного на нас, исходящего от нас – будет и дальше, будет и дальше расти». Лучший способ с этим справиться – чтобы OpenAI удвоила усилия в

пиаре, укрепила отношения с правительствами и твёрдо держалась своих убеждений в своём видении. «Думаю, в целом у нас это неплохо получается, – добавил он, – но нас здесь снова будут проверять на прочность».

То, как OpenAI повела себя в истории с Суцкевером, станет лишь микрокосмом продолжающегося хаоса, порождаемого сохранением империи. Просьба Суцкевера о заверениях вызовет новую грызню в руководстве, на этот раз с немного другим составом действующих лиц, повторяя движимую самолюбием динамику, которая терзала компанию с самого начала. Александр Мадры, польский профессор из MIT, которого многие описывали как человека, рвущегося к власти, за сравнительно недолгое время работы успел сколотить внутри компании изрядную вотчину. Мадры не считал возвращение Суцкевера хорошей идеей. Суцкевер пользовался слишком большим восхищением и преданностью исследователей. Это могло бы ослабить влияние Мадры – и его доброго друга Пахоцкого. За несколько часов опасения Мадры успешно посеяли сомнения и раскололи руководство. Как всегда, Альтман устранился от решения в ту или иную сторону, чтобы не выглядело, будто он с кем-то не согласен.

Не прошло и суток после того, как руководители побывали у него дома, как Суцкеверу позвонил Брокман. Любое обсуждение возвращения Суцкевера, сказал ему Брокман, теперь полностью снято с повестки.

Глава 18

Формула империи

Однажды Альтман сказал со сцены, что лучшей книгой, прочитанной им в 2018 году, был «Разум Наполеона» (The Mind of Napoleon) – собрание высказываний Наполеона Бонапарта больше чем на триста страниц. Наполеон – французский военачальник, который совершил переворот и захватил власть во Франции, провозгласил себя императором французов и затем попытался завоевать Европу.

– Очевидно, глубоко порочный человек, но, чёрт возьми, впечатляет, – сказал Альтман.

– Какие у него были прозрения? – спросил Тайлер Коуэн, профессор экономики Университета Джорджа Мейсона, который вёл мероприятие.

– Невероятное понимание человеческой психологии, – ответил Альтман, которому оставались считаные недели до перехода в OpenAI на полную ставку и который ещё был президентом YC. – Это мы видим у многих наших лучших основателей.

Затем Альтман пересказал отрывок, который поразил его больше всего. Это были размышления Наполеона о девизе Французской революции (ставшем впоследствии национальным девизом страны) – «Свобода, равенство, братство» – и о том, как его можно переосмыслить и использовать для укрепления собственной власти. В конечном счёте именно под этим знаменем Наполеон сделал противоположное: ограничил свободу, отбросил братство – философию, основанную на единстве и солидарности, – и даровал равенство только французским мужчинам, но не женщинам, а заодно восстановил колониальное рабство, стремясь воссоздать французскую империю.

– И он говорил о том, как построить систему… в которой можно, в общем, контролировать людей, – размышлял Альтман. – Я подумал: «Ого. Как хорошо, что он не управляет Соединёнными Штатами, потому что этот чувак понимает что-то глубокое, чего не понимал я, и явно сумел использовать это для власти».

Через шесть лет после моего первоначального скептицизма насчёт альтруизма OpenAI я пришла к твёрдому убеждению, что миссия OpenAI – добиться того, чтобы AGI приносил пользу всему человечеству, – возможно, и начиналась как искренний порыв идеализма, но с тех пор превратилась в исключительно мощную формулу для сосредоточения ресурсов и построения властной структуры имперского типа. В этой формуле три ингредиента.

Во-первых, миссия сосредоточивает таланты, сплачивая людей вокруг грандиозной амбиции, – ровно так, как это сделал Джон Маккарти, придумав выражение «искусственный интеллект». «Самые успешные основатели вовсе не ставят себе целью создать компанию, – писал Альтман в своём блоге в 2013 году.

– Их миссия – создать нечто, гораздо более похожее на религию, и в какой-то момент выясняется, что проще всего это сделать, основав компанию». Во-вторых, миссия сосредоточивает капитал и другие ресурсы, устраняя препятствия, регулирование и несогласие. Инновации, современность, прогресс – чего бы мы не отдали, чтобы их достичь? Тем более перед лицом страшных, несогласованных с нами корпоративных и государственных конкурентов, которые якобы существуют. «Кто будет контролировать будущее ИИ?» – писал Альтман в колонке для The Washington Post в июле 2024 года, когда ещё не утихли толчки «Омникризиса». «Будет ли это будущее, в котором Соединённые Штаты и союзные страны продвигают глобальный ИИ, распространяющий блага технологии и открывающий к ней доступ, или авторитарное будущее, в котором страны или движения, не разделяющие наших ценностей, используют ИИ, чтобы укрепить и расширить свою власть?»

А самое главное – миссия остаётся настолько расплывчатой, что её можно толковать и перетолковывать – точно так же, как Наполеон поступил с девизом Французской революции, – направляя сосредоточение талантов, капитала и ресурсов туда, куда пожелает тот, кто их сосредоточивает. Что значит «польза»? Что такое AGI? «По-моему, это нелепый и бессмысленный термин, – сказал Альтман The New York Times всего за два дня до того, как совет его уволил. – Так что прошу прощения, что продолжаю им пользоваться».

В этом последнем ингредиенте дрейф OpenAI был поистине примечательным. В 2015 году её миссия означала быть некоммерческой организацией, «не стеснённой необходимостью приносить финансовую отдачу», и открывать исследования, как писала OpenAI в объявлении о своём создании. В 2016 году она означала, что «все должны получить пользу от плодов ИИ, после того как он будет создан, но совершенно нормально не делиться наукой», как писал Суцкевер Альтману, Брокману и Маску. В 2018 и 2019 годах она означала создание структуры с ограниченной прибылью, чтобы «мобилизовать значительные ресурсы», избегая при этом «конкурентной гонки, в которой нет времени на надлежащие меры безопасности», как писала OpenAI в своём уставе. В 2020 году она означала отгородить модель и построить «API как стратегию открытости и распределения пользы», как писал Альтман в ответ на мой первый портрет компании. В 2022 году она означала «итеративное развёртывание» и гонку, чтобы как можно быстрее развернуть ChatGPT. А в 2024 году Альтман написал в блоге после выпуска GPT-4o: «Ключевая часть нашей миссии – бесплатно (или по отличной цене) передавать в руки людей очень мощные инструменты ИИ».

Даже во время «Омникризиса» Альтман начинал заново переписывать свои определения.

На общем собрании 15 мая 2024 года, после ухода Суцкевера и Лейке, Альтман подчеркнул, что OpenAI скоро выйдет на новый уровень возможностей ИИ, который потребует от компании всё переосмыслить и перестроить. «Теперь

мы будем исходить из того, что мы, типа, вступаем в эпоху AGI», – сказал он. Во имя своей миссии OpenAI придётся ещё сильнее закрыться и удвоить усилия в глобальном лоббировании и публичных заявлениях. «Есть много вещей, к которым мы сейчас не готовы, – сказал он. – Стандарты безопасности, политические планы, которые у нас должны быть, а ещё созыв правительств, который нужно будет провести, чтобы к этому подготовиться; план, история, будущее, в котором люди смогут увидеть себя, когда речь пойдёт о социально-экономических последствиях этого».

В то же время во имя своей миссии OpenAI не будет замедляться коммерчески. «Это не значит, что мы не будем выпускать отличные продукты. Это не значит, что мы не будем и дальше делать отличные исследования. Это не значит, что мы не будем заключать всевозможные партнёрства и делать другие классные вещи». В конце того месяца в СМИ появятся сообщения, что OpenAI заключила крупную сделку о встраивании своих моделей в продукты Apple; обе компании подтвердят новость через две недели.

Более того, отметил Альтман на общем собрании, OpenAI и Microsoft пересматривают условия партнёрства, чтобы коммерциализация продолжалась. «Сразу после этого я лечу в Сиэтл, чтобы об этом поговорить. Это придётся развивать, – сказал он. – Когда мы изначально заключали сделку с Microsoft, мы придумали так называемую оговорку о достаточном AGI» – оговорку, определявшую момент, когда OpenAI перестанет делиться своей интеллектуальной собственностью с Microsoft. «Теперь мы все думаем иначе», – добавил он. Чёткой точки отсечения, когда OpenAI достигнет AGI, больше не будет. «Мы считаем, что это будет непрерывный процесс». Две компании и дальше будут сотрудничать и выпускать всё более совершенные технологии – по отличной цене.

Это была странная и непоследовательная стратегия, имевшая смысл лишь при одном прочтении: OpenAI будет делать всё, что нужно, толкуя и перетолковывая свою миссию соответственно, чтобы закрепить своё господство.

За кулисами Альтман готовил почву и для того, чтобы закрепить собственный контроль. Кризис в совете ясно показал, что структура OpenAI – некоммерческая организация, управляющая коммерческой, – сделала его смещение практически неизбежным. Такое устройство не только узаконило две противоборствующие силы компании – бумеров и думеров, – соперничающие за контроль над развитием ИИ, но и дало совету широкие полномочия уволить его, исходя из собственного, а не его толкования того, наилучшим ли образом он служит миссии.

Пока эта структура сохранялась, такой конфликт неизбежно повторился бы.

28 мая, меньше чем через неделю после того, как руководители пытались вернуть Суцкевера, один сотрудник снова написал в канале Slack #i-have-a-question. «Не знаю, стоит ли/как об этом спрашивать», – начинался вопрос: в дебрях последнего акционерного соглашения OpenAI обнаружились новые

подробности, которые, похоже, допускали роспуск некоммерческой организации. «Насколько прочна некоммерческая организация? – писал сотрудник. – План – и дальше оставаться под управлением некоммерческой организации?»

На следующий день The Information сообщило, что, похоже, такого плана нет. В списке главных приоритетов Альтмана на год была реструктуризация организации, чтобы она больше походила на обычную компанию. В следующем месяце издание подтвердило новые подробности. Альтман рассматривал несколько сценариев: один – превращение OpenAI в обычную коммерческую компанию; другой – превращение её в коммерческую общественно-полезную корпорацию, как Anthropic и xAI. Оба сценария сохраняли некоммерческую организацию как отдельное юридическое лицо, но лишали её совет контроля над бизнесом компании. В рамках новой структуры инвесторы к тому же давили на Альтмана, чтобы он взял долю в компании и его стимулы теснее совпали с их собственными.

* * *

В следующие месяцы, пока OpenAI готовила планы перехода, две силы, действовавшие во время «Омникризиса», продолжали работать. Думеры усилили публичное давление на компанию. 4 июня The New York Times опубликовала портрет Дэниела Кокотайло и рассказ о его новой кампании, призывавшей передовые ИИ-компании взять на себя обязательства большей прозрачности и защиты информаторов, сохраняющей право сотрудников предупреждать общество о рисках, которые они видят в компании. В открытом письме с изложением требований к Кокотайло присоединились ещё двенадцать подписантов, десять из которых в разное время принадлежали к клану безопасности OpenAI. Через месяц The Washington Post сообщила, что группа из OpenAI также подала жалобу в SEC, утверждая, что компания нарушила федеральные гарантии защиты информаторов чрезмерно широкими соглашениями об уходе. Позже в том же месяце пятеро американских сенаторов направят Альтману письмо с вопросами, требуя большей ясности по различным обвинениям Лейке, Кокотайло и других, а также по материалам Пайпер о пренебрежении OpenAI безопасностью ИИ и подавлении критики со стороны сотрудников.

Тем временем хаос в руководстве и раздражение Альтманом не утихали. Вкупе с продолжающейся вознёй бумеров и думеров это вело к постоянным изменениям в структуре подчинённости компании. Брокман перестал подчиняться Мурати и стал подчиняться напрямую Альтману; Александра Мадры вскоре после письма сенаторов перевели с поста руководителя команды готовности на менее значимую роль в исследованиях. OpenAI также приглашала более опытных руководителей, в том числе Сару Фрайар, бывшего CEO соседской соцсети Nextdoor, на должность финансового директора и Кевина Вейла,

бывшего руководителя продуктов в Facebook, Instagram и Twitter, на должность директора по продукту.

Довольно скоро компания лишится целой череды своих самых давних руководителей. Первым ушёл Джон Шульман, объявивший об уходе 5 августа 2024 года: он написал о желании «глубже сосредоточиться на согласовании ИИ» и о решении заниматься этим в Anthropic. В тот же день Брокман объявил, что берёт творческий отпуск до конца года, представив его – отчасти ставший итогом накопившихся у сотрудников претензий к его руководству – как необходимую передышку после девяти лет бега в компании.

В следующем месяце, 25 сентября, о внезапном уходе объявила и второй руководитель, высказывавший совету серьёзные опасения насчёт Альтмана, – Мира Мурати. «Шесть с половиной лет с командой OpenAI были для меня исключительной честью», – написала она, поблагодарив Альтмана и Брокмана и отметив, как дорога ей компания и как она и дальше будет за неё болеть. «Я ухожу, потому что хочу найти время и пространство для собственных поисков». Через несколько часов о своём уходе заявили ещё двое ключевых руководителей: директор по исследованиям Боб Макгрю и вице-президент Баррет Зоф, руководивший командой дообучения вместе с Шульманом. Все трое подчёркивали перед коллегами и публикой, что сейчас подходящий момент, чтобы уйти из OpenAI на высокой ноте, после того как компания достигла очередной важной вехи: в середине сентября вышла её новейшая модель Strawberry – под новой схемой названий компании, o1, – выросшая из одного из последних вкладов Суцкевера в компанию.

* * *

На самом деле момент был ужасный. После «Омникризиса» конкуренция, с которой сталкивалась OpenAI, только ускорилась. Маск с пугающей скоростью наращивал вычислительные мощности для xAI. Новейшая версия Claude от Anthropic уводила клиентов у ChatGPT. Суцкевер официально основал свою новую конкурирующую компанию Safe Superintelligence («Безопасный сверхинтеллект») и только что объявил о стартовом финансировании в миллиард долларов.

В то же время после больше чем года работы OpenAI всё ещё не могла добиться от Orion той производительности, которая оправдала бы её выпуск. Компания начинала смотреть в лицо неприятной перспективе: её испытанная формула масштабирования, похоже, больше не работала; чтобы и дальше развивать свои ИИ-системы, ей, вероятно, нужны были принципиально новые исследовательские идеи. Сказать это вообще было гораздо проще, чем сделать, а тем более после того, как OpenAI годами выстраивала найм и организацию команд вокруг эксплуатации существующих исследований, а не изучения неизведанной науки.

За два дня до объявления Мурати Альтман опубликовал свою самую напыщенную на тот момент запись в блоге – на фоне очередного привлечения средств OpenAI. Она называлась «Эпоха интеллекта» (The Intelligence Age) и была полна захлёбывающихся обещаний «невообразимого» грядущего процветания. «Как мы оказались на пороге следующего скачка в процветании? – писал Альтман. – Если в пятнадцати словах: глубокое обучение сработало, предсказуемо улучшалось с масштабом, и мы вкладывали в него всё больше ресурсов».

На общем собрании Мурати объяснила сотрудникам, почему её объявление было таким внезапным. «Я хотела, чтобы новость о моём уходе вы все узнали сначала от меня, а не от своих руководителей, не от кого-то ещё и уж тем более не из прессы». При таком пристальном внимании к OpenAI она не видела другого способа сделать это, не преподнеся всем сюрприз. Узнав о решении Мурати, Макгрю тоже решил, что пора уходить, сказал он. «Я понял, что на самом деле добился большей части главного, чего хотел, когда пришёл сюда».

Как и в случае с уходом Суцкевера, OpenAI постаралась сгладить уход этой тройки. «Мира, Боб и Баррет приняли эти решения независимо друг от друга и по-доброму, – написал Альтман во внутренней записке, которую затем выложил в Twitter, – но время решения Миры было таким, что имело смысл сделать всё это сейчас разом, чтобы мы могли вместе плавно передать дела следующему поколению руководителей».

Вместо Макгрю новым старшим вице-президентом по исследованиям, который будет руководить подразделением вместе с Пахоцким, станет Марк Чен – один из руководителей исследований, выступавший рядом с Мурати и Зофом на демонстрации 4o. Ещё один давний исследователь OpenAI, Джошуа Ачиам, займёт новую должность руководителя по согласованию с миссией. Лиам Федус, один из выходцев из Google, пришедших вместе с Зофом, вскоре возглавит работу над дообучением. Пока, сказал Альтман, OpenAI не будет искать нового технического директора на смену Мурати.

При всех переменах в руководстве и запланированных структурных изменениях компания обнаруживала свою единственную константу: это была и по-прежнему оставалась империя ИИ Сэма Альтмана.

* * *

В начале октября 2024 года OpenAI закрыла новейший раунд финансирования на 6,6 миллиарда долларов – крупнейший венчурный раунд в истории, оценивший компанию в 157 миллиардов. В нём была загвоздка: инвесторы могли потребовать деньги обратно, если компания в течение двух лет не превратится в коммерческую.

До конца 2024 года ключевые сотрудники продолжали утекать из OpenAI: Люк Мец, третий выходец из Google, пришедший вместе с Зофом и Федусом; Майлз Брандедж, руководитель исследований политики; Лилиан Венг,

унаследовавшая работу Дэйва Уиллнера по доверию и безопасности и недавно повышенная до вице-президента по исследованиям и возглавившая направление безопасности; и Алек Рэдфорд – тот самый исследователь, который когда-то направил OpenAI по пути моделей GPT.

Anthropic запустила рекламную кампанию Claude с дерзкой надписью на билбордах в Сан-Франциско: «Тот, что без всей этой драмы». На фоне утечки кадров Брокман досрочно вернулся из творческого отпуска. Энни направила Сэму официальное письмо с уведомлением о намерении подать иск. Маск, теперь союзник вновь избранного президента Дональда Трампа, усилил свой иск, возражая против ожидаемого превращения OpenAI в коммерческую компанию, и обнародовал новые ранние письма времён основания OpenAI, в том числе те, где рассказывалось, как Альтман за спиной плохо отзывался о Брокмане и Суцкевере («Признал, что за этот процесс сильно потерял доверие к Грегу и Илье. Считает, что их позиция была непоследовательной и порой казалась ребяческой»), и где проявлялась его неприязнь к прозрачности («Считал, что это отвлекает команду»). В неожиданном союзе его поддержал Цукерберг, давно враждовавший с Маском: в письме генеральному прокурору Калифорнии Meta тоже призвала заблокировать преобразование OpenAI. «Поведение OpenAI может иметь для Кремниевой долины сейсмические последствия», – писала Meta. Преобразование могло создать опасный прецедент, после которого гораздо больше стартапов стали бы объявлять себя некоммерческими, получая для себя и своих инвесторов налоговые вычеты от государства, пока не станут прибыльными.

В конце года, затерявшись в праздничном потоке новостей, OpenAI официально объявила о планах новой структуры. Она превратится в коммерческую общественно-полезную корпорацию, а некоммерческая организация сохранится как отдельное юридическое лицо с долей в коммерческой. Такая структура, доказывалось в объявлении, – лучший способ обеспечить и коммерческую, и некоммерческую части нужными ресурсами для выполнения их задач, служа при этом миссии. «Нам снова нужно привлечь больше капитала, чем мы представляли, – говорилось там. – Мир переходит к строительству новой инфраструктуры энергетики, землепользования, чипов, центров обработки данных, данных, моделей ИИ и ИИ-систем для экономики XXI века. Мы стремимся эволюционировать, чтобы сделать следующий шаг в нашей миссии – помочь построить экономику AGI и добиться того, чтобы она приносила пользу человечеству».

В начале нового года Альтман снова красовался на публике. «Теперь мы уверены, что знаем, как создать AGI в том смысле, в каком его традиционно понимали, – писал он в новой записи в блоге 6 января 2025 года. – Мы начинаем устремлять свой взгляд дальше – к сверхинтеллекту в подлинном смысле слова».

Как рушится империя

В 2021 году я наткнулась на историю, непохожую ни на одну из тех, о которых писала прежде: историю общины коренного народа Новой Зеландии, которая с помощью ИИ возрождала те рео маори – язык народа маори.

Как и многие коренные народы по всему миру, маори поколениями терпели чудовищное обращение при колониальном правлении: в 1867 году по Закону о туземных школах, сделавшему английский единственным языком, которому можно учить в школах, детей маори стыдили и даже били за то, что они говорили на родном языке. Когда в начале 1900-х страну захлестнула стремительная урбанизация, общины маори распались и рассеялись, что ослабило очаги сохранения культуры и языка. Доля говорящих на те рео среди маори рухнула с 90 до 12 процентов. К тому времени, когда Новая Зеландия – или Аотеароа, как изначально назвали свою землю маори, – 120 лет спустя отказалась от этой политики, учителей те рео, способных возродить умирающий язык, почти не осталось. Как и очень многие языки до него, те рео едва не исчез с лица земли.

Трудно в полной мере передать трагедию утраты языка. По тем же причинам, по которым исследователи ИИ изначально потянулись к языку, создавая свои технологии, утрата языка – это намного больше, чем утрата средства общения. В каждом языке закодированы богатая история, культура, знания; это коллективное творение миллионов людей на протяжении времени, которые искали звуки и письменные формы, чтобы запечатлеть тончайшие наблюдения о мироздании, о жизни, о человеческом опыте; чтобы делиться друг с другом ошеломляющей красотой и мучительными неудачами; чтобы учить ребёнка и учиться у старших; чтобы выражать любовь.

Утрата языка – мировая трагедия, но и личная. Когда тебя отрывают от твоего наследия и заставляют под угрозой побоев хранить чужое, это самым грубым образом устанавливает чёткую иерархию: чья история, чья культура, чьё знание заслуживают того, чтобы их передавали дальше, а чьи настолько ничтожны, что заслуживают стирания.

Большие языковые модели ускоряют утрату языков. Даже для моделей на несколько поколений старше, вроде GPT-2, в мире есть лишь немного языков, на которых говорит достаточно людей и которые задокументированы в сети в достаточном масштабе, чтобы удовлетворить потребность этих моделей в данных. Из более чем семи тысяч языков, существующих сегодня, почти половина, по данным ЮНЕСКО, находится под угрозой исчезновения; примерно у трети есть хоть какое-то присутствие в интернете; меньше 2 процентов поддерживаются Google Переводчиком; а по собственным тестам OpenAI, GPT-4 поддерживает с точностью выше 80 процентов лишь пятнадцать языков, или 0,2

процента. По мере того как эти модели становятся цифровой инфраструктурой, доступность интернета для разных языковых сообществ – и доступность экономических возможностей, которые он даёт, – будет и дальше сокращаться, подталкивая всё больше таких сообществ к тому, чтобы учить доминирующий язык вроде английского и говорить на нём в ущерб собственному.

Именно перед лицом этой надвигающейся экзистенциальной угрозы – совсем иного понимания того, что такое «экзистенциальное», – пара коренных жителей, Питер-Лукас Джонс и Кеони Махелона, впервые обратилась к ИИ как к возможному инструменту, который поможет новому поколению носителей вернуть те рео былую жизненную силу. Джонс – маори, Махелона – коренной гаваец; они партнёры и в работе, и в жизни. Двое мужчин встретились и полюбили друг друга, рассказывает Махелона, после того как ему во сне пришло видение: если он переедет в Новую Зеландию, то встретит парня-маори, с которым разделит жизнь.

В 2012 году они переехали из Веллингтона обратно в город, где родился Джонс, – Кайтаиа на севере Аотеароа. Джонс стал CEO Te Hiku Media, общественной радиостанции, вещающей на те рео, части более широкой сети СМИ и других организаций, занимающихся возрождением языка. На новом посту Джонс увидел возможность. За двадцать с лишним лет вещания Te Hiku накопила богатый архив записей людей, говорящих на те рео, в том числе запись его собственной бабушки Раихи Моэроа, родившейся в конце XIX века, чей выговор ещё не был искажён влиянием английского языка колонизаторов. Джонс к тому же мечтал записать ещё много интервью со старейшинами маори, чтобы задокументировать их устные предания и исконный те рео, пока они не ушли. Эти записи, как считал Джонс, могли стать драгоценным ресурсом для изучения языка, порталом в прошлое, через который новые поколения носителей те рео услышат изначальное звучание своего языка и прикоснутся к мудрости предков.

Трудность была в том, чтобы расшифровать записи и помочь учащимся следить за речью, ведь людей, свободно владеющих те рео, было очень мало. Поэтому в 2016 году, как раз когда начинала свою работу OpenAI, Джонс обратился к Махелоне, который переделывал сайт Te Hiku, с просьбой найти решение. Махелона, человек разносторонний, изучал машиностроение в Колледже Олина, управление бизнесом в первой магистратуре, а физику и вычислительные нанотехнологии – во второй, будучи стипендиатом программы Фулбрайта в Новой Зеландии. Он быстро придумал использовать ИИ: с тщательно обученной моделью распознавания речи на те рео Te Hiku сможет расшифровать свой аудиоархив силами всего нескольких носителей.

Здесь история Te Hiku полностью расходится с моделью разработки ИИ, которой следуют OpenAI и Кремниевая долина. Хорошо знакомые с опустошительными последствиями колониального лишения, Джонс и Махелона были полны решимости браться за проект, только если смогут на каждом этапе

разработки гарантировать три вещи: согласие, взаимность и суверенитет народа маори. Это означало, что ещё до начала проекта они получат разрешение общины маори и её старейшин, спросив, хотят ли те вообще этого начинания; для сбора обучающих данных будут привлекать только тех, кто полностью понимает, для чего будут использованы данные, и готов участвовать; чтобы модель принесла как можно больше пользы, будут прислушиваться к общине, узнавая, какие ресурсы для изучения языка будут полезнее всего; а когда появятся средства, купят собственные GPU Nvidia и серверы, которые будут стоять у них на месте, чтобы обучать модели, не завися от облака какого-либо техногиганта.

Самое главное – Te Hiku создаст порядок, при котором собранные данные будут и дальше служить ресурсом ради будущей пользы, но никогда не будут присвоены для проектов, на которые община не давала согласия, которые могут её эксплуатировать и вредить ей или иным образом нарушать её права. Согласно принципу маори кайтиакитанга, то есть попечительства, данные останутся под опекой Te Hiku, а не будут свободно выложены в интернет; затем Te Hiku будет выдавать на них лицензии только организациям, уважающим ценности маори и намеренным использовать их для проектов, которые община одобрила и сочла полезными.

«Данные – последний рубеж колонизации», – сказал мне Махелона: империи прошлого отнимали землю у коренных общин, а потом заставляли их покупать её обратно на новых, ограничительных условиях, если они хотели вновь ею владеть. «ИИ – это просто очередной захват земли. Большие технологические компании любят собирать ваши данные более или менее бесплатно – чтобы строить что им вздумается, какой бы ни была их конечная цель, – а потом разворачивают это и продают вам обратно как услугу».

От начала и до конца Джонс и Махелона осуществили проект без компромиссов. В какой-то момент они начали просветительскую кампанию, чтобы рассказать большему числу маори об ИИ, и общественный конкурс, чтобы собрать пожертвования данных и разметку силами общины. За десять дней Te Hiku собрала триста десять часов качественно расшифрованных записей из примерно двухсот тысяч фрагментов, записанных примерно двумя с половиной тысячами человек. Такой уровень вовлечённости был неслыханным для многих исследователей ИИ – и свидетельствовал о том, какое доверие и воодушевление вызвал подход Te Hiku в общине. Люди более чем охотно отдавали свои данные, как только понимали проект и давали на него согласие, полностью доверяя Te Hiku и дальше надлежащим образом заботиться об этих данных.

Этот объём данных бледнел по сравнению с шестьюстами восемьюдесятью тысячами часов аудио, которые OpenAI выдрала со всего интернета для обучения своего инструмента распознавания речи Whisper. Но ещё один урок истории Te Hiku в том, что трёхсот десяти часов всё же хватило, чтобы создать самую первую модель распознавания речи на те рео с точностью 86 процентов. Если OpenAI

стремится создавать единые гигантские модели ИИ, которые будут делать что угодно, и такая цель неизбежно требует выгребать как можно больше данных, то Te Hiku просто хотела создать маленькую специализированную модель, превосходно делающую одно дело. Кроме того, Te Hiku выиграла от трансграничного сообщества ИИ с открытым кодом: в качестве отправной точки она взяла бесплатную модель распознавания речи DeepSpeech от фонда Mozilla, которая сама была плодом иного видения развития ИИ. Как и Te Hiku, Mozilla обучала модель только на данных, переданных с полного согласия, а построила её на архитектуре нейросети, разработанной исследовательской лабораторией китайской компании Baidu в районе залива Сан-Франциско. В общей сложности Te Hiku использовала всего два GPU.

* * *

Я писала о работе Te Hiku до того, как ChatGPT стремительно утвердил господствующую парадигму развития ИИ, почти выбросив за борт согласие, взаимность и суверенитет. Но в последующие годы я стала видеть в радикальном подходе Te Hiku нечто ещё более актуальное и жизненно важное. Критика видения OpenAI и Кремниевой долины в целом, которую я излагаю в этой книге, ни в коем случае не означает отказа от ИИ как такового. Я отвергаю опасную мысль о том, что широкая польза от ИИ может быть получена только – и вообще когда-либо будет получена – из такого видения технологии, которое требует полной капитуляции нашей частной жизни, нашей свободы действий и нашей ценности, включая ценность нашего труда и искусства, ради проекта централизации, в конечном счёте имперского.

Te Hiku показывает нам другой путь. Она показывает, что ИИ и его разработка могли бы быть прямо противоположными. Модели могут быть маленькими и предназначенными для конкретных задач, их обучающие данные – ограниченными и известными, что устраняет стимулы к повсеместной эксплуатации труда, наносящей вред психике, и к всепоглощающему экстрактивизму, без которого не построить и не запустить огромные суперкомпьютеры. Создание ИИ может быть делом общины, основанным на согласии и уважении к местному контексту и истории; его применение может поддерживать и укреплять маргинализированные общины; управление им может быть инклюзивным и демократическим.

Te Hiku – не единственная организация, ищущая новые пути развития ИИ. Работая над этой книгой, я снова и снова черпала вдохновение у множества организаций и движений по всему миру, которые расцвели, чтобы противостоять империям ИИ, отстоять своё право на самоопределение и представить себе новый путь вперёд.

После того как Тимнит Гебру выгнали из Google, в декабре 2021 года она основала некоммерческую организацию, чтобы продолжить свои исследования. Она назвала её DAIR – Distributed AI Research Institute, Институт

распределённых исследований ИИ: «распределённых» – наперекор централизации. «Это было первое слово, пришедшее мне в голову», – говорит Гебру. Она задумала собрать команду исследователей со всего мира, которые оставались бы в своих общинах, привнося в работу института богатый опыт и взгляды своего местного контекста и в то же время используя эту работу на благо этих общин. «Технологии из Кремниевой долины влияют на весь мир, но у всего мира нет возможности влиять на технологии», – говорит она.

Социолог Алекс Ханна, одна из соавторов статьи о «стохастических попугаях» из Google, первой присоединилась к Гебру в DAIR как директор по исследованиям. Первым делом Ханна написала исследовательскую философию, подробнее раскрывающую дух работы организации. Для этого Гебру и Ханна наняли третьего человека в DAIR – Милагрос Мисели, тоже социолога и специалиста по информатике, изучавшую эксплуататорские трудовые практики отрасли ИИ. Вместе они сформулировали свою философию: «Наши исследования призваны приносить пользу сообществам, которым ИИ обычно не служит, и прокладывать пути к тому, чтобы вместе отвергать, подвергать сомнению и перекраивать ИИ-системы».

Они определили семь опор для воплощения этой философии, в том числе: ставить в центр сообщества, на которые влияет ИИ, но которые обычно не представлены в его исследованиях, и выстраивать с ними содержательные отношения, относясь к ним как к настоящим партнёрам в производстве знаний; справедливо оплачивать любые формы труда, задействованные в создании исследований и технологий; ставить под вопрос системы, лежащие в основе развития ИИ и маргинализирующие тех, кого исторически всегда маргинализировали; и вместе с этими сообществами придумывать альтернативы, которые шаг за шагом могли бы переделать мир в тот, в котором они хотели бы жить.

Отсюда Мисели перешла к новому исследовательскому проекту, чтобы воплотить эту философию на практике. Она создала проект Data Workers' Inquiry («Исследование работников данных») и пригласила работников по разметке данных со всего мира самим сформулировать исследовательские вопросы об индустрии разметки данных и о том, как сделать её лучше. Где бы они ни жили, она платила им стандартную зарплату исследователя в Германии, где работает сама, – двадцать пять евро в час, – чтобы отразить ценность их работы.

«Вокруг работы с данными всегда эта ложная логика: какой минимум мы можем платить этим людям? Это идёт от колониалистской логики: выбираешь место, где можно сделать больше всего с самым дешёвым бюджетом и где можно по-настоящему обворовывать людей, красть ресурсы задёшево, – говорит Мисели. – Вопрос в том, почему эти компании платят два доллара в час, если работа приносит им миллиарды или триллионы выручки? Почему бы нам не посмотреть, сколько эти компании могут заплатить, вместо того чтобы смотреть, на сколько меньше могут согласиться эти работники?»

Среди пятнадцати работников, участвовавших в первом раунде исследования, были Оскарина Вероника Фуэнтес Анайя из Венесуэлы и Мофат Окиньи из Кении. Для своего проекта Фуэнтес вместе с художником-аниматором создала видео о пережитом и вместе с другими работниками по разметке данных рассказала об их общих трудностях: нехватке заданий на платформе, непредсказуемом и неподвластном им рабочем времени и ничтожной оплате. Сейчас Фуэнтес работает одновременно на пяти платформах разметки данных, чтобы получать чуть больше минимальной зарплаты в Колумбии – около 335 долларов в месяц. Каждое задание приносит в среднем от одного до пяти центов; она по-прежнему заставляет себя просыпаться, когда задания приходят посреди ночи. «Мы призраки для общества, и, осмелюсь сказать, мы дешёвая, одноразовая рабочая сила для компаний, которым годами служим без гарантий и защиты», – написала она для своего проекта. После Data Workers' Inquiry она продолжает рассказывать о своём опыте в онлайн-выступлениях и вебинарах, надеясь надавить на компании и политиков, чтобы те добились лучшего обращения с работниками.

На другом континенте Окиньи тоже занимается организацией. В мае 2023 года, через год с небольшим после внезапного прекращения контракта OpenAI с Sama, он стал одним из организаторов кенийского Африканского профсоюза модераторов контента, который борется за более высокую оплату и лучшее обращение с африканскими работниками, выполняющими самую тяжёлую работу интернета. Полгода спустя, публично рассказав о своём опыте с OpenAI в моей статье для The Wall Street Journal, он вместе с бывшим коллегой по Sama Ричардом Матенге основал и собственную некоммерческую организацию – Techworker Community Africa (TCA).

В августе 2024 года, когда мы снова поговорили, Окиньи мечтал превратить TCA в ресурс и для африканского сообщества работников, занятых данными для ИИ, и для международных групп и политиков, желающих их поддержать. Он организовывал онлайн-конференции и школьные собрания, чтобы рассказать работникам и учащимся, особенно женщинам, об их трудовых правах и праве на защиту данных, а также о том, как устроена отрасль ИИ. Он искал финансирование, чтобы открыть учебный центр для повышения квалификации. Он встречался с американскими представителями, приезжавшими в Найроби, чтобы лучше понять опыт работников, обслуживающих американские технологические компании. Он отвечал на разные запросы международных организаций, в том числе Equidem, правозащитной организации, поддерживающей трудящихся глобального Юга, и проекта Fairwork Оксфордского института интернета.

Для Data Workers' Inquiry он брал интервью у работников Remotasks в Кении, которых Scale без разговоров отрезала от платформы, так что последние заработки, которые они не успели вывести, просто исчезли. Часть пожертвований, собранных TCA, он потратил на то, чтобы поддержать их в этом

финансовом кошмаре. «Когда пыль на этой главе оседает, одно остаётся ясным: человеческая цена бесконтрольной власти и безудержной жадности, – писал он. – Голоса этих работников несут надежду на более светлое и справедливое будущее… это призыв к действию, чтобы с работниками повсюду обращались с тем достоинством и уважением, которых они заслуживают».

В его собственной жизни достоинство и уважение, которые принесла ему правозащитная работа, вдохнули в Окиньи новую надежду и сильно улучшили его психическое состояние, говорит он. Незадолго до нашего разговора он узнал, что войдёт в ежегодный список ста самых влиятельных людей в ИИ по версии журнала Time. «Я чувствую, что мою работу ценят», – говорит он. Это не значит, что работа обходится без трудностей. В марте 2024 года он ушёл со штатной должности в аутсорсинговой компании, где работал после Sama. По его словам, руководству компании не нравилась его организаторская деятельность. «Они думали, что я буду подбивать сотрудников становиться активистами». Та же компания перевела часть проектов в Гану, когда профсоюз и TCA стали громче. Он слышал, что кенийские чиновники жалуются, будто волнения работников отпугивают инвестиции и оставляют без работы ещё больше кенийцев.

Глобальный характер отрасли сделал Окиньи ещё решительнее в стремлении привлечь к африканским работникам, занятым данными, международное внимание. Даже если бы кенийское правительство его поддерживало, одни кенийские законы мало что могли бы сделать, чтобы ограничить поведение ИИ-компаний. Большинство этих компаний – из США, а именно из Сан-Франциско, говорит он. Нужны согласованные международные усилия, чтобы призвать их к ответу.

В Уругвае Даниэль Пенья приходит к тем же выводам. Цепочка поставок отрасли ИИ запутанна и обширна. «Энергию берут здесь, данные идут туда, минералы добывают где-то ещё, работников привозят откуда-то ещё», – говорит он. Перед лицом этих расползающихся последствий и стоящих за ними огромных, могущественных компаний каждая община, ведущая свою местную борьбу, может чувствовать себя одинокой и бессильной, особенно когда ей связывают руки собственные правительства, которым «нужны компании, чтобы поддерживать видимость стабильной экономики». Вскоре после нашей встречи он узнал, что правительство проигнорировало его петицию, подписанную больше чем четырьмястами людьми, с требованием более широко изучить социальное и экологическое воздействие центра обработки данных Google в их стране. Вместо этого министерство окружающей среды тихо одобрило проект и, по его словам, сообщило о решении лишь после того, как истекло тридцатидневное окно для публичного оспаривания. Пенья не сдаётся. Он общается с MOSACAT в Чили и связывается со всеми другими общинами, какие может найти, которые тоже сопротивляются эксплуатации и экстрактивизму технологической отрасли. Объединяя движения поверх границ, делясь друг с другом информацией и стратегиями сопротивления, он видит путь к созданию

более сильной коллективной власти, которая сможет давить на отрасль и подталкивать её к лучшему. «Нам нужно бороться на глобальном уровне», – говорит он.

* * *

Если миссия OpenAI – формула построения империи, то какова формула её роспуска? Пока я пишу эту книгу, невозможно знать в подробностях, как будут дальше развиваться эта компания и стремительная отрасль ИИ. Возможно, кто-то из многочисленных конкурентов OpenAI отнимет у неё лидерство; весьма вероятно, тактики этих империй ИИ будут меняться – в том, как они разрабатывают модели, эксплуатируют труд и расширяют вычислительную инфраструктуру. Но как бы ни повернулись события через два года или через десять лет, есть вещи, которые нам нужно делать и которые не должны меняться.

В своём выступлении на NeurIPS в 2019 году, на семинаре Queer in AI, Рия Каллури, исследовательница ИИ из Стэнфорда, предложила меткую альтернативу вопросу о том, как добиться, чтобы ИИ творил «добро». Добро, польза для человечества – оценка этих понятий всегда будет зависеть от того, кто смотрит. Вместо этого нам следует спрашивать, как ИИ перераспределяет власть: сосредоточивает он её или рассредоточивает? Если говорить в рамках этой книги: продолжает ли он укреплять империю или начинает возвращать нас к демократии?

Выступая перед технической аудиторией, Каллури сосредоточилась на фундаментальных исследованиях ИИ – на том, как учёные могли бы с помощью этого вопроса решать, какие формы ИИ создавать и в каких направлениях развивать эту область. Но её вопрос столь же важен для всех остальных сторон ИИ. Как нам разрабатывать приложения ИИ; как ими пользоваться; и в конечном счёте, как я спрашивала в начале этой книги, как нам управлять этой технологией, чтобы вернуть власть людям?

Работа Te Hiku, DAIR, Окиньи, Фуэнтес и Пеньи – примеры той работы, которую можно и нужно делать, чтобы перераспределить власть. Но вопрос управления – это вопрос о том, как создать условия, при которых такой работы станет больше и она сможет процветать.

В своём выступлении Каллури говорила о разных осях власти. Эта книга касается трёх: знаний, ресурсов и влияния. Сейчас OpenAI и конкурирующие с ней империи контролируют все три: сосредоточивая таланты, подтачивая открытую науку и закрывая свои модели от общественного контроля, они контролируют производство знаний; накапливая финансирование, данные, труд, вычисления, энергию и землю, они контролируют и истощают ресурсы других людей; создавая и укрепляя идеологии и устраивая невероятно популярные демонстрации, захватывающие воображение всего мира, они обладают огромным влиянием. Каждая ось укрепляет другую. Контроль над

производством знаний подпитывает влияние; растущее влияние накапливает ресурсы; накопленные ресурсы обеспечивают производство знаний.

Значит, формула роспуска империи требует перераспределения власти по каждой оси. Предложения и рекомендации, которые я здесь излагаю, – лишь примеры, отнюдь не исчерпывающие. Во-первых, чтобы перераспределить знания, нужно больше финансирования для их производства вне империи. Это означает поддерживать исследователей, способных проводить независимую оценку корпоративных моделей, чтобы в понимании их возможностей мы не зависели только от самих компаний. Это означает поддерживать организации вроде DAIR, которые могут вести исследования в совершенно новых направлениях – например, создавать новые формы ИИ помимо больших языковых моделей, более экономные в отношении данных и энергии. Это означает поддерживать организации вроде Te Hiku, которые могут создавать предназначенные для конкретных задач, ведомые общинами приложения ИИ, укрепляющие маргинализированные общины. К независимому производству знаний относится и работа журналистов и групп гражданского общества, которые могут погружаться в жизнь общин и работать на месте, помогая нам понять, а не просто гадать, какова многослойная реальность воздействия этих технологий.

Для перераспределения знаний понадобятся и законы, обязывающие компании раскрывать ключевые сведения об обучающих данных и технических характеристиках своих моделей и суперкомпьютеров. Только тогда независимые оценщики корпоративных моделей смогут делать свою работу. Исследовательница из Калифорнийского университета в Беркли Дебора Раджи, которая и после форумов Шумера продолжает работать с политиками разных стран, говорит, что это ещё и необходимый минимум для того, чтобы гарантировать безопасность корпоративных систем в реальном мире. То есть не от гипотетического вреда со стороны вышедшего из-под контроля ИИ, о котором говорит думерство, а от уже существующего реального вреда – от дискриминации до дезинформации и автоматизации рабочих мест, – с которым потребители и общины могут столкнуться уже сейчас, если широко развёрнутые модели не проверены как следует. «У нас есть CFPB, который следит за потребительскими финансовыми продуктами. У нас есть FDA, которое следит за медицинскими изделиями. Но почему-то, когда дело касается продуктов ИИ, никакого надзора просто нет», – говорит Раджи. На самом деле модели ИИ должны требовать большей прозрачности, чем обычный продукт. «Это системы, определяемые данными. Они не детерминированы. Поэтому нам нужно знать о них больше, чтобы понимать, что они делают».

Такая прозрачность к тому же крайне важна для измерения воздействия ИИ на окружающую среду. И здесь другие продукты уже проходят оценки, которых не проходят продукты ИИ. «Если вы пользуетесь машиной, если покупаете бытовой прибор, у вас есть рейтинг Energy Star, – говорит Саша

Луччони из Hugging Face. – А ИИ так глубоко встроен в наше общество, так широко используется в продуктах, а у нас нет никакой информации об экологичности этих систем».

С такой прозрачностью мы начали бы перераспределять власть и по второй оси – ресурсам. Пряча составляющие своих моделей как интеллектуальную собственность, империи ИИ до сих пор безнаказанно присваивали чужую интеллектуальную собственность без упоминания, согласия и вознаграждения. Возможность заглянуть в обучающие данные компаний сильно затруднила бы такое хищническое и эксплуататорское поведение. Как и возможность заглянуть в цепочки поставок компаний, в том числе узнать, где они нанимают работников и где договариваются об аренде новых земель под строительство электростанций и центров обработки данных, что так часто происходит через подставные компании.

Перераспределение ресурсов требует также более сильной защиты трудящихся повсюду – не только работников по разметке данных, напрямую нанятых отраслью, но и всех, чей труд рискует быть присвоенным как обучающие данные или чья работа может быть автоматизирована. Голливудские забастовки, которые позволили сценаристам и актёрам добиться защиты от определённых способов использования ИИ, показали, какую важнейшую роль будут играть профсоюзы в сопротивлении обесцениванию человеческого труда, снижению зарплат и перетеканию денег от работников в руки ИИ-компаний.

Наконец, чтобы перераспределить власть по третьей оси, влиянию, нам нужно широкое просвещение. Противоядие от мистики и миражей ажиотажа вокруг ИИ – учить людей тому, как работает ИИ, каковы его сильные стороны и недостатки, какие системы определяют его развитие, каковы мировоззрение и слабости людей и компаний, создающих эти технологии. Как сказал в 1960-х годах Джозеф Вейценбаум, профессор MIT и создатель чат-бота ELIZA, «стоит разоблачить конкретную программу, стоит объяснить её внутреннее устройство языком достаточно простым, чтобы всё стало понятно, – и её волшебство рассыпается». Я надеюсь, что эта книга станет хотя бы одним вкладом в такое понимание. Она опирается на труд множества учёных, журналистов, активистов и просветителей, работавших до меня и посвятивших себя просвещению общества. Пусть она станет новой почвой, на которой вслед за ними поднимутся и продолжат строить многие другие.