СМИРНОВ СЕРГЕЙ АЛЕВТИНОВИЧ

Ведущий научный сотрудник
Института философии
и права СО РАН,
доктор философских наук.
Новосибирск. Россия

Главный редактор
гуманитарного альманаха
«Человек.RU».

E-mail: smiroff1955@yandex.ru

УДК 004.8

ПРЕДЕЛЫ ЭТИКИ ДЛЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация. В статье ставится вопрос о возможности применения этической проблематики к области разработок моделей искусственного интеллекта. Автор разбирает гуманитарные и этические допущения, которые кладутся в основания данных моделей. Ставится под сомнение сама возможность создания так называемых «моральных машин», то есть систем искусственного интеллекта, способных к обучению этическому поведению. Делается вывод о невозможности этого обучения, поскольку этическое поведение не строится по таким же основаниям, по каким строится математическое формально-логическое доказательство. Ставится под сомнение, что этическое поведение человека, равно как и моральной машины, можно строить и объяснять по таким же формально-логическим правилам, по каким строится алгоритм, закладываемый в основание моделей AI. В статье подвергается критике так называемая «теория вагонетки», которая стала популярной в западной литературе и постепенно вытесняет классический этический дискурс. Согласно этой теории, этическая проблема ответственности поступка и свободы воли может строиться так же, как и любая логическая задача. В этой связи допускается сама возможность построения модели AI, в которую встроен алгоритм этического поведения. Дается краткий обзор уже существующих этических кодексов, разработанных в разных странах для регулирования этических проблем в сфере AI. Делается вывод, что эти кодексы в большинстве своем приняты не из этических и гуманных соображений, а сугубо исходя из соображений, связанных с проведением маркетинговой политики, с тем, чтобы клиенты и простые пользователи сервисов, использующих разные системы AI, были убеждены, что эти модели AI дружественны и не опасны для пользователей.

Ключевые слова: искусственный интеллект, непредсказуемость, этика, теория вагонетки, этический кодекс, этика для сферы искусственного интеллекта.

Исследование выполнено за счет гранта Российского научного фонда, проект No. 21-18-00103, https://rscf.ru/project/21-18-00103/

Sergey A. Smirnov

Institute of Philosophy and Law of the SB RAS
E-mail: smiroff1955@yandex.ru

ETHICAL BORDERS FOR ARTIFICIAL INTELLIGENCE

Abstract. The article raises the question of the possibility of applying ethical issues to the development of artificial intelligence models. The author examines the humanitarian and ethical assumptions that underlie these models. The very possibility of creating so-called "moral machines", that is, artificial intelligence systems capable of teaching ethical behavior, is questioned. The conclusion is made about the impossibility of this learning, since ethical behavior is not built on the same basis as a mathematical formal logical proof is built on. It is questioned that the ethical behavior of a person, as well as a moral machine, can be built and explained according to the same formal-logical rules that are used to build an algorithm that is the basis of AI models. The article criticizes the so-called "trolley theory", which has become popular in Western literature and is gradually replacing the classical ethical discourse. According to this theory, the ethical problem of the responsibility of an act and free will can be built in the same way as any logical problem. In this regard, the very possibility of building an AI model, in which an algorithm of ethical behavior is built in, is allowed. A brief overview of the already existing ethical codes developed in different countries to regulate ethical issues in the field of AI is given. It is concluded that these codes are mostly adopted not for ethical and humane reasons, but purely for marketing policy reasons, so that customers and ordinary users of services using different AI systems are convinced that these models AI friendly and not dangerous to users.

Keywords: artificial intelligence, unpredictability, ethics, trolley theory, code of ethics, ethics for the field of artificial intelligence.

The work was written within the framework of the grant project supported by the Russian Science Foundation Project No. 21-18-00103. https://rscf.ru/project/21-18-00103/

DOI: 10.32691/2410-0935-2023-18-40-54

Введение

Разработка и внедрение программ искусственного интеллекта (далее – ИИ) не просто набирают обороты в мире. Это направление становится приоритетным на самом высоком уровне. Приоритетность зафиксирована в принятых стратегических документах развития, в том числе и в нашей стране. Это вполне объяснимо, поскольку предполагается, что лидерство в этом направлении делает страну в целом конкурентоспособным игроком на мировом рынке передовых технологий.

Вместе с тем, при таком темпе развития мало кто задумывается о главной фигуре в этом тренде – о самом человеке, который разрабатывает модели ИИ. Понимаем ли мы, что происходит с самим человеком при развитии этого тренда? Задаём ли мы себе следующие вопросы? И как на них отвечаем?

Какие гуманитарные и антропологические основания закладываются при разработке моделей ИИ? Думают ли про это разработчики проектов ИИ?

Каковы не технологические, а гуманитарные последствия могут быть от внедрения проектов ИИ?

При развитии модели ИИ всё более мощно расширяется и углубляется технологический и функциональный аутсорсинг от человека к машине, то есть передача человеком все более умных функций умной машине. Возникает вопрос: что человек оставляет себе при разработке технологий ИИ и что передаёт умному устройству?

Где и когда при таком аутсорсинге наступает граница, за пределами которой передавать умному устройству уже будет нечего, поскольку человек, перейдя эту границу, перестаёт быть самим собой, поскольку все свои умные функции и работы он передал умной машине? И тогда действительно возникает вопрос о замене человека постчеловеком. Наступит ли когда-нибудь такая ситуация и мы упрёмся в границу?

Как и какой должен будет выстраиваться кентавр человека и ИИ в будущем? Какие здесь могут быть более предпочтительными модели? Какие должны быть новые интерфейсы человека и машины, в которых и человек, и машина предстают уже в новом функциональном качестве?

Сложилась ли уже ситуация, при которой возникает необходимость нового этического кодекса для ИИ и создания института гуманитарной экспертизы внедрения умных технологий? Что сейчас необходимо делать в сфере международного права с точки зрения гуманитарного этического регулирования в сфере разработки и внедрения технологий ИИ?

Подходы к модели ИИ

Прежде, чем начать отвечать на эти вопросы, необходимо остановиться на основном вопросе: какие модели человека закладываются при разработке ИИ и какие при этом гуманитарные допущения позволяют себе разработчики?

Определений ИИ в настоящее время достаточно много. Но самым продвинутым выступает определение, согласно которому под ИИ понимается такая программа (или алгоритм), которая на основе обработки данных научается «принимать решения в ситуации неопределенности» [Искусственный интеллект 2019: 13].

Что это значит? Это значит, что модель ИИ должна не только решать задачи по алгоритму, но и уметь *вести себя в непредвиденных ситуациях*, которые не были предусмотрены в заложенном в неё алгоритме. Интеллектуальной системе, в таком случае, ставится не только задача, но моделируются и сами ситуации, при которых эти задачи могут возникнуть. В модель тогда встраивается не только набор данных, не только система задач, но моделируются и сами ситуации, способные порождать те задачи, которые в алгоритме отсутствуют (рис. 1).

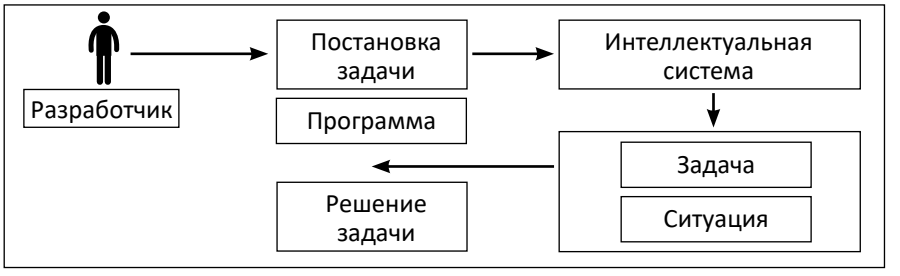

Рис. 1. Модель ИИ

Но такая система, принимающая решение, для разработчика выглядит как черный ящик, внутри которого что-то происходит. Разработчик не знает, почему и как система принимает те или иные решения. Да и сама система не знает, почему она принимает такие решения, находясь за занавесом неведения [Искусственный интеллект 2019: 157]. Такая модель ИИ решает задачи, но не знает, как она это делает. То есть модель ИИ в принципе лишена рефлексии. Но пока разработчика интересует другой вопрос – умеет ли модель ИИ работать с непредсказуемостью? Возникает вопрос – умеет ли сам человек работать с непредсказуемостью и как он это делает?

В таком случае приведём пример, взяв качестве аналогии модели непредсказуемости и неопределенности модель регулирования поведения водителей и пешеходов на дорогах большого города. Фактически город выступает реальным полигоном-моделью системы с большой степенью непредсказуемых ситуаций и последствий. В силу сталкивания многих тысяч воль и действий жителей городское пространство обитания становится реальным полигоном, на котором уже апробируются модели ИИ (например, система умных светофоров, система наблюдения и штрафов, использование беспилотников в качестве умных средств мониторинга дорожного движения и т. д.).

Каждый водитель, едущий в машине, не знает другого водителя в другой машине и не знает, как тот себя поведёт. Но и этот, и другой водитель, допускают, что все водители знают правила дорожного движения (далее – ПДД), то есть нормы, разработанные заранее в качестве всеобщего, принятого всеми *нормативного регулятора поведения* на дорогах. ПДД вводятся для того, чтобы минимизировать степень непредсказуемости поведения водителей и пешеходов на дорогах. ПДД вводятся для всех участников движения в качестве дополнительного над физическим полем – нормативного поля, в виде *системы знаков*. Но именно это поле, знаковое, становится главным регулятором, более важным, нежели само реальное пространство передвижений. Ориентируясь в нормативном поле, заданном через ПДД и систему знаков, человек

может передвигаться по физическому полю. Если он не умеет ориентироваться в нормативном поле или пренебрегает его правилами, то он рискует попасть в аварию. Таким образом, водитель или пешеход не только видят дорогу и другие машины, но также видят знаки, регулирующие движение, читают их, видя в них заложенные правилами нормы. Одни водители тем самым допускают, что и другие водители будут выполнять правила, закодированные в этих знаках. Тем самым и водители, и пешеходы начинают видеть это нормативное поле, посредством которого они видят и физическое поле. Тем самым жители города, передвигающиеся по его улицам, видят не просто передвижения людей и транспорта, но они видят *регулируемое передвижение с помощью знакового опосредования*.

Таким образом само наличие этого нормативного регулятора, вынесенного во вне водителей, во вне машин, становятся условием движения на дорогах и условием вхождения в это пространство передвижения. Без них будет стихия, хаос и непредсказуемые последствия (рис. 2).

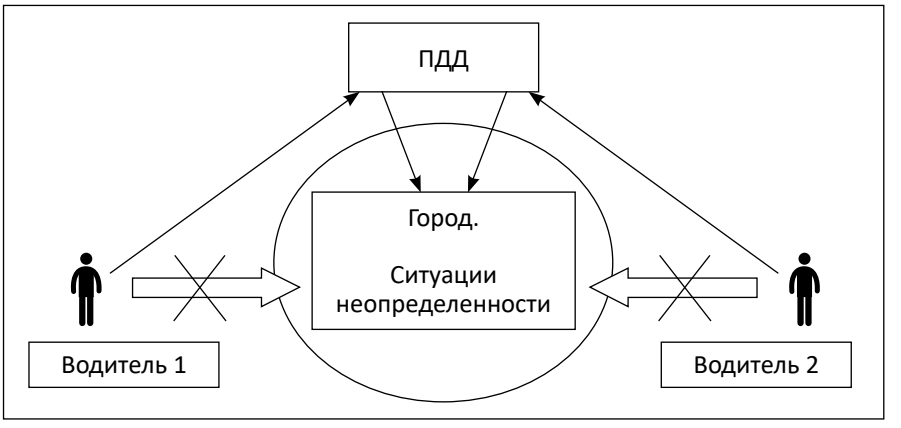

Рис. 2. Модель города как аналог ИИ. Модель поведения в нормативной рамке в ситуации неопределённости.

Если перевести эту ситуационную модель в область разработок ИИ, то при разработке моделей ИИ необходимо учитывать встраивание в них подобных нормативных систем, дабы эти модели «умели» предсказывать и действовать в ситуации непредсказуемости. Если подобная нормативная система будет встроена в модель ИИ, то поведение последней становится как бы «разумным». Или нет?

В этом допущении скрыто ключевое условие: соблюдение правил (шире – норм) становится условием безопасности и в целом функционирования всей системы, здесь – системы дорожного движения. Если водители и пешеходы не соблюдают правила, то происходят аварии и катастрофы. Заметим, что фактически все крупные техногенные аварии (включая Фукусиму и Чернобыль, ставшие джокерами в своё время) были рукотворными, связанными с нарушением правил безопасности.

Если перейти к теме ИИ в виде разработок беспилотников, новых участников ДД, то его создатели вынуждены закладывать в их модели как ПДД, так и все возможные сценарии поведения разных агентов на дорогах (машин, пешеходов и др.). Но возможно ли заложить в модель ИИ *весь репертуар непред-*

сказуемости поведения пьяных водителей и нерадивых пешеходов? Пока нет. Поэтому появились первые случаи аварий с участием беспилотных машин[1].

Заметим, что по большому счету, сам въезд в город, само вождение в городе (см. рис. 2) возможно лишь при условии знания водителем ПДД и сдачи соответствующих экзаменов. Что в мире и делается, хотя люди сами же эти правила постоянно нарушают. Но виноваты здесь не правила и не машины, а люди, позволяющие себе действия, не соответствующие этим правилам. Но само знание ПДД становится пропуском вхождения в большой город.

В таком случае создание ИИ и действие с ним должно выстраиваться по такой же схеме. Параллельно разработкам ИИ создаются и правила, этика для ИИ, а потом уже выстраиваются манипуляции и действия с ним. Знание и принятие норм выступает *условием* управления автоматическим транспортным средством. Знание и принятие нормы и прохождение соответствующей аттестации на знание нормы выступает условием вхождения в ситуацию со многими степенями непредсказуемости, с которыми сталкиваются автономные субъекты.

Главным фактором этой неопределенности выступает наличие действий многочисленных акторов, игроков, агентов, в данном случае водителей и пешеходов, вступающих в контакт. Но у них у всех в таком случае должен быть пропуск на вхождение в ситуацию неопределённости и непредсказуемости.

В таком случае ключевой проблемой при разработке ИИ становится не сама по себе разработка ИИ как умного технического устройства. Проблема не в технике, не в чёрном ящике, не в том, что и как делает нейросеть, выдающая разработчику непредсказуемые решения. Проблема в самом человеке, принимающем или нарушающем принятые им самим правила. Если их толковать шире – проблема в сохранении или размывании границ этоса, то есть глубинных устоев обитания, принятых человеком для своего собственного сохранения. Если он не будет удерживать и воспроизводить этос, то устои рухнут. Если сами горожане будут нарушать ПДД, принятые ими же, то город погрузится в хаос.

Достаточное ли это основание для определения ИИ? То есть такое, согласно которому модель ИИ может считаться разумной, если умеет работать *с правилами поведения и предвидеть непредсказуемые ситуации?*

Но проблема заключается в том, что сам человек нарушает собственные нормы и правила. Человек есть не тот, кто принимает разумные решения, а тот, кто их же и нарушает. Или тот, кто постоянно выходит за собственные границы. И это есть его собственное онтологическое качество – выходить за собственные границы и обустраивать новые сферы обитания. А выходит он за границы не в силу родовой деструктивности, а потому, что он так устроен, будучи существом трансцендентным. Это его способ обитания – раздвигание границ, преодоление ранее положенной нормы, с тем, чтобы далее в новой реальности воссоздавать себя и свои границы уже в новых условиях.

---

[1] В 2014 году автомобиль-беспилотник Google столкнулся с женщиной, находящейся в инвалидной коляске с электрической тягой, прогоняющей метлой проходящих через дорогу уток. Беспилотник не мог распознать данный экзотический для него объект [Искусственный интеллект 2019: 157]. Кстати, в Германии впервые в 2016 году был предложен законопроект о создании первой правовой базы для автономных транспортных средств. В России также разрабатываются изменения для ПДД в связи с беспилотниками. В 2016 году в США был принят первый документ, регламентирующий федеральную политику в отношении беспилотников.

Но в таком случае ИИ *невозможен в принципе*. Потому что граница, поставленная ранее, будет всегда нарушаться тем, кто эти границы поставил – человеком. Не успеет ИИ научиться работать с поставленными границами, завтра эти границы будут нарушены. Или в таком случае необходимо другое представление об ИИ: это такая программа, которая сама же и нарушает свой собственный алгоритм.

Но в таком случае главной проблемой становится не то, возможен ли ИИ или нет, заменит он человека или нет, какие интеллектуальные задачи он сможет решать в будущем. Главным становится сугубо этическая проблема – проблема границ этоса человека, способа и мира его обитания, проблема им же поставленных границ и того, как регулируется удержание этих границ. В данном случае речь идёт о границах человека и ИИ, человека и умной машины. То есть не проблема техническая, математическая, инженерная, информационная, а проблема сугубо гуманитарного, этического свойства. Здесь и рождается собственно этическая, шире – антропологическая, проблематика. То есть проблема границы человека и техники.

Теория вагонетки или этический тупик

Необходимо признать, что в основание своих рассуждений разработчики-инженеры закладывают допущение, что, учитывая рост непредсказуемых ситуаций, необходимость как-то встроить в свои модели ИИ этическую составляющую, связанную с безопасностью человека, возникает задача решить этическую проблему так же, как решаются формально-логические и математические задачи. Тем самым допускается, что в программу, в алгоритм ИИ реально заложить этические нормы, которые понимаются так же, как математические формулы. Иными словами, программу можно обучить этическому поведению. Допускается, что этическому поведению можно обучить и нейросеть. Допускается, что этическую норму или правило (как правильно поступать с точки зрения морали?) можно обосновать формально-логически, что машину можно обучить поступать этически так же, как и научить её совершать вычислительные операции. Значит, допускается, что этический выбор так же рационален и логичен, как решение математической задачи.

Такие допущения породили развитие так называемой теории вагонетки, или «вагонеткологии» (trolleyology), внедрённой в практику этических дискуссий с лёгкой руки разных западных авторов [Эдмондс 2020]. Ситуация или проблема вагонетки описывается мысленным экспериментом.

Представим себе, вы стоите на пути. По нему несётся неуправляемая вагонетка, у которой отказали тормоза. Вы видите, что впереди на пути лежат пятеро человек, они не могут убежать с пути и рискуют погибнуть под вагонеткой, поскольку они привязаны к рельсам. У вас есть возможность перевести стрелку, чтобы вагонетка пошла по другой ветке. Но на ней лежит один человек, который тоже не может уйти с пути, будучи привязан к рельсам. Вы можете перевести стрелку, но он погибнет, хотя пятеро останутся живы. Ваши действия?

Мне кажется, ситуация слишком надумана. Это, разумеется искусственная модельная ситуация, каковых много строится при разработке моделей взаимодействия людей и техники, попадающих в ситуации с большой степенью

рисков и неопределённости. Подобные модели полезны для развития практик мысленных экспериментов. Но по сути ситуация, выстроенная в этой модели, предельно цинична. Она предлагает участникам подобных мыслительных экспериментов поместить себя, хотя и в придуманные, но нечеловеческие условия. И представить – способен ли я убить одного ради спасения пятерых? Выбор однозначно тупиковый. Выбора здесь вообще быть не должно. Ведь этика – не арифметика. Нельзя там ставить минус, а там – плюс. Так бывает на войне – когда генерал жертвует в одном месте батальоном, чтобы в другом месте сделать прорыв. Но это на войне. Жизнь нельзя строить по стратегиям войны. Если уж строить модели, то такие, которые бы не загоняли человека в изначально тупиковую ситуацию, в которой любое решение – заведомо убийственное. Человек заведомо ставится в нечеловеческую зависимость от тупой вагонетки и необходимости делать выбор, однозначно и одинаково убийственный. Ведь непричинение вреда другому, и тем более невозможность спасения одного ради смерти другого – это непреложное этическое правило. Оно обязательно к исполнению.

Д. Эдмондс приводит в своей книге 10 таких ситуаций с вагонеткой, в которых ставится вопрос о выборе между тем, пустить ли вагонетку по пути, и ты задавишь пятерых, или пустить по другому пути, и ты задавишь одного. Также он приводит различные примеры выбора, с которыми столкнулись врачи при лечении больных, с которыми столкнулись военные на фронте во время боевых действий, спасатели во время чрезвычайных ситуаций. Но заметим, что все примеры, которые приводит Эдмондс, и на которых строится теория вагонетки, взяты из медицины, на войне, на происшествиях и чрезвычайных ситуациях, во время катастроф. То есть примеры заведомо чрезвычайные. Но жизнь человека нельзя рассматривать как сплошное происшествие, как постоянную войну и бесконечную болезнь. Жизнь человека в норме не может рассматриваться как жизнь на войне или в больнице. Авторы же теории вагонетки строят свои мысленные эксперименты сугубо на примерах чрезвычайных и далее пытаются распространить их на жизнь человека в норме, в нормальных повседневных условиях, когда нет войны, когда никто не болеет и когда нет пожаров и наводнений.

Тем не менее такая с позволения сказать прикладная этика или даже моральная философия пустила корни в западном философском дискурсе и вошла в списки различных опросов. Например, Д. Бурже и Д. Чалмерс не преминули вставить в свое социологическое исследование вопросы из теории вагонетки [Bourget, Chalmers 2021]:

«Задача вагонетки. Ехать прямо, тогда наедешь на пятерых, или переключить стрелку на другую линию, тогда наедешь на одного? Ваш ответ».

Какие ответы были получены?

Переключить – 63,42% опрошенных.

Не переключать – 13, 31%.

Другое – 24,88%.

Удивительно. В какую циничную ситуацию ставятся респонденты, на какой вопрос им предлагают ответить! И несмотря на все моральные муки и угрызения все же 13,31% ответили так, что не надо переключать стрелку. Мол, пусть едет вагонетка и давит всех пятерых. И авторы опроса, и респонденты

допускают вообще саму возможность рассуждать по поводу таких ситуаций. Это и есть этическая проблема? Куда делся Кантовский ригоризм? Куда ушла великая метафизика нравов? Моральный долг? Нравственный поступок? Самопожертвование ради другого? Почему в этой вагонеткологии не допускается третий вариант – тот, кто находится у стрелки или тот, кто сидит в вагонетке вообще вынужден себя просто взорвать, если движение неминуемо. И прямо, и направо – одинаково пути убийственные. А поэтому, если ты находишься в такой ситуации, ты жертвуешь собой, поскольку иного пути нет, и только ты сам ценою своей жизни и можешь остановить эту страшную вагонетку, которая становится слепым орудием убийства. Но авторы вагонетковедения почему-то сидят и рассуждают, но никак не допускают и мысли о таком варианте.

В книжке Д. Эдмондса рассуждения о вагонетках и иных ситуациях морального выбора перемежаются рассказами из истории моральной философии. Интересно, познавательно. Но мысленный эксперимент с вагонеткой, мне кажется, перечеркивает эту увлекательность и ставит вообще крест на проблеме нравственности. Этика исчезает под грудой обломков мчащейся слепой вагонетки. Последняя становится метафорой всего происходящего тренда: за мчащимся поездом технического прогресса мы перестали видеть человека, перестали понимать самих себя. Увлекательная игра с нейросетями несёт нас в тупик, раздавливая человека.

В то же время налицо явно выраженная тенденция, описанная в литературе, состоящая из двух направлений: с одной стороны, машинизация человека, превращение его в функциональное устройство, редукция его поведения до функции, с другой стороны, уподобление машины человеку, превращение её в некое человекоподобное устройство [Смирнов 2022; Смирнов 2023].

Подобная тенденция воплощается в том, что разработчики и исследователи допускают возможность моделей так называемых «моральных машин», в которых возможен учёт этических норм, как будто этические нормы могут быть включены в алгоритм поведения машины, вплоть до того, что проблема свободы воли может быть как-то учтена в моделях ИИ, если не сейчас, то не в столь отдалённом будущем [Карпов и др. 2018; Разин 2019].

Исследователи допускают, что обсуждать этику ИИ нельзя без обсуждения проблемы свободы воли [Разин 2019: 58]. Но они допускают в своих рассуждениях и способность ИИ к научению этическому поведению, способность ИИ принимать этически взвешенные решения. Вводится очередная метафора – «моральные машины». Разработчики и исследователи всерьез допускают, что ИИ можно научить не только решать нетривиальные задачи, но и научить этическому поведению [Moral Machine 2023; Bostrom 2011][2]. Это означает, что этическое поведение можно описать так же строго, математически точно. Последнее, разумеется, связано с выше приведёнными примерами из теории вагонетки. Коль скоро этическое действие так же можно рассматривать, как поведение рациональное, логически выверенное, и его можно рассчитать и по-

[2] В публикациях, посвященных моральным машинам, при этом обучение этическому поведению ставится в том же духе, что и в теории вагонетки. Например, для испытания беспилотных машин ставятся задачи типа: какое из двух зол выберет самоуправляемый автомобиль – гибель двух пассажиров или пяти пешеходов? Обучение этическому поведению строится по той же теории вагонетки. Цинизм задачи закладывается в алгоритм поведения моральной машины.

строить по этому поводу алгоритм, то в принципе машину можно научить совершать этически обоснованные поступки.

Исследователи допускают, что если человек есть субъект, обладающий свободой воли, с которым связана его моральная и правовая ответственность, то именно эти качества должны быть заложены в модели ИИ как параметры. Это значит, что модель ИИ должна быть способна реагировать на случайные (не входящие в алгоритм) события, неповторимые ситуации, уметь принимать адекватные решения, быть способной на нравственные действия. Но коль скоро этому всему человек сам научается поэтапно, в процессе культурного развития в рамках онтогенеза, то тем самым ИИ надо этому тоже учить. Поэтому необходима модель поэтапного обучения ИИ этическому поведению, так же, как есть поэтапное обучение человека [Разин 2019]. Полагаю, что упование на создание подобных моральных машин, способных на этическое поведение – тоже тупик, ещё одно безмерное увлечение разработчиков и исследователей.

Этика для сферы ИИ

Итак, мы не будем идти по пути вагонеткологии и по пути создания моральных машин. Зададим себе вопрос еще раз – где и когда рождается этическая проблематика в ситуациях с разработкой моделей ИИ? Ответ такой: там и тогда, когда разрабатываются правила и нормы, регулирующие взаимодействие людей в ситуации разнообразия воль и действий, непредсказуемости и высокой степени сложности и рисков. Этика рождается не в инженерии, а в ситуации множественности свободных воль и действий. В самом ИИ этики нет и быть не может. ИИ научить этике нельзя по определению, поскольку ИИ не является субъектом действия. Этика выступает регулятором, с помощью которого регулируются отношения между людьми, а не между человеком и машиной. Здесь – ключевая проблема доверия не к ИИ, а к человеку, автору ИИ.

Этика рождается тогда, когда мы сталкиваемся с рисками и угрозами, связанными с разработкой и внедрением ИИ в повседневность, в результате чего растёт степень непредсказуемости, а поэтому необходимо выработать кодексы и конвенции, регулирующие деятельность людей в сфере ИИ. Таких конвенций выработано в последние годы уже много [Кодекс 2021; Принципы этики 2021; Proposal 2021; AI Principles 2017; ЮНЕСКО 2021].

Казалось бы, человечество всерьез задумалось о том, что пора осмыслять этическую проблематику при разработке умных систем и не увлекаться одними инженерными задачами. То есть пора вспомнить о человеке.

Но зададим себе два вопроса:

1. Какая модель человека при этом закладывается в эти принятые проекты конвенций и этических деклараций?
2. Не стоит ли за разработкой этих конвенций и деклараций всего-навсего маркетинговая стратегия? Может, авторы этих конвенций, в числе которых крупные корпорации (в России это компания «Сбер»), стремятся просто убедить клиентов в том, что ИИ не обманет, он дружественный, его не надо бояться? Но тогда это всего-навсего маркетинговый ход, который продавец той или иной умной технологии использует для того,

чтобы убедить клиента и покупателя купить его замечательную технологию.

Что касается первого вопроса, то пока ситуация выглядит следующим образом.

При разработке фактически всех этических конвенций и деклараций их авторы делают два фундаментальных допущения:

- человек не меняется, он всё такой же, каким был и тысячу лет назад,
- история цивилизации есть история освобождения человека, становления его автономной личностью, он достоин лучшей, то есть комфортной и безопасной жизни. Последнее ему и должен обеспечить в том числе ИИ.

При этом допускается, что это благополучие и комфорт требуется защитить от новых рисков, связанных с развитием умных технологий. Всегда при разработке и внедрении сложных инженерных систем требованиям безопасности уделялось серьёзное внимание, будь то строительство АЭС или ракетных комплексов, или разработка ИИ. В данном случае она и должна быть выработана и принята цивилизованным сообществом.

Но проблема в том, что ИИ, внедряясь в реальность, в повседневность человека, реально начинает менять и его устои жизни, его этос, его среду обитания. И речь идёт уже не о правилах поведения и технической безопасности, а о готовности самого человека отказаться от этих устоев. В силу чего отказ ставит под вопрос саму привычную норму человека, его представления о самом себе.

В этой связи задачей становится не сама по себе разработка ИИ и не только выработка мер защиты человека от возможных угроз и рисков, связанных с внедрением ИИ, а построение новой онтологии человека в новой реальности, новой гибридной среде и восстановление нормы человека в этой новой реальности, в которой необходимо перестраивать интерфейс человека и ИИ. Поэтому проблема не в том, чтобы выработать очередной этический кодекс для сферы ИИ, а в том, чтобы всякий раз восстанавливать этос человека, норму человека.

Стремление же разработать кодекс выступает фактически следствием редукции человека, сведения его к отдельной особи, капризной и зависимой от умной техники. Чтобы умная техника не нанесла ему вред, вырабатывается система мер по защите слабого и беззащитного человека-индивида. Представление об атомарной личности, венце западной цивилизации, свелось к отдельной, зависимой от умной техники особи, отдельном индивиде, испытывающем соблазн от разного рода желаний и потребностей, которыми представлена окружающая жизнь, воспринимаемая им в виде большого и яркого гипермаркета.

По большому счёту эти кодексы защищают человека как обывателя жизни. А задача заключается в том, чтобы человек оставался субъектом, способным в себе вырабатывать систему защиты от капризов и соблазнов, связанных с потреблением умной техники. Строго говоря, именно человек, отказывающийся от бытия как нормы, и стремится защитить себя от придуманного им же самим интеллектуального монстра, ИИ.

Что касается второго вопроса, то здесь ситуация сложнее. С одной стороны, принимаются декларации, согласно которым вводится запрет на разработку

и внедрение определенных систем ИИ. Например, в Проекте Регламента Еврокомиссии от 2021 года введен так называемый класс запрещенных практик ИИ:

«Искусственный интеллект должен быть не самоцелью, а инструментом, который служит людям с конечной целью повышения их благосостояния. Правила для искусственного интеллекта, распространенного на рынке Евросоюза или влияющего на граждан другим образом, должны ставить людей в центр. Люди должны доверять использованию технологий, знать о ее безопасности и соответствии закону, включая уважение основных прав <…>.

Запрещаются следующие виды практики искусственного интеллекта:

a) размещение на рынке, ввод в эксплуатацию или использование системы искусственного интеллекта, которая применяет сублиминальные методы вне сознания человека с целью существенного искажения поведения человека таким образом, что причиняет или может причинить этому человеку или другому лицу физический или психологический вред;

b) размещение на рынке, ввод в эксплуатацию или использование системы искусственного интеллекта, которая использует любое из уязвимых мест определенной группы лиц в силу их возраста, физических или психических недостатков, с целью существенного искажения поведения лица, относящегося к этой группе, таким образом, что это лицо или другое лицо причиняет или может причинить физический или психологический вред;

c) размещение на рынке, ввод в эксплуатацию или использование систем ИИ государственными органами или от их имени для оценки или классификации благонадежности физических лиц в течение определенного периода времени на основе их социального поведения или известных или прогнозируемых личных или личностных характеристик <…>.

d) использование систем дистанционной биометрической идентификации "в реальном времени" в общедоступных местах в целях обеспечения правопорядка, за исключением случаев» [Proposal 2021][3].

С другой стороны, например, компания «Сбер», принимая свой Кодекс этики для сферы ИИ, введя принципы этики, явным образом показывает, что за этим стоит всего-навсего маркетинговая политика [Принципы этики ИИ 2021].

Можно упомянуть также Азиломарские принципы ИИ[4]. Но что здесь интересно? В основании этих принципов также вольно или невольно закралось уподобление поведения ИИ – поведению человека. Например, названы такие принципы:

«… Схожесть ценностей. Высоко автономные системы ИИ должны быть разработаны таким образом, чтобы их цели и поведение были схожи с человеческими ценностями на протяжении всей их работы.

---

[3] Proposal for a Regulation of the European Parliament and of the Council, Title II, article 5, p. 43-44. https://digital-strategy.ec.europa.eu/en/node/9756/printable/pdf

[4] Принципы были разработаны и приняты по итогам конференции разработчиков и исследователей в сфере ИИ, прошедшей в январе 2017 года в Азиломаре, США. На данный момент под ними поставили свои подписи свыше 3 500 ученых, разработчиков, предпринимателей и экспертов. Среди них Илон Маск, Стивен Хокинг, Рэй Курцвейл, представители Google, Apple, Facebook, IBM, Microsoft и т.д.

Человеческие ценности. Системы ИИ должны разрабатываться и работать таким образом, чтобы быть совместимыми с идеалами человеческого достоинства, его прав и свобод, многообразия культур.

… Общее (всеобщее) благо. Суперинтеллект должен разрабатываться только для служения широко разделяемым этическим идеалам и на благо всего человечества, а не одного государства или организации» [AI Principles 2017]

Можно также назвать недавно принятый в России Кодекс этики для сферы ИИ, принятый в 2021 году на форуме по этике ИИ [Кодекс 2021]. Согласно этому Кодексу *человеко-ориентированный и гуманистический* подход является основным этическим принципом и центральным критерием оценки этичного поведения акторов в сфере ИИ.

Его расшифровка звучит следующим образом: «При развитии технологий ИИ человек, его права и свободы должны рассматриваться как наивысшая ценность. Разрабатываемые Акторами технологии ИИ должны способствовать или не препятствовать реализации всех потенциальных возможностей человека для достижения гармонии в социальной, экономической, духовной сфере и наивысшего расцвета личности, учитывать ключевые ценности, такие как сохранение и развитие когнитивных способностей человека и его творческого потенциала; сохранение нравственных, духовных и культурных ценностей; содействие культурному и языковому многообразию, самобытности; сохранение традиций и устоев наций, народов, этносов и социальных групп» [Кодекс 2021].

Далее в Кодексе приводятся уже известные принципы этики для сферы ИИ (непричинение вреда, ответственность, безопасность, поднадзорность и т. д.).

Возникает вопрос: кто и как контролирует соблюдение названных принципов? Каковы механизмы контроля, мониторинга соблюдения принципов? Кто и как несёт ответственность за несоблюдение принципов? Эти вопросы пока остаются открытыми. В принятых декларациях и кодексах вопрос о механизмах контроля и наказания вообще фактически отсутствует.

Заключение

Учитывая сказанное, пока необходимо констатировать следующее.

При разработке этических кодексов и деклараций сам человек не рассматривается как источник риска. Источник рисков почему-то видится в ИИ, как будто он выступает субъектом действия. Сам же человек (и не только разработчик) не рассматривается как субъект изменений и источник рисков.

До сих пор сохраняется допущение сходства человека и машины, допускается аналогия и уподобление человека и ИИ.

В принятых декларациях и конвенциях присутствует больше маркетинга, нежели этики.

Предложенные декларации и кодексы для сферы ИИ не стали регулятивами, поскольку в них отсутствуют механизмы контроля и мониторинга соблюдения названных принципов и правил.

В разрабатываемых моделях ИИ доминирует допущение, что этический выбор совершается так же логически и рационально, как и решение математической задачи. А с этим связано и допущение, что ИИ можно обучить этическому поведению так же, как и научить решать задачи.

Возникает ощущение, что мы сами, люди, уже запутались в своих собственных допущениях и представлениях. Наверное, надо начинать вновь с чистого листа. И с самих себя.

Библиография

Принципы этики 2021 – *Принципы этики* искусственного интеллекта Сбера. URL: https://www.sberbank.com/ru/sustainability/principles-of-artificial-intelligence-ethics (дата обращения 20.08.2023).

Искусственный интеллект 2019 – *Искусственный интеллект.* Что стоит знать о наступающей эпохе разумных машин. Пер. с англ. О. Д. Сайфутдиновой. М.: Изд-во АСТ, 2019.

Карпов и др. 2018 – *Карпов В Э., Готовцев П. М., Ройзензон Г. В.* К вопросу об этике и системах искусственного интеллекта // Философия и общество. 2018. № 2. С. 84–105.

Кодекс 2021 – *Кодекс этики* в сфере искусственного интеллекта https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf (дата обращения 21.08.2023).

Разин 2019 – *Разин А. В.* Этика искусственного интеллекта // Философия и общество. 2019. № 1. С. 5 –73.

Смирнов 2022 – **Смирнов С. А.** Наше бесчеловечное будущее, или уловка трансгуманизма» // «Человек». 2022. Том 33 № 1. С. 61–79.

Смирнов 2023 – **Смирнов С. А.** Соблазн не быть, или Онтологические корни технологического аутсорсинга // Человек. 2023. Том 34. № 1. С. 28–50.

Эдмондс 2020 – *Эдмондс Д.* Убили бы вы толстяка? Задача о вагонетке: что такое хорошо и что такое плохо? М.: Изд-во Института Гайдара, 2020. 264 с

ЮНЕСКО 2021 – *Рекомендации ЮНЕСКО* об этических аспектах искусственного интеллекта. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus (дата обращения 20.08.2023).

AI Principles 2017 – *AI Principles.* URL: https://futureoflife.org/ai-principles/ (дата обращения 20.08.2023).

Bourget, Chalmers 2021 – *Bourget D., Chalmers D. J.* Philosophers on Philosophy: The 2020 PhilPapers Survey. 2021. URL: https://philarchive.org/archive/BOUPOP-3 (дата обращения 21.08.2023).

Bostrom 2011 – *Bostrom N., Yudkowsky E.* The Ethics of Artificial Intelligence Draft for Cambridge Handbook of Artificial Intelligence, eds. William Ramsey and Keith Frankish (Cambridge University Press, 2011. Pp. 1–20. URL: https://nickbostrom.com/ethics/artificial-intelligence.pdf (дата обращения 28.08.2023).

Moral Machine 2023 // *Moral Machine.* URL: https://www.moralmachine.net/ (дата обращения 28.08.2023).

Proposal 2021 – *Proposal* for a Regulation laying down harmonised rules on artificial intelligence. https://digital-strategy.ec.europa.eu/en/node/9756/printable/pdf (дата обращения 20.08.2023).

References

AI Principles 2017 – *AI Principles.* URL: https://futureoflife.org/ai-principles/ (accessed 20.08.2023). In Russian.

Artificial Intelligence 2019 – *Artificial Intelligence.* What you should know about the coming era of intelligent machines. Per. s Engl. O. D. Sayfutdinova. Moscow: AST Publishing House, 2019. In Russian.

Bostrom 2011 – *Bostrom N., Yudkowsky E.* The Ethics of Artificial Intelligence Draft for Cambridge Handbook of Artificial Intelligence, eds. William Ramsey and Keith Frankish (Cambridge University Press, 2011. Pp. 1–20. URL: https://nickbostrom.com/ethics/artificial-intelligence.pdf (accessed 28.08.2023).

Bourget, Chalmers 2021 – *Bourget D., Chalmers D. J.* Philosophers on Philosophy: The 2020 PhilPapers Survey. 2021. URL: https://philarchive.org/archive/BOUPOP-3 (accessed 21.08.2023).

Code 2021 – *Code of Ethics* in Artificial Intelligence https://ethics.a-ai.ru/assets/ethics_files/2023/05/12/Кодекс_этики_20_10_1.pdf (accessed 21.08.2023). In Russian.

Edmonds 2020 – *Edmonds D.* Would you kill a fat man? The task of the wagon: what is good and what is bad? Moscow: Gaidar Institute Publishing House, 2020. In Russian.

Karpov et al. 2018 – *Karpov V.E., Gotovtsev P. M., Roizenzon G. V.* To the question of ethics and artificial intelligence systems // Philosophy and Society. 2018. № 2. P. 84–105. In Russian.

Moral Machine 2023 // *Moral Machine.* URL: https://www.moralmachine.net/ (accessed 28.08.2023).

Principles of Ethics 2021 – *Sber's Principles* of Ethics for Artificial Intelligence. URL: https://www.sberbank.com/ru/sustainability/principles-of-artificial-intelligence-ethics (accessed 20.08.2023). In Russian.

Proposal 2021 – *Proposal* for a Regulation laying down harmonised rules on artificial intelligence. https://digital-strategy.ec.europa.eu/en/node/9756/printable/pdf (accessed 20.08.2023).

Razin 2019 – *Razin A. V.* Ethics of artificial intelligence // Philosophy and Society. 2019. № 1. P. 5–73. In Russian.

Smirnov 2022 – *Smirnov S. A.* Our inhuman future, or the trick of transhumanism" // "Man". 2022. Vol. 33. № 1. P. 61–79. In Russian.

Smirnov 2023 – *Smirnov S. A.* The temptation not to be, or Ontological roots of technological outsourcing // Man. 2023. Vol. 34. №. 1. P. 28–50. In Russian.

UNESCO 2021 – *UNESCO* Recommendations on the Ethical Aspects of Artificial Intelligence. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455_rus (accessed 20.08.2023).