Радио Свобода

ПОПУЛЯРНАЯ АМЕРИКА

Отличить Никсона от Кеннеди. AI – от векторов до политики

Валентин Барышников

менее минуты назад

"Популярная Америка" с Максимом Рагинским

Администрация США ввела экспортные ограничения на новейшие модели искусственного интеллекта компании Anthropic, которые якобы способны выявлять уязвимость компьютерных систем. Столкновение AI и политики на глазах становится одной из главных проблем нашего времени.

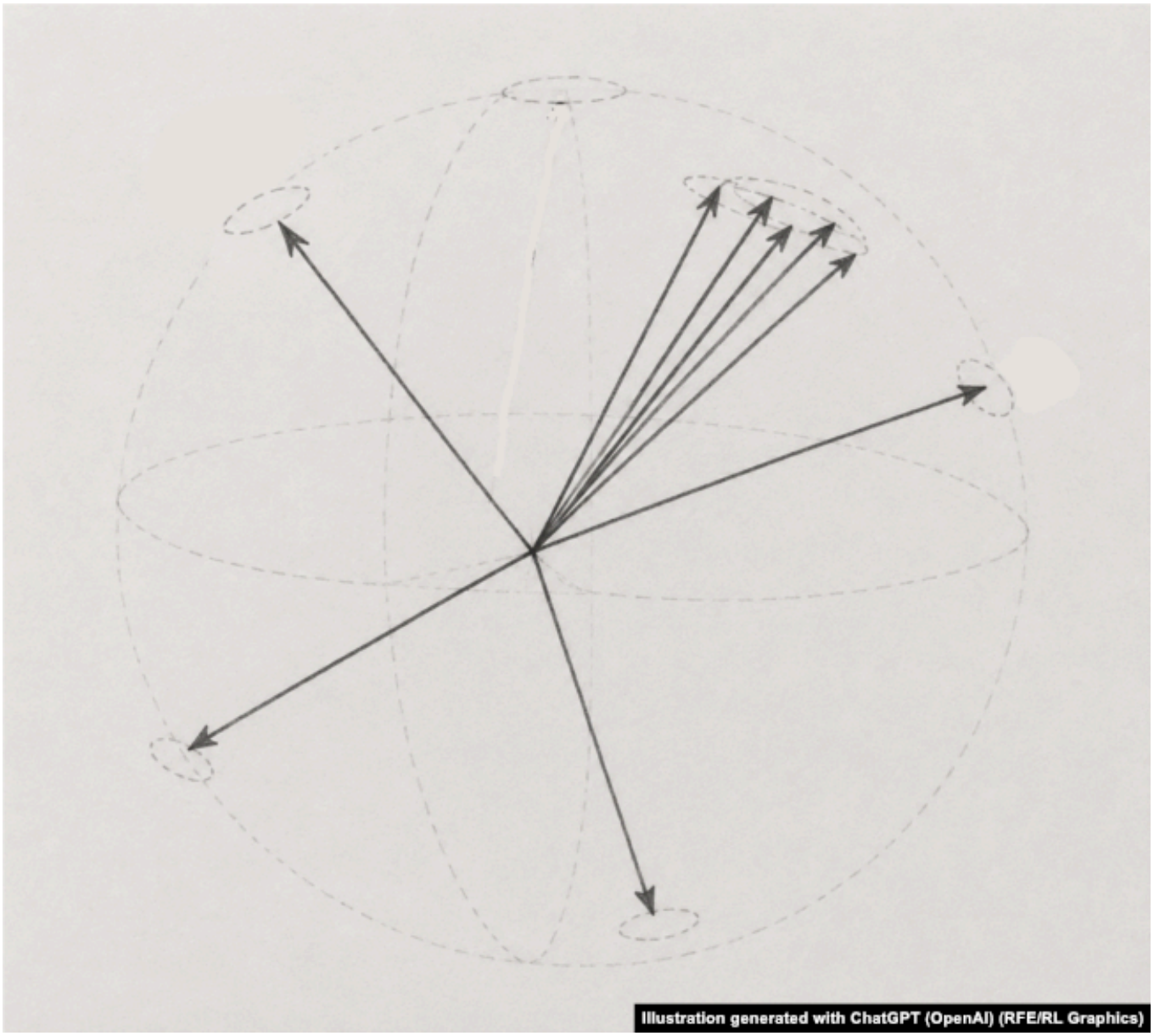
Как устроены модели AI: векторы в 12-тысячемерном пространстве, как отличить Никсона от Кеннеди, использование AI в военной сфере и конституция Anthropic, кому принадлежит AI: правительству, олигархам или народу, – все это в новом выпуске "Популярной Америки" обсуждают **Максим Рагинский**, профессор департамента электротехники и компьютерной инженерии Иллинойсского

университета, и ведущий подкаста Радио Свобода "Читая новости" **Валентин Барышников**.

Выдержки из подкаста:

– Я хочу поговорить о столкновении искусственного интеллекта и политики, это на глазах становится одной из главных проблем нашего времени, что видно хотя бы по взаимоотношениям администрации США и Anthropic. Но чтобы разобраться как следует, нужен краш-курс в AI. Большие языковые модели, чат-боты, с которыми мы теперь чатимся все время, правильно ли я понимаю, что каждое слово или ключевая часть слова, токены, представлены векторами в двенадцатитысячемерном пространстве (векторы тут можно считать стрелками, торчащими из одной точки в разных направлениях). Когда я задаю вопрос AI, алгоритм смотрит на эти отдельные слова-векторы, и потом строит некий новый вектор , который показывает, как надо преобразовывать это 12000-мерное пространство, чтобы слова-вопросы, крутясь, производили слова-ответы. Это верное представление?

– Примерно так, да. Токены – единицы, на которые разбиваются слова, иногда это могут быть целые слова, представлены векторами в пространстве размерности примерно 12 тысяч. У каждого слова (токена) есть такое смысловое облако, определяется соотношением с другими словами, часто встречающимися и употребляемыми вместе с ним. Векторы образуют кластеры слов, понятий, которые направлены примерно в одну и ту же сторону в этом пространстве огромной размерности.

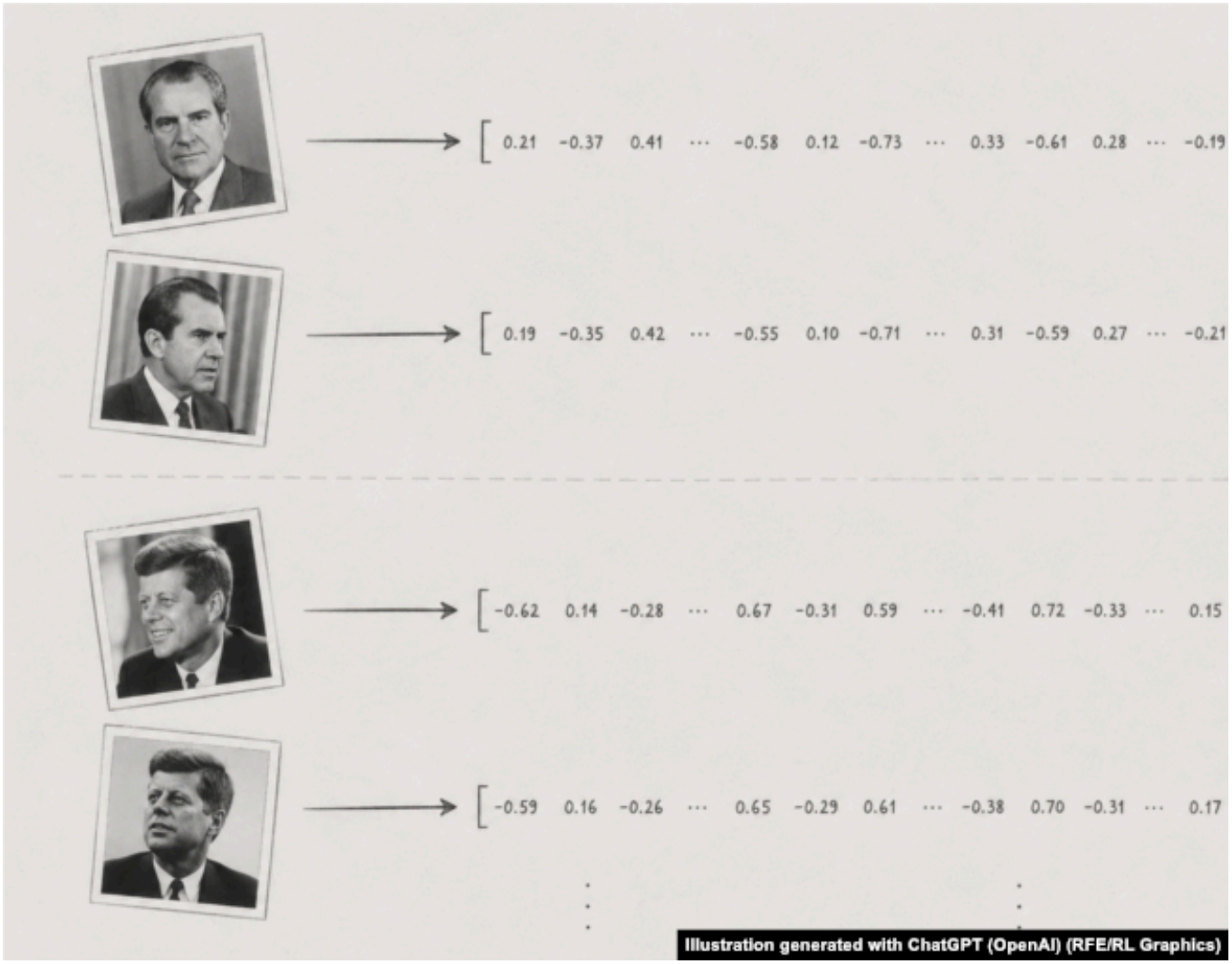


Иллюстрация, созданная ChatGPT (OpenAI).

Чат-бот получает то, что называется "контекстным окном", которое сейчас в самых продвинутых моделях, наверное, состоит из миллиона токенов, этот контекст создает некоторое поле притяжения, и вектора-токены, к которым притяжение больше в этом контексте, становятся вариантом ответа, который выдаёт чат-бот на запрос.

– Поговорим о машинном обучении, о том, как AI тренируют понимать картинки. Например, алгоритму дают десяток фотографий президента Никсона, десяток фотографий президента Кеннеди, алгоритм смотрит на последовательности пикселей (в типичном фото на смартфоне 12 миллионов пикселей, по три числа на каждый пиксель из-за цвета, это последовательность из 36 миллионов чисел). В этой цифровой развертке фотографии алгоритм ищет закономерности, которые позволяют ему потом в произвольной фотографии узнавать Никсона или Кеннеди – или отличать кошку от собаки. Мы, люди, прекрасно отличаем кошку от собаки без всяких вычислений, но про AI точно не знаем, как выглядит это отличие в

последовательности чисел, эти закономерности. И для меня вопрос, а кто-нибудь понимает, как выглядит с точки зрения AI в этих числовых последовательностях отличие кошки от собаки или Никсона от Кеннеди?



Иллюстрация, созданная ChatGPT (OpenAI).

– Скорее всего, нет. Примерно в том же смысле, что мы довольно плохо понимаем, как отличия Никсона от Кеннеди или кошки от собаки выглядят на уровне нейронных представлений в человеческом мозгу. Возможно, есть структуры, нацеленные на какие-то определенные типы признаков, и потом эти признаки объединяются, и таким образом цепочка абстракции нас приводит к зрительному образу. С моделями искусственного интеллекта, с одной стороны, проще, потому что мы можем проследить, как векторные представления эволюционируют, когда проходят через слои модели, а с другой стороны, система действительно сложная, однозначно понять, как возникают эти абстракции, тоже не очень возможно.

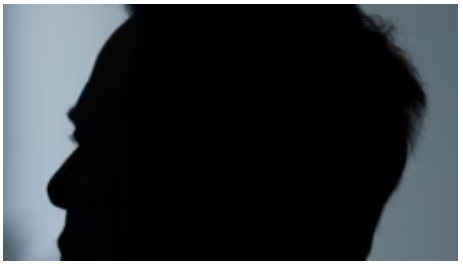
– То есть инженер, который тренирует AI-модель, дает ей необходимые данные, картинки, примерно указывает схему для алгоритмов, как действовать. Но в остальном, это черный ящик? И даже инженер, скажем, Anthropic или OpenAI,

тоже не может посмотреть на куски последовательности и сказать: "А, вон она, кошка"?

– Механического понимания нет в том смысле, что мы не знаем точно, каким образом определенное представление, понятие будет реализовано в системе. Психологи, чтобы изучить процесс формирования понятий, процесс закрепления привычек и поведенческих паттернов, использовали методы статистики, многофакторный анализ. Эти же методы применяются инженерами Anthropic и OpenAI для того, чтобы проникнуть вглубь систем и понять, как образуются понятия, можно ли направлять эти модели в какое-то заданное русло или наоборот попытаться удержать эволюцию в нежелательные области, где смысловые понятия противоречат каким-то этическим установкам или соображениям безопасности.

– С вопроса о безопасности начинается политика. Anthropic создал некие новые модели, которые якобы способны очень хорошо взламывать компьютерные системы и, соответственно, могут помочь при взламывании и могут помочь при защите от взламывания. И администрация США запретила их экспорт. Для начала я не понимаю, как это устроено.

– Теперь эти самые продвинутые модели Anthropic подпадают под действие экспортного контроля, запрещены к пользованию вне границы Соединенных Штатов. А поскольку у Anthropic нет возможностей проследить каждого пользователя и проверить, действительно ли это гражданин Соединенных Штатов, то они просто отключили вот эти модели для всех вообще. Anthropic был основан выходцами из OpenAI, которые пришли к выводу, что отношение OpenAI к проблеме безопасности недостаточно серьезное, и основали свою собственную компанию, где в основу легли их философские представления о том, как вообще должна выглядеть этика отношений человека и искусственного интеллекта.



СМОТРИ ТАКЖЕ

"Кольцо власти" Альтмана. Кто контролирует Искусственный Интеллект

Эволюцию модели Клод нужно рассматривать в этом свете. Их тактика состоит в определенном нагнетании страха перед тем, что эти модели могут из себя представлять, если выйдут из-под контроля, у них имидж ответственных инженеров, которые не дадут им выйти из-под контроля, они будут воспитывать эти модели. В основу систем типа Клод легла так называемая конституция. Это понятие, введенное одной из сотрудниц Anthropic Амандой Аскелл, получившей

философское образование в Оксфорде. Оксфордский подход к проблеме искусственного интеллекта основан на концепциях утилитаризма, то есть на том, что нужно причинить как можно меньше вреда и максимизировать пользу. На основе этого была выведена идея Конституции AI. Система Клод должна сверяться с этим документом, чтобы не причинить вреда, чтобы быть максимально полезной и удовлетворять другим этическим требованиям. И вот на этом фоне [было сделано] заявление Дарио Амодея [об ограничениях на развитие AI].

*

– Искусственный интеллект – это новое средство производства, все люди в конечном итоге будут его использовать. Если использовать марксистский подход, сейчас он принадлежит владельцам крупных AI-компаний, они являются новыми капиталистами. Кому будет принадлежит на самом деле AI? Интернет, например, раньше принадлежал отдельным лабораториям, но теперь его все используют, и доступ к интернету, кажется, – народное достояние. Искусственный интеллект тоже будет народным достоянием?

– Хотелось бы. Эта мысль была высказана папой Львом XIV. Противопоставление олигархической модели совместному проекту, когда все участвуют в освоении технологий и привносят свои ценности и взаимодействуют. В его последней энциклике были две метафоры: Вавилонская башня, олигархический проект унификации, когда все человеческие ценности сводятся к единому языку, который будет находиться под контролем олигархов.И восстановление стен Иерусалима в книге Неемии в Ветхом Завете, общественный проект, когда у каждого есть участок стены, который они восстанавливают, взаимодействие и сотрудничество.

Валентин Барышников

Этот контент также в категориях

Выбор Свободы

Популярная Америка