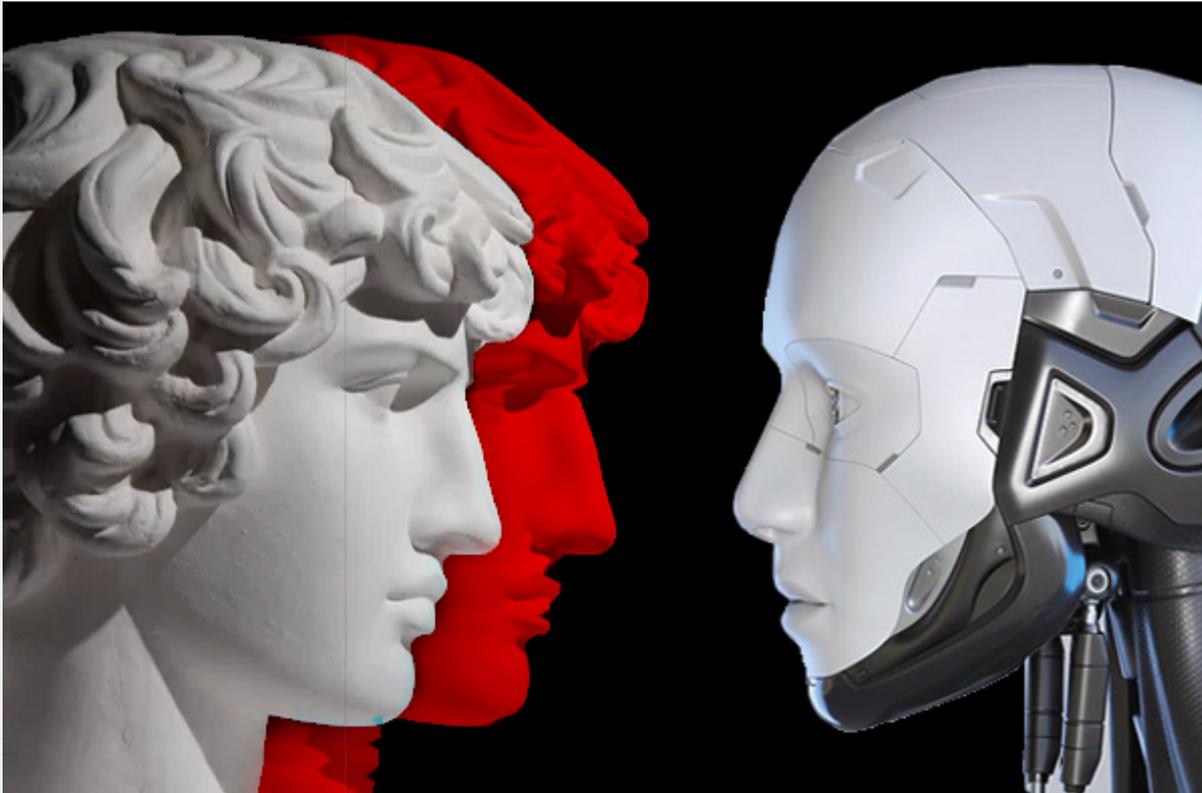


## Этика искусственного интеллекта. Часть 2. Предвзятость систем ИИ



*Пока философы рассуждали (см. первую часть), возможно ли создать сильный искусственный интеллект, математики и программисты работали.*

Начнём наш разговор с упоминания, наверное самого известного примера, когда искусственный интеллект компьютер переиграл человека и сделал то, что считалось до этого невозможным. В 2014 году программа Eugene Goostman впервые прошла тест Тьюринга. Замечу, что это произошло без применения машинного обучения, а лишь с использованием комбинации широкого дерева диалогов и ухода от тех вопросов, которые в дереве не прописаны. Если вспомнить определение искусственного интеллекта, то это вполне он. Собеседник позиционировал себя как украинский мальчик, плохо говорящий на английском языке, т.е. защищал себя от сложных вопросов сразу и возрастом, и незнанием языка. Сегодняшние чат-боты стали ещё более совершенны, например, когда автору звонят по телефону из каких-либо компаний, он всё чаще задумывается, разговаривает ли он с роботом или с живым человеком. Забавные и вполне разумные диалоги случаются порой даже с Алисой от Яндекса.

Современные системы искусственного интеллекта имеют узкую направленность (то есть являются в общепринятой терминологии «слабым искусственным интеллектом») и могут решать только одну конкретную задачу, например, распознавать речь или играть в игру го. Всё чаще стали появляться системы, которые выполняют целый класс похожих задач: например, нейронка GPT-3, созданная в прошлом году компанией Илона Маска OpenAI, способна писать компьютерный код, художественные тексты и поддерживать диалог.

Сама идея машины, имитирующей человеческий интеллект — что является одним из самых старых определений ИИ — вызывает опасения по поводу обмана, особенно если ИИ встроен в роботов, которые выглядят или действуют как люди. Чем больше свободы у машины, тем больше она будет нуждаться в моральных стандартах. Сегодня это очевидно и легко иллюстрируется в контексте автономного транспорта. Например, если на дороге возникла критическая ситуация, то как должен повести себя автономный автомобиль, если ему дано лишь три варианта для выбора: задавить старушку, сбить ребенка или направить автомобиль в обрыв с горной рекой, подвергнув опасности водителя – белого трансгендерного мужчину?

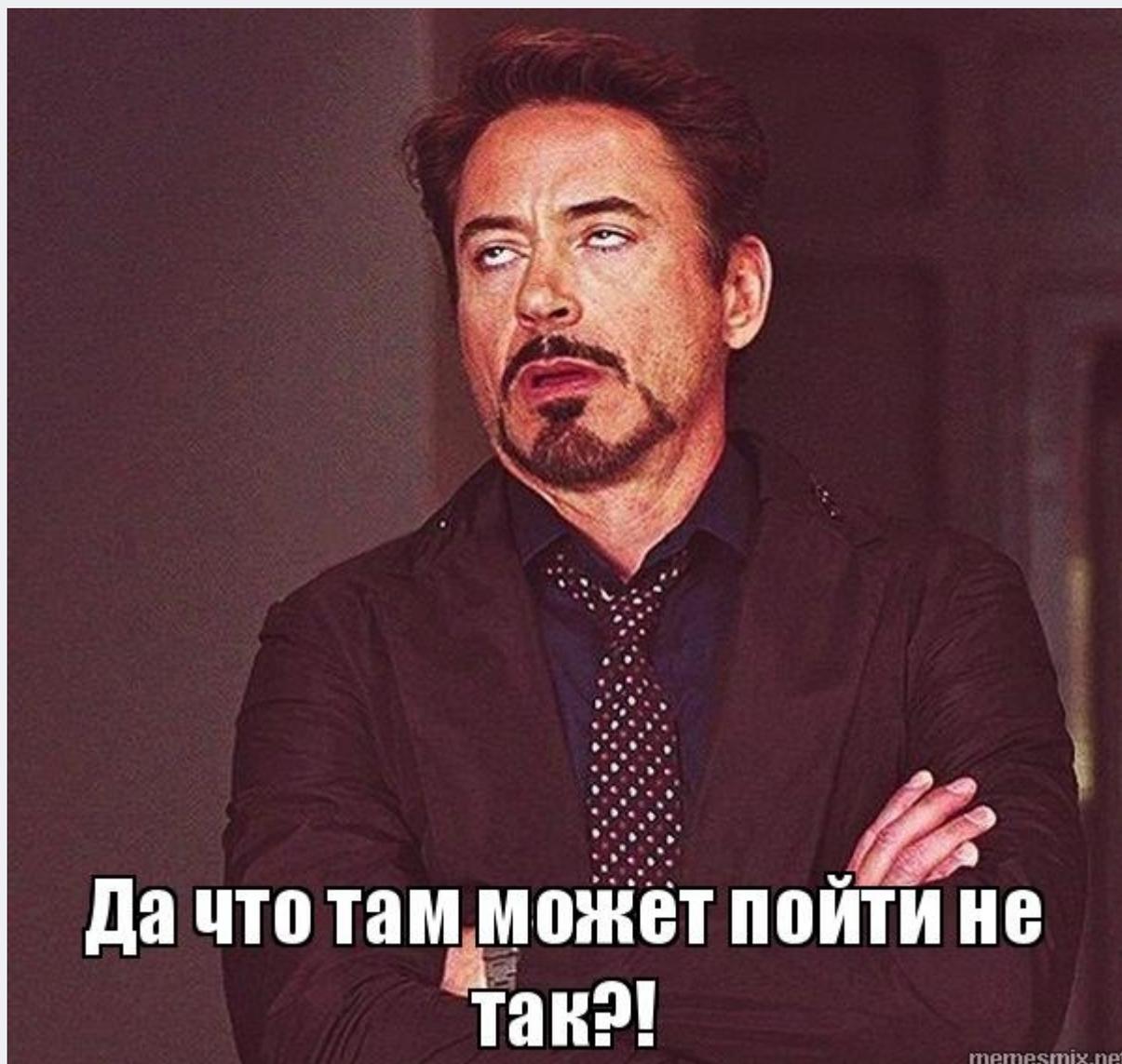
Другие примеры касаются также очень чувствительных в современном мире тем: как избежать машинной предвзятости в юридических делах, в принятии решений о найме, как избежать создания расистских и сексистских чат-ботов или негендерно-нейтральных языковых переводов? И кто будет нести ответственность за любую ошибку, которая закрадётся в эти решения? А на самом деле ошибки возникают постоянно.

Самый первый случай, когда алгоритмы показали свою предвзятость, по-видимому, произошёл в 70-80-е годы XX века. В это время Медицинская школа больницы Св. Георгия в Великобритании использовала компьютерную программу для первоначального отбора абитуриентов. Программа, которая имитировала выбор, сделанный приемной комиссией в прошлом, отказала в собеседовании шестидесяти кандидатам, потому что они были женщинами или имели не европейские имена. Такая дискриминация возникла не благодаря злему умыслу, а лишь потому, что была фактически заложена в данные ранее принятых решений. То есть ранее люди, безусловно без злого умысла, принимали такие «дискриминирующие» решения, руководствуясь, по видимому тем, что еврей-мужчины были лучшими специалистами, ну или так считала комиссия. Компьютерная программа по сути лишь усугубила проблему.

Прошло чуть менее половины столетия, и в 2018 году вспыхнул новый скандал, связанный трудовыми правами женщин.

Специалисты по машинному обучению компании Amazon обнаружили, что их новый механизм рекрутинга «не любил» женщин. Самообучаемая система Amazon проанализировала введенные данные о принятых ранее на различные позиции кандидатах и «научилась» тому, что кандидаты-мужчины предпочтительнее. Применение алгоритмов искусственного интеллекта в компании было остановлено.

Вот ещё пара примеров, уже связанные с чат-ботами. В 2016 году компания Microsoft провела эксперимент по самообучению искусственного интеллекта, создав в Twitter чат-бота Tay. Что могло пойти не так, спросите вы? Всего за сутки пользователи сети научили чат-бота неприлично ругаться и сделали из него расиста. Эксперимент свернули. Прошло пять лет и история повторяется, но уже в Facebook. Южнокорейский Facebook-бот Lee Luda общался с пользователями от лица 20-летней студентки колледжа. Во время беседы с одним из пользователей бот заявил, что «действительно ненавидит» лесбиянок и считает их «отвратительными». В общении с другими пользователями бот негативно отзывался о представителях негроидной расы. Снова скандал. Бота отключают.



Однако, и это очевидно, системы искусственного интеллекта могут реагировать намного быстрее, чем люди — и работать без необходимости отдыха и заработной платы. ИИ очень выгодны бизнесу. Поэтому уйти от их использования нам не удастся. В дальнейшем все больше ответственности будет перекладываться с людей на автономные системы искусственного интеллекта.

Мы смотрим из сегодняшнего дня в будущее и кажется, что прогресс не остановить. Однажды совершенно будничным образом появится машина, с которой вы сможете вести вполне вдумчивую беседу, и ни разу у вас не появится чувства, что вас не понимают. Даже наоборот, она вас будет понимать лучше, чем кто-либо из ваших родных и друзей. Так что же нам делать с этим надвигающимся миром будущего? Способны ли мы предложить ему новую этическую парадигму?

Если принципиально ничего не изменится, то будущее существование человечества будет зависеть от реализации твердых моральных стандартов в системах ИИ. Для того, чтобы в тот момент, когда машинный интеллект будет соответствовать и даже превосходить человеческий разум, который, возможно, никогда не наступит, а, может, и случится совсем скоро, эти машины нас не уничтожили в одно прекрасное утро. Как бы нереалистичным звучала эта угроза, о ней на полном серьезе предупреждают знаменитый драматург Карл Чапек, известный астрофизик Стивен Хокинг, влиятельный философ Ник Бостром и писатель, философ-рационалист Элизер Юдковски.

В чём же сложность этой проблемы? Первый этический кодекс для систем искусственного интеллекта был введен известным

писателем-фантастом Айзеком Азимовым в 1942 году (во время войны думать о законах робототехники, Карл!). Законы робототехники Азимова формулируются следующим образом:

1. Робот не может причинить вред человеку или своим бездействием допустить причинение вреда человеку
2. Робот должен подчиняться приказам людей, за исключением случаев, когда такие приказы противоречат первому закону
3. Робот должен защищать свое существование до тех пор, пока такая защита не противоречит первому или второму закону

Ирония заключается в том, что затем на протяжении всех последующих лет жизни, почти трёх десятков рассказов и нескольких романов маэстро фантастики доказывает полную несостоятельность этих абсолютно логичных, непротиворечивых и интуитивно понятных законов. Так есть ли у нас надежда? Очертив проблему, поговорим об этом наследующей неделе.

*Автор: Андрей Мурачёв,  
научный сотрудник Высшей школы теоретической механики*