УДК 004.8:17

МОРАЛЬ ЧЕЛОВЕКА И КОД ИСКУССТВЕННОГО ИНТЕЛЛЕКТА – В ЧЁМ РАЗНИЦА?

Филипенкова А. А., научный руководитель Ершова О. В.

РЭУ имени Г. В. Плеханова, Москва, e-mail: filipenkova@internet.ru, O.V.Ershova@yandex.ru

В данной работе рассматриваются проблемы, возникающие при попытках внедрения моральных принципов и установок поведения в искусственный интеллект. Даётся подробное определение морали. Рассматривается противоречивость применения морали в различных дилеммах, в частности, в дилемме по «делу Ричарда Паркера». Приводится неоднозначность определения морали по причине культурных различий, отдельно рассматриваются два вида – восточная и западная. Выясняются отличительные особенности данных двух видов. Даются характерные черты русской морали в сравнении с западной и восточной системами нравственных принципов. При этом выделяются особенности русской этики, не наблюдаемые в обеих странах, либо являющиеся комбинацией двух моделей одновременно, с учётом современных тенденций. На основании характерных черт вышеуказанных моделей морали приводятся гипотетические примеры использования морали человеком в реальной жизни. Описывается процесс написания кода искусственного интеллекта, принципы осуществления заданных алгоритмов. Приводится сравнение нравственного процесса и процесса исполнения кода. Даются возможные гипотезы о возможности соблюдения морали человека в результате работы искусственного интеллекта. Даётся вывод о невозможности абсолютного соответствия деятельности искусственного интеллекта нравственным принципам человечества.

Ключевые слова: этический кодекс искусственного интеллекта, необходимость морали в искусственном интеллекте, моральные дилеммы, противоречивость морали, культурные особенности морали

HUMAN MORALITY AND ARTIFICIAL INTELLIGENCE CODE – WHAT IS THE DIFFERENCE?

Filipenkova A. A., scientific supervisor Ershova O. V.

The Plekhanov Russian University of Economics, Moscow, e-mail: filipenkova@internet.ru, O.V.Ershova@yandex.ru

This paper examines the problems that arise when trying to introduce moral principles and behavioral guidelines into artificial intelligence. A detailed definition of morality is given. The inconsistency of the application of morality in various dilemmas, in particular, in the dilemma in the "Richard Parker case" is considered. The ambiguity in the definition of morality due to cultural differences is presented; two types are considered separately - Eastern and Western. The distinctive features of these two types are clarified. The characteristic features of Russian morality are given in comparison with Western and Eastern systems of moral principles. At the same time, features of Russian ethics are highlighted that are not observed in both countries, or that are a combination of two models at the same time, taking into account modern trends. Based on the characteristic features of the above models of morality, hypothetical examples of the use of morality by a person in real life are given. The process of writing artificial intelligence code and the principles of implementing given algorithms are described. A comparison is made between the moral process and the code execution process. Possible hypotheses are given about the possibility of complete observance of human morality as a result of the work of artificial intelligence. The conclusion is drawn about the impossibility of full compliance of the activities of artificial intelligence with the moral principles of humanity.

Keywords: ethical code of artificial intelligence, the need for morality in artificial intelligence, moral dilemmas, inconsistency of morality, cultural characteristics of morality

Определение грани между моралью и искусственным интеллектом (ИИ) необходимо для решения насущного актуального вопроса: нужно ли «привить» нравственность искусственному интеллекту, и если да, то как?

Нравственное поведение является основой жизни среднестатистической личности. И если искусственный интеллект внедряется в массы, то он должен учитывать и обстоятельства, в которой живет большая часть населения. Одной из важных частей окружающего мира для любого индивида является именно мораль. Значит, «вживлять» её необходимо. Однако для того, чтобы её «привить», необходимо разобраться, чем на самом деле она является.

Мораль — это нравственные нормы поведения, отношений с людьми, а также сама нравственность; логический, поучительный вывод из чего-нибудь; нравоучение, наставление [1]. Её изучением в свою очередь занимается такая наука как этика. И этика выделяет ряд сложностей в изучении морали, которые связаны: а). с её противоречивостью, б). с многообразием её определения.

Противоречивость морали ярко раскрывается в особых случаях, которые носят название дилемм. Одни из самых известных примеров дилемм, который действительно имел место быть — «дело Ричарда Паркера». Ричард Паркер — юнга, попавший на борт корабля Mignonette 1884 года. Корабль угодил в шторм, и капитан корабля Томас Дадли принял решение спустить шлюпку, в которой оказался он, Ричард Паркер и еще двое моряков — Брукс и Стивенс; однако в шлюпке оказалось мало провизии, и совсем не оказалось воды. Ричард Паркер вскоре заболел — начал лихорадить, и его кожа покрылась струпьями; члены команды были уверены, что юнга должен был скоро умереть. И спустя несколько дней, когда стало ясно, что Ричард Паркер не выживет, было принято решение его убить, чтобы спасти жизнь членов команды.

Как действует закон морали, когда можно спасти троих ценой жизни одного, находящегося уже на грани смерти? Что лучше — убить одного или умереть всем?

Может показаться, что этот пример далек от применения искусственного интеллекта, однако тот же самый выбор может встать не только во время рискованного морского путешествия: подобный выбор зачастую встречается в военном деле, сфере опасной промышленности, медицины и в зонах с повышенным риском технологических или природных катастроф.

Для того чтобы вживить мораль в искусственный интеллект, необходимо разобраться в двух проблемах: в содержательной части и технической. Содержательная часть подразумевает, какое именно наполнение выбрать для искусственного кода. Техническая часть касается того,

каким образом осуществляется процесс прописывания морали коду. Начинается рассмотрение темы с содержательной трудности.

Сложность в данном аспекте относится к многообразию значений морали: оно кроется в том, что многие культуры по-разному определяют критерии морали.

Мораль также зависит от морального кода и менталитета страны. При этом на Востоке страны по-разному трактуют основные положения нравственности, в зависимости от своего исторического опыта, что может быть связано с их историей и культурой. Поскольку восточные страны более изолированы друг от друга, чем западные, формирование их прошлого имело более обособленный и самоидентичный характер. Однако это послужило также и формированию одной яркой общей черты: так, в восточных странах главные принципы морали заключаются в добровольном и искреннем соблюдении долга перед обществом, в соблюдении добродетелей своей собственной страны. В то же время западные страны взаимосвязаны друг с другом, что позволило сформировать общие положения, которые могут охарактеризовать общую картину в целом для Запада, без деления его на части.

В качестве примера для рассмотрения ситуации на Востоке будет взята Япония, как одна из самых восточных стран в мире.

Мораль в Японии — до:току 道徳 буквально звучит как «путь добродетели». При этом мораль до:току — это не простое соблюдение правил, как на Западе. На Западе – это простое подчинение сводам законов, но в Японии это имеет более глубокий смысл— здесь мораль находится внутри человека, является его сердцевиной. Принципы японской морали: самоотречение и бескорыстие; взаимопомощь и готовность идти на компромисс; скромность и сдержанность; самообладание и добровольное желание отплатить за добро нации; моральный и нравственный долг (даже перед теми, кого ненавидят, но с кем работают вместе). Последний принцип, который также известен как гири-ниндзё — является основополагающим в японской культуре, поскольку исполнение долга перед другими людьми является полностью добровольным выбором и не может принуждаться другими; однако тот, кто не исполняет принцип гири-ниндзё, попадает под презрение общества. Важную роль в подобном деле исполняет добровольность принципов морали на традиционном пути. Мы можем увидеть, что многие моральные ценности ясно и точно сформулированы, имеют рациональное основание, но при этом направлены на духовную часть личности в коллективе; каждый член общества точно знает свои место и роль [2].

Что касается Запада? После опубликования работ Иммануила Канта ситуация на Западе сильно поменялась из стороны апологии в сторону критики морали. В том числе к числу критикующих относятся Ницше, у которого эта тема находится в центре философских

рассуждений, и Маркс, у которого это является второстепенной темой раскрытия. Однако человек, сильный волей, теперь считается аморальным, а сама мораль считается конструктом для слабых, их ответом на собственное поражение перед сильными. Западная модель направлена на увеличение эффективности индивида, на усиление личностных характеристик. Это означает, что многие моральные ценности рациональны, материальны, направлены на служение личности. Ранее эти практики заботы личности и самопознания: труд, поиск своего места в жизни, своего собственного «Я» должны были служить для поддержания государства, общественных интересов, однако постепенно они стали выполняться ради самого индивида, превратились в автономную самоцель [3].

Понимание морали в нашей стране находится на стыке западной и восточной. Исходя из своих исторических корней, она связана в первую очередь с православием и с иерархией – иерархией материального и духовного, Бога и человека. Многие нормы берутся в первую очередь из религии, и даже основные моральные нормы формируются под её довлеющим влиянием. Как следствие, мораль в современной России основана не на личности, как это видно на Западе, а на коллективе. В первую очередь — коллективное согласие, коллективный успех, совместное движение вперед, большое чувство общности. При этом ее другое главное отличие от западной и от восточной, уникальное понимание — сложность для рационального осмысления. Она намного сложнее, поскольку не поддается общему теоретическому обобщению и функциональному расчёту, как в других культурах, потому что она базируется на иррациональном, тесно связанным с религией и на истории, и иногда носит характер бессознательный (продолжение советской идеологии). Моральные ценности направлены на коллектив и на личность в коллективе, не всегда являются четко обозначенными, но являются духовными, но при этом данное описание больше подходит под классическое понимание отечественной морали — в современном мире всё более заметной становится тенденция склонения к морали, служащей личности [4].

Как же это разнообразие влияет на создание искусственного интеллекта? Чтобы продемонстрировать эту сложность содержательной части проблемы, потребуется привести пример, и в качестве наилучшего иллюстрирования сравнения будут выбраны машины Тесла.

Машины Тесла — машины, обладающие искусственным интеллектом. Благодаря этому искусственному интеллекту машина может принимать сложные решения без прямого вмешательства человека в процессе экстренных ситуаций. Одно из таких решений представляет сложный нравственный выбор — сбить ребенка, выбежавшего на дорогу? Или спасти жизнь ребенку, но пожертвовать собственником транспортного средства? Представим, что существует два вида машины Тесла. Одна под названием «Эгоист» будет нацелена на то,

чтобы обезопасить владельца машины, но сбить ребёнка. Другая, «Альтруист», будет спроектирована спасти ребёнка, но, вероятнее всего, пожертвовать владельцем автомобиля.

Если бы выдвигалось такое предположение, что люди своей культуры полностью перенимают её моральный портрет, то автомобиль «Альтруист» пользовался бы повышенным спросом в восточной культуре, а автомобиль «Эгоист» — в западной. Этот пример является тривиальным, поскольку нельзя рассчитывать на полное соответствие моральному портрету. Однако рассмотренные различия ясно показывают, что не существует одного эталона морали.

Ранее были рассмотрены трудности, связанные с содержательной частью. Теперь будет совершен переход к обсуждению технической проблемы.

Прежде чем от обсуждения морали переходить к обсуждению кода искусственного интеллекта, необходимо развеять некоторые смутные подозрения, вызванные самим термином «искусственный интеллект».

В русском языке обозначение этого феномена вызывает подозрение и волнение, и это может быть связанно с тем, что «интеллект» является понятием антропоморфным. Применение этого определения к машине даёт восприятие искусственного интеллекта, как прототипа человекоподобного существа.

Наиболее точное описание искусственного интеллекта складывается в английском языке. В определении «artificial intelligence» используется слово «intelligence», а не «intellect». И в первую очередь «intelligence» означает «способность мыслить разумно». Это означает, что первоочередной задачей искусственного интеллекта является выполнение поставленных целей, которые в него запрограммированы. При этом машина действует в рамках узкой спецификации: к примеру, если она нацелена на изучение истории философии, является невозможным заставить этот механизм решить задачу по высшей математике. Отсюда следует, что искусственный интеллект подразумевает несколько этапов: программирование, получение цели, сбор данных из внешнего мира, использование для конкретной цели, усовершенствование используемых техник на основании получаемых данных и разумных рассуждений и достижение целей.

Одно из частных определений интеллекта, общее для человека и «машины», можно сформулировать так: «Интеллект — способность системы создавать в ходе самообучения программы для решения задач определённого класса сложности и решать эти задачи» [5, с. 46].

Следовательно, самая главная сложность, связанная с программированием ИИ, заключается в постановке целей — цель задаёт человек, создатель машины. Создание новых целей за рамками спецификации невозможно. Создание новых спецификаций без вмешательства человека также невозможно. Далее «создание нового» и «творческий процесс»

будет подразумевать собой как раз ту область, которая закладывается в искусственный интеллект человеком — новые цели и спецификации.

Подводя итог вышесказанному, необходимо отметить, что мораль является результатом труда человека, она отличается многообразием и сложностью понимания. Изобретатель морали — живое существо, которое обладает креативностью мышления — способностью создавать новое и изменять уже существующие придуманные им правила — в частности, создавать новые цели и новые предметные области, не существовавшие ранее; ставить новые цели. Поэтому мораль имеет свойство меняться и добавлять вновь придуманные и сочинённые конструкты в новых целях и из новых намерений.

Искусственный интеллект — прописанный код. Код имеет только одну возможность — исполнять алгоритм — осуществлять заданные действия — достигать поставленной цели. Исполнять не то же самое, что и придумывать: создание чего-то нового подразумевает выход за рамки алгоритма. Выход за рамки является невозможным в процессе постоянного исполнения кода, которому подвержен искусственный интеллект. Следовательно, искусственный интеллект может осуществить заданные ему действия, соответствующие нормам морали, но не создавать новые.

Вопрос встаёт о самообучении искусственного интеллекта. Кто-то может возразить, сказав, что искусственный интеллект способен самостоятельно обучаться. Это действительно так, но это не влияет на способность к креативности. Искусственный интеллект, согласно прописанному ему коду, может запоминать всё более сложные конструкции самостоятельно. К примеру, ему можно загрузить возможность изучать японский язык самостоятельно, и постепенно он сможет самостоятельно охватывать все более сложные конструкции.

Однако у человека есть одно важное преимущество — умение находить ассоциации. До сих пор функция мозга, отвечающая за непрерывный и бесперебойный процесс нахождения ассоциаций не изучен до конца и вызывает множество споров. Причиной тому являются случайность образов и создание случайных комбинаций из пережитого опыта. Если человеку необходимо придумать новый конструкт в морали, ему необходимо задействовать все органы чувств, весь предыдущий накопленный опыт из различных сфер жизни — бытовой, культурной, религиозной и др. Благодаря умению находить соответствующие ассоциации из собственной памяти и на их основе создавать новое, не противоречащее уже существующему и содержащее элементы предыдущих знаний, необходимо понимание и соответствие абсолютно на всех уровнях.

Допустим, что будет создан искусственный интеллект, который, подобно ребёнку, накопил первоначальные знания о мире, а также обладает всеми органами чувств и гормональной системой. Предположим, что в его коде ассоциации будут выведены

коэффициенты корреляции (взаимозависимости) между возникающими случайными образами на основе случайной выборки. Однако в таком случае существует один важный недочёт системы: несмотря на высокую корреляцию, он может выдать образы, которые на самом деле не имеют ничего общего. Исходя из приведенных рассуждений, можно доказать корреляцию между людьми, любящими сыр, и количеством удушенных во сне. Наличие корреляции еще не говорит о наличии связи. В таком случае возникающие ассоциации будут мало связаны друг с другом в понимании человека. Будет нарушена логическая связь между ними. Из-за этого что-то новое, что может придумать искусственный интеллект, будет иметь мало практической пользы и ценности. Возможно, такой робот в конце концов и сможет создать что-то - к примеру, культ сыра — но это что-то будет оторвано от человеческого мира и не будет представлять явной и неоспоримой ценности [6].

Код искусственного интеллекта — математические формулы и расчёты. Однако не все математические процессы могут описать деятельность человеческого мозга. Возможно, искусственный интеллект превзойдёт человека, но только к тому моменту, как человек полностью познает себя и свой вид и сможет описать его с помощью цифр. Предположительно, если это и произойдёт, то человеку уже будет не так страшна идея захвата искусственным интеллектом человечества, как истина собственного существования.

Таким образом, главное отличие между этическим кодом человека и машины – в том, что человек задаёт основы морали, а машина исполняет чужие прихоти. Искусственный интеллект, который вполне может оказаться толчком к новой промышленной революции, остаётся инструментом в руках человека. Такая же история происходила с машинами «Дженни» и пароходами, сколько бы ужаса ни вызывало их появление. И на основе этого можно сделать вывод, что код искусственного интеллекта, в котором прописана мораль, — алгоритм по исполнению предписанных норм морали, математический расклад. Мораль вне кода, как результат мыслительной деятельности человека, — постоянно развивающийся и дополняющийся конструкт, обрастающий всё более новыми традициями, которые имеют сложную, неповторимую историю создания человеком.

Список литературы:

1. Даль В. И. и др. Толковый словарь русского языка. Эксмо, 2005.
2. Скворцова Е. Л. К вопросу о влиянии глобализации на процессы духовного развития в Японии // Вестник Института востоковедения РАН. 2021. Т. 4. №. 18. С. 67-74.
3. Черняк Н. А. «КУЛЬТУРА СЕБЯ»: К ИСТОРИИ СТАНОВЛЕНИЯ ЗАПАДНОГО СУБЪЕКТА // Вестник Омского университета. 2021. Т. 26. №. 3. С. 30-36.

4. Тарбеев Д. В. ИСТОРИКО-СОЦИАЛЬНЫЕ АСПЕКТЫ, ВЛИЯЮЩИЕ НА ЛИЧНОСТНЫЕ КАЧЕСТВА НА ПРИМЕРЕ ЗАПАДА, СССР И РОССИИ // Международный журнал гуманитарных и естественных наук. 2023. №. 3-1 (78). С. 137-144.
5. Ильясов Ф. Н. Разум искусственный и естественный // Известия АН Туркменской ССР, серия общественных наук. 1986. № 6. С. 46–54.
6. Васильев А. П., Абрамов А. Х. Искусственный интеллект на основе нейронных сетей // Academy. 2018. №. 5 (32). С. 15-17.