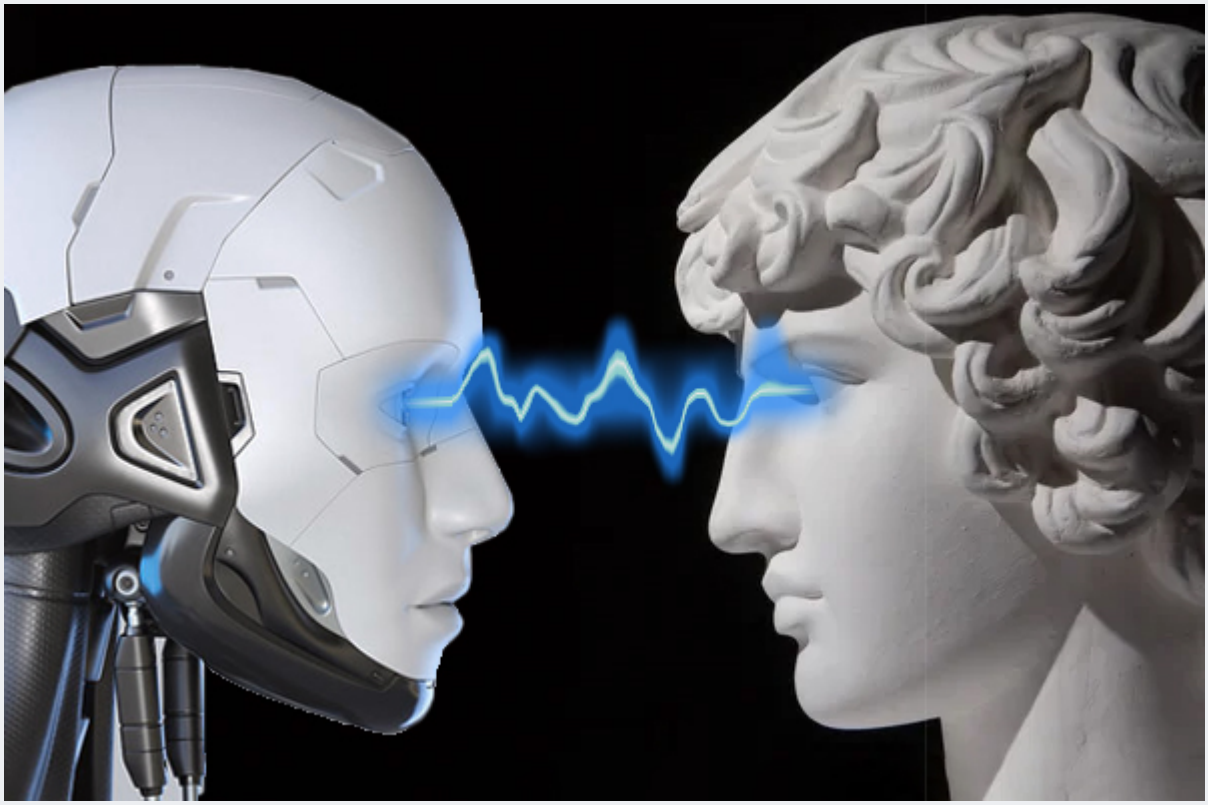
Этика искусственного интеллекта. Часть 1. Машинный разум



Около месяца назад в Москве на I Международном форуме «Этика искусственного интеллекта: начало доверия» был подписан кодекс, который, по задумке авторов, должен внести хоть какое-то морально-этическое регулирование в разработку систем ИИ. Сдержитесь, пожалуйста, от фейспалма. Организациями, которые присоединились к Кодексу этики ИИ, стали крупнейшие российские технологические компании и университеты, такие как Яндекс, Газпром нефть, Университеты Иннополис и ИТМО и другие.

Это событие могло бы быть хорошим поводом поговорить о морали и ответственности в эпоху искусственного интеллекта, если бы оно хоть что-нибудь значило. То, что кодекс служит скорее для пиара, чем для чего-то действительно важного, можно понять даже из пункта 2.1:

> *Присоединение к Кодексу является добровольным. Присоединяясь к Кодексу, Акторы ИИ на добровольной основе соглашаются следовать его рекомендациям. Присоединение и следование положениям настоящего Кодекса может приниматься во внимание при предоставлении мер поддержки или ином взаимодействии с Актором ИИ или между Акторами ИИ.*

Так что давайте согласимся, что повод плохой, однако важности затронутой проблемы, которая уже не первый год обсуждается во всех развитых странах, этот вывод не принижает. Наоборот, я считаю, что величие нации определяется проблемой, которую эта нация способна решить, а вынесение проблемы в общественное пространство для дискуссии является хоть каким-то сигналом к тому, что в нашем богоспасаемом отечестве не всё потеряно. В конце концов без этого кодекса не было бы этих заметок.

Если говорить о суперкомпьютерных мощностях, которые достаточно сильно коррелируют с разрабатываемыми решениями в области ИИ, то Россия хоть и входит в топ-10 стран с наибольшим количеством суперкомпьютеров, отставание от лидеров у нас колоссально. Рейтинг возглавляет, конечно, Китай – 173 суперкомпьютера. Ему в затылок дышит США – 149 суперкомпьютеров. Потом идут Япония, Германия, Франция, Великобритания, Нидерланды, Канада, Южная Корея, а мы замыкаем список со своими семью суперкомпьютерами. К слову, этот рейтинг неплохо иллюстрирует правило 20/80: Китай и США вместе располагают 72% всех суперкомпьютеров в первой десятке. То, что мы попали в топ-10, – заслуга Яндекса, которому, видимо, надоело терпеть насмешки Сбера в свой адрес и он в ноябре 2021 ввёл в строй сразу три машины. Как говорится, vive la concurrence.

Конечно, компьютерные мощности не определяют мощность созданного ИИ, но что-то мне подсказывает, что какой бы критерий для оценки ИИ мы бы не выбрали, будь то количество стартапов, разрабатывающих решения для ИИ, или количество обучаемых параметров в самой мощной нейронке, в итоге мы получим упорядоченный список стран, не сильно отличающийся от вышеприведённого. Таким образом, разговор, конечно, интересный, но *Россия пока физически не способна повлиять на мировые процессы в этой области*. Однако мы вполне можем адаптировать на своей почве иностранные решения и надеяться, что когда-нибудь всё изменится.

Давайте теперь вернёмся к заявленной в заголовке этого текста теме. На сегодняшний день вполне очевидно, что среди стран-лидеров в области ИИ нет устоявшегося мнения ни о том, как достичь и что считать сильным искусственным интеллектом — рукотворным разумом по мощности сравнимым с человеческим (та голубая мечта, о которой грезит всё человечество благодаря золотой эпохе фантастики 60-х годов), ни о том, как прописать моральные нормы для роботов, и кто понесёт ответственность,

если роботы вдруг начнут их нарушать. Продолжая разговор, мы вступаем на очень зыбкую почву философствования и рискуем увязнуть ещё больше. Однако я считаю, что полезно хоть изредка отходить от принципа “Shut up and calculate”. По большому счёту, если мы хотим разобраться в моральных дилеммах ИИ, у нас нет другого выбора, ибо современная наука не способна дать окончательное решение этого вопроса, а науке будущей, которая сможет это сделать, надо начинать с аксиом, выведенных философией.

Давайте начнём с основ. *Способна ли машина думать? Может ли она быть разумной?* Вспомните о какой-нибудь игрушке, которую очень любили в детстве, а сейчас к ней даже не прикасаетесь или вообще выбросили. Чувствуете ли вы свою ответственность перед этой игрушкой? Может, у части из вас в груди родилось какое-то щемящее чувство жалости к этой игрушке. Если вы начнёте разбираться в себе, вы обнаружите, что это чувство вызвано тем, что вы переносите человеческие эмоции на определённо неодушевленный предмет, вещь. Вы словно наделяете её чувствами и смотрите на происходящие её глазами. Как бы она чувствовала себя в данной ситуации? В этот момент произошла антропоформация игрушки. Вы сделали её в своих глазах человеком.

Сегодня вы не сомневаетесь в том, что игрушка ни мыслить, ни испытывать эмоции не способна. А робот может? Ведь говоря о морали роботов мы неизбежно наделяем их разумом. Глупо требовать морального поведения от камня, облака или старой игрушки. Глупо наделять их субъектностью. *В чём же отличие робота от игрушки и в чём отличие робота от живого человека?* Алан Тьюринг в своей основополагающей работе «Вычислительные машины и разум» 1950 года задаётся этими же вопросами. Он говорит, что роботы, поведение которых вычислимо с помощью универсальной вычислительной машины («машины Тьюринга»), не могут «думать» в человеческом понимании этого слова, они могут лишь совершать действия, неотличимые от разумных. Из этого умозаключения родился всем известный тест Тьюринга. Уильям Росс Эшби, английский психиатр и кибирнетик, в эссе «Что такое разумная машина» 1963 года пишет, по сути, о том же самом: *«разумен тот, кто разумно действует»*.

С самого начала Тьюрингу возражали, что языковое поведение – это еще не все мышление. Тьюринг бы, наверное, на это ответил *«не спорю, но придумайте критерий лучше»*. Прошло уже 70 лет, а мы так и не придумали. Между тем, надежды на создание кибернетических големов оставались.

В конце 20 и в начале 21 века конечной целью многих компьютерных специалистов и инженеров было создание надежной системы искусственного интеллекта, которая не отличалась бы от человеческого интеллекта ни в каком аспекте, кроме своего машинного происхождения. В 1980 году выдающийся американский философ Джон Сёрл представил так называемый аргумент китайской комнаты, призванный показать, что как бы близко мы не подошли к созданию «разумных» агентов, всегда будет оставаться неопределённость в том, обладает этот агент «внутренней разумной жизнью» или лишь демонстрирует разум. Главный тезис Сёрла заключался в том, что какой бы сложной ни была машина, она, тем не менее, наверняка не будет иметь «сознания» или «разума», что является необходимым условием для способности к пониманию. Сёрлу возражали так называемые функционалисты тем, что разум не требует определенного материального носителя, такого как, например, белковые структуры, и что он так же способен развиваться на основе кремния, если система будет достаточно сложна.

На этом дискуссия застопорилась. Ответ на этот вопрос нам может дать – неожиданно – астрофизика, если удастся обнаружить, например, кремниевую жизнь на Венере или Титане. В следующей части мы приступим к обсуждению непосредственно этики ИИ.

Автор: Андрей Мурачёв, научный сотрудник
Высшей школы теоретической механики