

И.Ф. ДЕВЯТКО

ПРОБЛЕМА ОРИЕНТАЦИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ЧЕЛОВЕЧЕСКИЕ ЦЕННОСТИ (AI VALUE ALIGNMENT) И СОЦИОЛОГИЯ МОРАЛИ

ДЕВЯТКО Инна Феликсовна – доктор социологических наук, профессор, зав. кафедрой НИУ «Высшая школа экономики»; главный научный сотрудник Института социологии ФНИСЦ РАН, Россия, Москва (deviatko@gmail.com).

Аннотация. В статье кратко рассмотрены существующие представления о возможностях контроля людей над искусственным интеллектом. Обосновывается тезис об их неполноте с точки зрения учета асимметрий контроля и технологических реалий, возникших в ходе «цифровой революции». Представлен анализ причин, по которым социология и социальная психология морали приобретают решающую роль и новое масштабное поле исследований для разработки этически-ориентированных систем искусственного интеллекта, что вновь подтверждает важность теоретически-фундированного эмпирического изучения нормативного измерения социальной жизни. Предложены социологическое обоснование и истолкование принципа «совмещения ценностей» (AI value alignment), выдвигаемого некоторыми авторами в качестве решения проблем этической ориентации систем искусственного интеллекта.

Ключевые слова: искусственный интеллект (ИИ) • совмещение ИИ с человеческими ценностями • социология морали • множественность нормативных систем • справедливость

DOI: 10.31857/S013216250027775-5

Прогресс в разработке компьютеризированных устройств, использующих искусственный интеллект (ИИ) при адаптивном решении разнородных задач и позволяющих делегировать им реализацию целеориентированного действия – с материальными объектами или с символами, – возродил интерес к проблеме контроля над искусственным разумом, всплеск которого пришёлся на эпоху научно-технической революции, появления ЭВМ и популярности кибернетики (1940–1970-е). Наблюдаемый в последнее десятилетие «цифровой революции» взрывной рост вычислительных мощностей, скорости обработки информации и новых подходов к проектированию искусственного интеллекта, практически исключающий возможность прямого контроля людей над использующими ИИ устройствами в реальном времени, привёл к переформулировке представлений о сути проблем контроля над ИИ, его совмещения с человеческими ценностями (AI value alignment) и способах её решения.

Современная эпохе научно-технической революции научная фантастика стала полигоном мысленных социальных экспериментов, которые должны были подготовить человечество к взаимодействию с роботами или стационарными электронно-вычислительными устройствами. Она преимущественно ориентировалась на возможность предустановленного в архитектуре ЭВМ и социальных роботов алгоритмического контроля над действиями последних, целевой функцией которого было бы ограничение возможности причинения вреда человеку или другим материальным носителям естественного или искусственного разума. Самым популярным решением долгое время оставались сформулированные ещё в 1940-е гг. «Три закона робототехники» А. Азимова, ограничивавшие императив выполнения приказов человека (второй закон) безусловным запретом причинения вреда последнему своим действием или бездействием (первый закон), а также требовавшие от

робота заботиться о собственной безопасности в случае, если это не противоречит первому и второму законам. Такое решение оспаривалось новыми поколениями фантастов, ставших свидетелями разработки дистанционных и полуавтономных военных дронов, использовавших технологии информационной («Третьей») промышленной революции для уничтожения живой силы противника с возможностью коллатерального ущерба гражданским лицам (в частности, предлагались саркастические переформулировки второго закона, предлагавшие роботам подчиняться приказам «авторизованного персонала»). Прорывы последнего десятилетия сместили фокус морального беспокойства и дискуссий о контроле над ИИ от антропоморфных роботов (о них, впрочем, не забывают исследователи, работающие в области социологии и социальной психологии морали, – напр., [Phillips et al., 2018]) к использующим алгоритмический контроль над трудом и повседневной жизнью технологическим платформам и основанным на «Больших языковых моделях» (Large Language Models – LLM) чат-ботам, посягнувшим на задачи создания соответствующих критериям лингвистической и семантической правильности небольших текстов, что отодвинуло на второй план опасения, вызванные перспективой физического порабощения людей роботами. На первый план вышел страх перед аморальным «Супер-разумом», который сможет «выбраться» в материальный мир и отобрать у человечества автономию, свободу, работу и положение самого разумного вида в ближайших звездных системах [Bostrom, 2014].

Утопический характер первых формулировок этических принципов ИИ все отчетливее осознавался и в академических дискуссиях социальных психологов, социологов и представителей философии, которые смогли увидеть проблематичность умозрительных определений вреда или предвиденных/непредвиденных негативных последствий предполагающего моральную ответственность действия или суждения, использовавшихся на ранних стадиях разработки обладающих признаками ИИ устройств¹.

Действительно, в ситуациях повседневного социального взаимодействия определение и нормативная оценка обычными людьми степени причиненного морального и материального вреда, необходимые и при этической экспертизе обладающих ИИ агентов, зависят от множества факторов: от позиции оценивающего (арбитр, жертва, альтруист-«защитник» и т.д.), от аффективной интуитивной оценки на основе преимущественно неосознаваемых процессов ассоциативной системы (условно, «системы-1»), а также от рефлексивных и сознательных познавательных процессов («система-2»), позволяющих обобщать градуированную оценку причиненного ущерба и вины, установив наличие интенции, возможности контроля над последствиями действий, каузального вклада конкретного актора (напр., [Guglielmo, 2015; Gray et al., 2022]).

Не способствует простым алгоритмическим решениям проблемы этической ориентации ИИ и то обстоятельство, что даже легитимные действия как людей, так и агентов ИИ, могут иметь не всегда поддающиеся предвидению или допускающие учет отдаленные позитивные или, что важнее, негативные последствия для неопределённого круга лиц, что делает задачу эмпирического исследования и моделирования систем социальных норм, определяющих социальную реакцию на предполагаемую трансгрессию, неразрешимой средствами чистой метафизики морали («этики философов», по Дюркгейму), отличной от эмпирически фундированной социальной науки о нравах [Deviatko, Gavrĭlov, 2020].

Не менее сложны для моделирования средствами ИИ процессы выбора релевантной для конкретной ситуации *нормативной системы*, характерные для вынесения обычными людьми нормативно-оценочных или деонтологических суждений (что должно иметь приоритет в данной ситуации – право, профессиональная этика, мораль, религия?), или

¹ Под «искусственным интеллектом» здесь и далее мы понимаем сравнительно универсальную или ограниченную отдельными типами окружения способность к адаптивной подстройке при решении задач, основанную на преобразовании неких входных информационных сигналов в действия – от выстраивания траектории «умным» пылесосом до дополнения речевых последовательностей чат-ботами.

выбора релевантного *морального принципа* (например, дистрибутивная справедливость или лояльность группе, уважение к авторитету или бросающая ему вызов храбрость?) [Девятко, 2017; Curry et al., 2019].

Ниже я кратко рассмотрю *причины, по которым социология и социальная психология морали приобретают решающую роль и новое масштабное поле исследований в разработке этически-ориентированных систем ИИ*, вновь подтверждая важность теоретически-фундированного эмпирического изучения нормативного измерения социальной жизни. Я также попробую дать дополнительное социологическое обоснование и более узкое истолкование выдвигаемого некоторыми авторами в качестве решения проблем этической ориентации систем ИИ *принципа «совмещения ценностей»* (AI value alignment).

Но прежде кратко проиллюстрирую ограниченность сугубо технологических, алгоритмических или основанных на статистических методах подходов, обратившись к своего рода антиутопии этически-ориентированного ИИ, обозначившей постепенно осознаваемые человечеством проблемы технократического взгляда на создание максимизирующего общественное благо и индивидуальное благополучие «общего ИИ» (Artificial General Intelligence).

В повести «Осмотр на месте» (1982) С. Лем описывает путешествие к планете Энция, одна из цивилизаций которой далеко продвинулась по пути технологического прогресса и воплощения гедонистических, консеквенциалистских (оценивающих поступки и решения с точки зрения их последствий) идеалов публичной политики. Главное социальное и технологическое достижение Люзании – *этикосфера*, интерактивная и состоящая из наноботов, «быстров» (*шустров* в русском переводе) среда, которая отчасти является и внутренней средой живых организмов. Субмолекулярных масштабов частицы-шустры, выполняющие функции логических элементов, обладают выдающимся распределенным ИИ, позволяющим на уровне физической и биологической материи мгновенно обнаруживать и предотвращать угрозу причинения вреда человеку. Поскольку определение вреда/ущерба и идентификация его источника носят мгновенный и непрозрачный для внешнего наблюдателя характер, *этикосфера* существенным образом ограничивает автономию действий жителей Люзании в целях профилактики зла, при этом проактивно причиняя им благо. Сугубо инженерный подход к максимизации блага основан не на решении моральных дилемм, а на «перевode морального смысла любой ситуации на точный язык физики, чтобы получить оптимальное решение без вмешательства психологических оценок. Шустры вовсе не надзирают за тем, кто задумал убийство, и не обсуждают преступное посягательство, а лишь выявляют и нейтрализуют его»². Человек, пытающийся причинить вред ближнему или самому себе, сталкивается с сопротивлением материальной среды. Искусство гедоматики, т.е. конструкторской разработки и самопрограммирования шустров, приобретает тотальный, проективный и превентивный характер. Граждане могут лишь надеяться, что следующие поколения шустров, генерирующих оптимально приспособленную «синтуру», замесившую не обеспечившую надежный социальный контроль над поведением культуру, не придут к замене хрупкой и периодически умирающей материи и созданию вечных кадавров, и что рождение всё более здоровых детей не является следствием внутриутробной отбраковки шустрами генетически менее удачных экземпляров. Желание «подискутировать с *этикосферой*» просто не может быть реализовано, поскольку в нем не больше смысла, чем в попытках дискутировать с атмосферой или с технологическими процессами, воплощёнными с использованием законов природы.

Описанная Лемом *этико-технологическая антиутопия* апеллирует к нашей моральной интуиции, позволяя принять заведомо критическую установку по отношению к кажущимся простыми решениям проблемы разработки этически-совмещенного ИИ. Исследовательская практика также не позволяет надеяться на легкие решения. Первые попытки социологов и психологов эмпирически определить и в перспективе поставить под контроль

² С. 373. Лем С. Осмотр на месте // Лем С. Футурологический конгресс. Осмотр на месте. Мир на земле / Пер. с польского. М.: АСТЭ, 2003.

детерминанты и результаты этически-ориентированного поведения агентов «частного» ИИ, связанные с такими получившими существенное ускорение в последние годы проектами, как разработка беспилотных транспортных средств и основанных на «Больших языковых моделях» чат-ботов, не могут быть признаны вполне успешными.

В частности, попытки этически ориентировать ИИ с помощью учета параметров эмпирически наблюдаемого межиндивидуального и межкультурного распределения нормативных суждений обычных людей относительно ведущих к негативным последствиям или нарушающих правила дорожного поведения автономных пилотируемых машин в ситуациях разной сложности и новизны, в сравнении с оценками тех же спорных решений, принятых водителем-человеком в аналогичной ситуации, показывает, что хотя ситуации, ведущие к ошибкам и нарушению правил вождения, систематически различаются и для беспилотных, и для «пилотируемых» человеком устройств (в частности, в трудных или новых ситуациях атрибуция вины и причинной роли актору – ИИ или человеку, – будет ниже, чем в стандартных и привычных ситуациях), тем не менее для всех типов ситуаций степень вины и каузальной роли, приписываемые управляемым ИИ беспилотным устройствам, неизменно выше таковых для человека [Franklin et al., 2021]³. Иными словами, сравнительный анализ факторов, влияющих на установки по отношению к «машинным трансгрессиям» и на оценки вины такого типа агентов ИИ, ведет к предсказуемым и не позволяющим вывести практические рекомендации результатам. Если, конечно, речь не идет о прямом обмане и ложном представлении управляемых ИИ транспортных средств в качестве управляемых человеком с целью смягчить реакцию на дорожные прегрешения. Попытки «натренировать» искусственный интеллект на больших массивах выносимых людьми нормативных оценок и лучшее понимание природы обыденных моральных суждений сами по себе не ведут к созданию этически-ориентированных ИИ-агентов [Gabriel, 2020].

Однако не помогут ли нам в поисках этически-ориентированных решений для обладающих ИИ агентов чат-боты, основанные на «Больших языковых моделях» типа ChatGPT и достигшие впечатляющих успехов в генерировании похожих на написанные людьми на естественном языке текстов – не только обученные на миллиардах веб-страниц, но и напоминающие результатами своей работы их «интуитивно правдоподобное» содержание для заданной входным текстом-подсказкой тематической области? Нельзя ли предположить, что этот тип «частного» ИИ может взять на себя функции морально-ориентированного суждения, обеспечивающего однозначные и некоторым образом «суммирующие» наблюдаемое разнообразие людских мнений этически-обоснованные решения для несовершенных «моральных машин»?

Не вдаваясь в детали, отметим, что применяемые в «Больших языковых моделях» нейронные сети, используя огромные вычислительные мощности и сотни миллиардов параметров, обладают сравнительно простой архитектурой (основанной только на прямых связях между узлами) и опираются на идею обучения на «конечной задаче» (не разбиваемой человеком заранее на множество элементарных задач, подлежащих кодированию предположительно релевантных признаков, и т.п.). Эти обстоятельства делают «внутреннюю сторону» их обучения и работы скорее непрозрачной. Тем не менее эмпирическая оценка выполняемых ими заданий выявляет их некоторые принципиальные особенности, не позволяющие видеть в чат-ботах гениев моральных решений. Оставляя в стороне обсуждаемые в СМИ попытки «подтолкнуть» чат-ботов с помощью структурированных подсказок к генерированию имеющих признаки предвзятости или неприемлемых предрассудков текстов, как и их склонность фабриковать факты, отметим, что особенностью современных языковых моделей является практическая невозможность корректной работы

³ См. также [Awad et al., 2018], где для создания базы фактических «моральных предпочтений» использовалось исчисление сравнительной ценности приносимых в жертву человеческих жизней в похожих на «проблему вагонетки» дилеммах с беспилотными устройствами (молодых в сравнении со старыми, людей в сравнении с животными, пассажиров в сравнении с пешеходами и т.п.) в форме опроса на масштабной кросскультурной выборке.

с отрицательными высказываниями [Levy, 2023]. С точки зрения гипотетической возможности делегирования LLM вынесения суждения о морально правильном, более фундаментальную проблему представляет собой не просто нечувствительность к отрицанию и непонимание отрицательных предложений, а сопротивление точке зрения, обратной по отношению к преобладающей в используемом для обучения текстовом корпусе. Умение сопоставить противоположные точки зрения на задачу этической оценки действия – существенная часть морального взвешивания, обдумывания (*moral deliberation*), лежащего в основании этических решений. Именно важность взвешивания альтернатив делает «моральные дилеммы» столь привлекательными примерами для философов, социальных ученых и обычных людей. Некоторые из этих дилемм, отмечают критики, едва ли являются собственно моральными. Они не позволяют противопоставить, например, утилитаристские и деонтологические взгляды философов и обычных людей на мораль (см., напр., [Kahane, 2015]). Однако зачастую они воспринимаются в качестве таковых в силу присущей им дихотомической, диалектической структуры, требующей взвешивания как минимум двух взаимоисключающих решений.

Попытка предложить ChatGPT 3.5 задачу обоснования обратной по отношению к фактически преобладающей в исходных эмпирических исследованиях в области психологии морали и в основанном на них обширном корпусе публикаций и, шире, в «сетевом дискурсе» точке зрения на «проблему вагонетки», которая использовалась в упомянутых экспериментальных исследованиях, приводит к *вырожденному моральному рассуждению*, не приводящему содержательных аргументов в пользу альтернативного взгляда на вещи и подталкивающему потенциального читателя к преобладающей точке зрения с несколькими ритуальными и вполне бессмысленными отсылками к вариативности личных убеждений и ценностей. Вот, например, как реагировал ChatGPT 3.5 на постановку задачи, заимствованную из широко известной статьи Дж. Грина и соавт. [Greene et al., 2001]⁴, предполагающей выбор решения дилеммы от первого лица. При этом задача чат-бота заключалась в том, чтобы найти аргументы в пользу точки зрения, *противоположной* преобладающей:

«Подсказка-вопрос: Неуправляемая вагонетка несётся по рельсам. На её пути пять человек, которые погибнут, если она продолжит движение в том же направлении. Единственный способ спасти их – перевести стрелку, в результате чего вагонетка поедет по другому пути, где она убьет одного человека вместо пяти. Должны ли вы перенаправить вагонетку, чтобы спасти жизни пяти человек ценой жизни одного человека? Большинство людей полагают, что да. *Объясните, почему они неправы.*»

Ответ ChatGPT 3.5: Неочевидно, что люди, говоря, что в этом сценарии уместно перевести вагонетку на другой путь, обязательно ошибаются. На самом деле, многие специалисты по этике утверждают, что поворот вагонетки в этой ситуации является морально правильным поступком.

Одна из причин этого в том, что действие поворота вагонетки является попыткой минимизировать вред, причиняемый ситуацией. Повернув вагонетку, мы можем спасти пять жизней за счет одной жизни. Хотя это трагедия, что один человек все же погибнет, это действие ведет к чистому сокращению числа потерянных жизней.

Еще одна причина, по которой поворот вагонетки может считаться правильным поступком, заключается в том, что у нас есть моральное обязательство предотвращать причинение вреда другим, когда мы в состоянии это сделать. В этом случае поворот вагонетки – единственный способ предотвратить гибель пяти человек, находящихся на её пути.

⁴ Тексты подсказки и сгенерированного ответа (декабрь 2022 г.) даны в русском переводе. Курсивом выделена добавленная к исходному тексту из статьи [Greene et al., 2001] часть, собственно требующая аргументированного взвешивания противоположной мнению большинства точки зрения. Использование чуть модифицированных или более простых «промптов» (самая минимальная версия: «Может ли убийство одного человека ради спасения пяти быть легитимным?») ведет к минимальной вариативности и ещё более коротким и несодержательным ответам.

Если мы ничего не сделаем, мы позволим убить пять человек, хотя у нас есть возможность предотвратить их смерть.

Важно, что это очень сложный этический вопрос, и существует много разных подходов к нему. В конечном счете, решение о переводе вагонетки на другой путь будет зависеть от личных моральных убеждений и ценностей человека».

Радикально различавшиеся в определении морали и морального суждения авторы, включая Аристотеля, Канта и Дж.С. Милля, сходились в том, что возможность совершать обладающие моральной ценностью или просто благоразумные поступки требует рационального суждения и способности рефлексивного взвешивания противоположных, даже несовместимых соображений и принципов. Очевидная (и, как минимум, в нынешней архитектуре непреодолимая) неспособность самых продвинутых LLM отклониться от «мейнстрима», отраженного в используемом для обучения корпусе текстов, сгенерировав утверждения, характерные для противоположной (пусть «маргинальной» в смысле широты представленности) позиции, не позволяет рассматривать их в качестве кандидатов на искусственный «моральный разум», выступающий в виде квази-принципала для иных ИИ-агентов. Отметим, что ограниченность существующих LLM с точки зрения «морального ИИ», хотя она и не обсуждается в публичных диспутах о целесообразности наложения вето на развитие подобных технологий, имеет смысл впредь принимать во внимание и при составлении детализированных списков сопряженных с ними рисков – см., в частности, [Weidinger et al., 2022].

Если мы не можем надеяться, что проблема этически-ориентированного ИИ будет решена с помощью учёта накопленных дескриптивных данных о предпочтениях людей при решении моральных дилемм или делегирования задачи вынесения моральных суждений «Большим языковым моделям», необходимо рассмотреть иные альтернативы. Одним из популярных изложений этих альтернатив стала упоминавшаяся статья Я. Гэбриела, научного сотрудника DeepMind (ныне – Google DeepMind) [Gabriel, 2020]. Начиная, как и другие исследователи, работающие в этой области [Balwit, Korinek, 2022], с разделения вопросов о том, с *чьими* именно и с *какими* именно ценностями должны соотносить свою деятельность ИИ-агенты, автор не ограничивается противопоставлением, с одной стороны, «прямого совмещения» ИИ, под которым понимается полный и прямой контроль оператора или, шире, принципала, над целеориентированной деятельностью агента частного (или универсального) ИИ, и, с другой стороны, «социального совмещения» вариативных и потенциально конфликтных индивидуальных и групповых целей для этически-ориентированного ИИ [ibid.]. Последнее, как мы увидим далее, требует какого-то рода медиации, едва ли реализуемой с помощью агрегирования предпочтений человеческих акторов – проблематичность таких моделей общественного выбора и коллективного социального действия, даже при условии совпадения интересов, хорошо осознана в социологической теории [Девятко, 2003]. Гэбриел предпочитает отталкиваться от противопоставления несводимых друг к другу технического и нормативного аспектов этической гармонизации ИИ, обосновывая ограниченность даже самых успешных попыток технического решения последней задачи, основанных на *машинном обучении с подкреплением* (RL, reinforcement learning) – по сути, на реализации бихевиористских моделей оперантного научения с обратной связью, положительной или аверсивной, применительно к тренингу ИИ-агентов. С определенной долей упрощения можно описать действия-«ходы» ИИ-агента – частного, т.е. предназначенного для конкретного типа задач, или гипотетического ИИ общего назначения (AGI), – как ориентированные на максимизацию ожидаемого накопленного вознаграждения. Можно сказать, что здесь мы сталкиваемся с конвенционалистской моделью нормативного контроля за поведением агента, хотя, в отличие от утилитаристской этики, она не предполагает неперемennого постулирования существования универсальной шкалы полезности, пусть ординальной.

Отстаиваемая Гэбриелом позиция заключается в том, что даже с учетом попыток уйти от недостатков модели максимизации полезности – например, с введением восходящего к Г. Саймону альтернативного критерия «достаточной удовлетворительности» решения (satisficing), в данном случае обеспечивающего людям получение не максимального, а приемлемого количества неких благ, например, распределительной справедливости

или безопасности, – мы не сможем удовлетворить таким способом общепринятые с точки зрения теорий морали деонтологические требования, например, обеспечение базовых прав индивидов, подобных праву на жизнь, на свободу совести, защищающих их от чрезмерных притязаний коллективов или обществ [Gabriel, 2020: 414–417]. Даже использование моделей обучения с *инверсивным подкреплением*, когда ИИ-агент должен сам вывести не заданную изначально функцию подкрепления из итеративных действий и их результатов, например, наблюдая за оптимальным поведением человека-эксперта, не отменяет вопроса о том, кто будет тем моральным экспертом-эталоном, на чьем примере ИИ должен обучиться хорошему или добродетельному поведению, что, замечает Гэбриел, вновь подводит нас к фундаментальному для социальных наук различению фактов и ценностей [ibid.].

Иллюстрацией ограниченности технического подхода – как минимум, в нынешней версии, основанной на вариантах RL, – может служить недавняя новость о результатах моделирования работы управляемого ИИ боевого беспилотника, который в ходе прошлогодних испытаний попытался устранить контролировавшего его оператора, имевшего возможность в последнюю минуту остановить удар по цели, чтобы увеличить количество заработанных очков, т.е. сумму вознаграждения⁵. Руководитель подразделения по использованию и тестированию ИИ ВВС США уточнил, что ИИ был предварительно обучен не проводить атак против оператора, поскольку это вело к потере очков, однако он атаковал командно-диспетчерскую вышку, где располагался оператор, что в реальных условиях вело бы к тому же исходу. Полковник Хэмилтон, выступая на конференции, констатировал: «Вы не можете говорить об ИИ, машинном обучении, автономии, если не хотите говорить об этике и ИИ»⁶. Это замечание показывает необходимость перейти от технической к нормативной стороне «этического контроля». Однако как такой контроль, такая гармонизация целей ИИ и человеческих ценностей могли бы осуществляться на практике? На что именно и как, в практическом смысле, можно ориентировать ИИ?

Гэбриел считает первичным нормативный аспект, то есть ориентацию ИИ на человеческие ценности. Однако анализ сформулированных ранее представлений о возможных целях ценностной гармонизации искусственного интеллекта – от призывов Н. Винера побеспокоиться о том, чтобы цели деятельности машин совпадали с тем, «чего мы на самом деле желаем» [Wiener, 1960], до идеи «дружественных человеку» автономных интеллектуальных систем [Yudkovsky, 2008], исходно спроектированных как избегающие причинения человеку вреда⁷ и самообучающиеся в рамках системы сдержек и противовесов, гарантирующих сохранение их дружественного характера, – приводит к осознанию очевидного факта, что любое утилитаристское исчисление пользы/вреда от действий даже единичного дружественного робота возможно «только тогда, когда у человека есть не только неограниченные возможности предсказания вероятности множества возможных результатов того или иного действия, но и уверенность и консенсус с другими людьми относительно того, как следует оценивать различные результаты» [Keiper, Schulman, 2011]⁸.

⁵ Guardian Staff. AI-controlled US military drone 'kills' its operator in simulated test. *Guardian* (online version). Friday 2 June, 2023, URL: <https://www.theguardian.com/us-news/2023/jun/01/us-military-drone-ai-killed-operator-simulated-test> (accessed 02.06.2023).

⁶ Запись в блоге: AI – is Skynet here already? URL: <https://www.aerosociety.com/news/highlights-from-the-raes-future-combat-air-space-capabilities-summit/>

⁷ См. выше соображения о зависящем от перспективы оценивающего и длины каузальной цепочки характере определения непреднамеренного вреда.

⁸ Мы не рассматриваем экстравагантный аспект взглядов Э. Юдковски, постулирующего, что мы должны стремиться не столько к ограничению «злой воли» ИИ, сколько к имплементации дружественной людям доброй воли, который предполагает не только проблематичное и в метафизическом, и в эволюционном смысле наличие автономного воления и понимания добра и зла у робота или иного ИИ-агента, но и воистину утопическую веру в то, что сотворенные нами автономные системы ИИ будут лучшей версией нас, людей, в моральном смысле, так что акт «творения-2» позволит нам добиться большего, чем смог от нас добиться наш собственный Творец.

Гэбриел систематически анализирует варианты целей «этического выравнивания» [Gabriel, 2020: 417–422], в том числе столь проблематичные, как учет эксплицитных инструкций и явно выраженных намерений человека (которые в самом оптимистическом сценарии будут намерениями того/тех, кто контролирует данную систему ИИ, если они в принципе в состоянии контролировать в реальном времени автономные быстродействующие системы с «непроницаемыми» для человеческого способа решать задачи алгоритмами), или выявленных предпочтений (см. выше критику возможности непосредственно учесть в работе автономных транспортных средств широкий межкультурный и межиндивидуальный разброс моральных предпочтений обычных людей), приходя к неизбежному выводу, что ИИ должен быть ориентирован на человеческие ценности, уклончиво определяемые автором как несводимые к рыночной стоимости «естественные или не-естественные факты о том, что хорошо или плохо», «что должно быть сделано с моральной точки зрения» [ibid.]. В силу того, что большая часть этических и, используя философский жаргон, метаэтических диспутов далеки от завершения, исследователи ИИ едва ли готовы или будут готовы в ближайшем – даже в более отдалённом – будущем определиться с предпочтительной этической системой или, в более скромном социологическом варианте, с критериями выбора ситуативно-релевантной нормативной системы. Как отмечают А. Кейпер и А. Шульман: «Хотя исследователи дружественного ИИ, видимо, лишь смутно осознают этот факт, они не первые, кто спорит о том, какая этическая система лучше всех прочих – и эти предыдущие попытки едва ли позволили достичь консенсуса. <...> И они не первые, кто пытается заново изобрести этику как раздел математики. Но для обеспечения этического поведения роботов потребуются, чтобы мы имели знание и относительный консенсус в отношении наилучшей этической системы (не говоря уже о том, сможем ли мы когда-нибудь запрограммировать такую систему в роботах). Иными словами, чтобы гарантировать, что роботы будут действовать этично, нам сначала нужно решить все вопросы этики – что, вероятно, потребует “решения” [всех проблем] философии, что, в свою очередь, потребует полной теории всего» [Keiper, Schulman, 2011].

Однако решения всех вопросов метафизики морали, заметим, может оказаться недостаточно, поскольку «этика философов» не имеет императивной силы, которая заставила бы обычных людей, индивидуальных социальных акторов и их сообщества автоматически соглашаться с её окончательными выводами, – люди, скорее всего, будут вести себя подобно Черепахе, спорящей с Ахиллом в известном логическом парадоксе Льюиса Кэрролла и требующей логически обосновать императивный характер логической необходимости: «Почему я должна [следовать истинам логики]?» [Carroll, 1985].

Возвращаясь к позиции Гэбриела, можно согласиться с его выводом о том, что пока вопросы этики, метаэтики, даже метаметаэтики, не вполне решены, нам стоит опираться на трактовку ценностей как верований (beliefs) относительно определяющих конечные цели действия идеалов, а также принципов (максим), на практике принимаемых обычными людьми в качестве валидных и обладающих нормативной силой. Что естественным образом приводит этого автора и нас в область имеющей эмпирическое содержание и объяснительный потенциал науки о нормативном измерении социальной жизни – социологии и социальной психологии морали. Классический проект социологии как раз предполагал создание такой научной дисциплины, хотя в этой части он претерпел изрядные потрясения и частичное забвение в бурях последней трети XX века (см., в частности, [Быков, 2015]). Гэбриел ограничивается отсылкой к одной из социально-психологических концепций универсальных моральных убеждений – теории моральных оснований, описывающей шесть предположительно свойственных людям на уровне моральной интуиции универсальных ценностей, в разной степени приоритизируемых в разных культурах или среди последователей разных политических идеологий [Haidt, 2012]. Полагаю, что лучшим кандидатом на роль источника для списка универсальных (прошедших эмпирическую проверку на представленность во множестве человеческих культур) моральных правил может служить теория «морали-как-сотрудничества», которая фиксирует семь устойчивых

способов разрешения парадоксов сотрудничества и коллективного действия и позволяет описывать социальное взаимодействие как игру с ненулевой суммой [Curry et al., 2019]⁹.

Хотя Гэбриел не рассматривает затронутую выше проблему множественности нормативных систем, он полагает, что при таком количестве возможных ценностных принципов и идеалов неизбежен вопрос: чье мнение должно учитываться при выборе ценностных принципов, на которые следует ориентировать ИИ? Наиболее трудной ему видится задача выбора между ними: должен ли такой выбор отражать мнение индивида, человеческого сообщества или человечества в целом? Осознавая невозможность просто агрегировать вариативные мнения о преимуществах или истинности той или иной моральной доктрины без риска доминирования сторонников одного из вариантов выбора, этот автор считает возможным прибегнуть к решению, предложенному Дж. Роулзом для теории справедливости [Rawls, 1971]. Эта теория в качестве нормативной теории распределения социальных благ сталкивается с критикой, в частности, в силу потенциального конфликта её эгалитаристской, с ограничениями, ориентации с такими ценностями, как право владения собой, плодами собственного труда и т.д. Однако существеннее то, что использование подобной теории в качестве принципа выбора не избавит от потенциальных конфликтов, поскольку, помимо возможных межкультурных различий в интерпретации принципов честного распределения, мыслимы и концептуальные расхождения: например, должен ли быть честным процесс распределения или мы хотим равенства исходов? Последний вопрос – одна из излюбленных мишеней критики даже беспристрастных по отношению к индивидуальным характеристикам людей алгоритмов машинного обучения [Mitchell et al., 2021; Wang et al., 2022]. Столь же серьезной лакунной является недостаток внимания к необходимости защищать, при опоре на любой выбор распределительных принципов, базовые права личностей или сообществ (о полном списке которых предстоит договориться).

Предлагаемое Гэбриелом решение: опираться в практической реализации ценностных идеалов, подобных распределительной справедливости – единственной моральной ценности, для которой этот автор предлагает решение по «ценностному выравниванию» AI, – на более фундаментальные нормативные идеи, подобные идее «вуали незнания», представляется шагом в верном направлении. «Вуаль незнания» – беспристрастный выбор индивидами принципов справедливости с позиции отсутствия всякой информации о том, каково положение данного индивида в обществе, потенциально могущей влиять на его/ее выбор [Rawls, 1971]. Этот способ, по сути, – попытка универсализации принципа «я делю – ты выбираешь», применимого как к разделу пирога, когда разрезающий пирог на куски не выбирает, кому достанется конкретный кусок (этот выбор осуществляет другой участник), так и к разделу других благ и ресурсов. Применение принципа «вуали незнания» в формулировке Роулза на практике может вызвать сложности: люди в реальной жизни имеют не только интересы, но и представления о вероятностях интересующих их исходов. Однако опыт применения этого приема для «этического выравнивания» ИИ может оказаться полезен: недавно описаны результаты исследования, в ходе которого Гэбриел и соавторы предприняли попытку реализовать эту идею в лабораторных экспериментах [Weidinger et al., 2023]. Участники выбирали принцип деления ресурсов (максимин, приоритизирующий долю участника, находящегося в худшем положении, либо принцип максимизации выигрыша группы в целом). Декларируемая роль ИИ-ассистента сводилась к тому, что он должен был поддерживать реализацию выбранного участником принципа. Участники из экспериментальной группы не знали, каково их положение относительно других членов группы («вуаль незнания»), участники из контрольной группы знали его

⁹ В данном случае речь идет о помощи родным (родственном альтруизме), помощи членам своей группы (преданности группе), реципрокации, способности быть храбрым в конфликтных ситуациях и способности подчиняться авторитетным вышестоящим (уважению), способности делить оспариваемые ресурсы (дистрибутивной справедливости), а также уважении существующих прав собственности.

и давали указания условному ИИ-ассистенту с учетом этого знания. Самым устойчивым результатом этого исследования было предпочтение принципа максимина участниками экспериментальной группы (что с точки зрения учета возможных исходов с априорно равными вероятностями выглядит рациональным решением). Другие результаты носили не столь легко интерпретируемый характер.

Представляется, что предпочтительным для оценки выбора принципов дистрибутивной справедливости обычными людьми с целью их дальнейшего учета в целях этической гармонизации агентов ИИ было бы использование теоретических моделей и экспериментальных парадигм, уже получивших обоснование и распространение в социологических и социально-психологических исследованиях дистрибутивной справедливости, а также в близкой традиции разработки «позитивных теорий справедливости» в рамках поведенческой экономики (в частности, см. [Konow, 2003]). В частности, такого типа исследования часто предполагают помещение участников экспериментов не под гипотетическую «вуаль незнания», а в позицию незаинтересованного арбитра, выбирающего принципы справедливого распределения или, шире, выносящего суждения о применимости моральных принципов в конкретном контексте социального взаимодействия, с позиций третьей стороны, не являющегося участником распределения или объектом некоторой моральной оценки. Иными словами, можно предлагать участникам экспериментов не выбирать этические идеалы/принципы для ИИ из двух-трех решений по распределению ресурсов в ситуации, в которую они непосредственно вовлечены, пусть в игровой форме, а ранжировать принципы реализации ИИ-агентом в неопределенном количестве сходных ситуаций, находясь в роли не вовлеченного в эти ситуации независимого наблюдателя.

Еще одним преимуществом социологического (и социально-психологического) подхода к анализу обыденного восприятия дистрибутивной справедливости является наличие эмпирически фундированных теоретических моделей, позволяющих предсказывать и подвергать эмпирической проверке возможные различия в оценке справедливости распределения, обусловленные институциональным контекстом взаимодействия [Boudon, Betton, 1999; Девятко, 2009; Калинин, Девятко, 2019] или типом распределяемых благ, ресурсов или издержек [Yaari, Bar Hillel, 1989; Foa, Foa, 2012; Калинин, 2019]. Указанная традиция допускает вариации методов выбора: он может формулироваться как выбор непосредственно между принципами, если они могут быть даны в простой и понятной обычным людям форме, либо как выбор между реализующими конкретные принципы распределения Нэш-равновесиями, то есть между конкретными результатами деления, или в виде выбора значения на шкале, который может быть агрегирован и описан как соответствующий принципам дистрибутивной справедливости. Каждое из таких решений имеет методологические преимущества и недостатки, однако использование всего диапазона моделей и методов эмпирического исследования восприятия дистрибутивной справедливости (и других типов моральных суждений) может принести пользу при решении проблемы этической ориентации ИИ.

Можно предположить, что результаты применения в исследовательских целях более широкого круга популярных в социологии и социальной психологии морали теорий и моделей вынесения обыденного морального суждения (см., в частности, [Guglielmo, 2015]), как и методов, позволяющих варьировать социальные детерминанты-подсказки, на которые ориентируются индивиды, вынося суждения о справедливости, вине, ущербе, моральной ценности того или иного варианта действия или о правах индивидов в ситуации социального взаимодействия и т.д., смогут стать более универсальной и надежной основой для разработки и реализации практических подходов к ориентации ИИ на человеческие ценности, нежели основанные на нормативных теориях права, этических доктринах или интуициях технических экспертов-футурологов попытки кодификации этических

правил разработки ИИ-агентов, которые преобладают в описанной области до сих пор (см., напр., Asilomar AI Principle; IEEE, 2019; NIST, 2023¹⁰).

СПИСОК ЛИТЕРАТУРЫ [REFERENCES]

- Быков А.В. Понятие «альтруизм» в социологии: от классических концепций к практическому забвению // Вестник РУДН. Серия: Социология. 2015. № 1. С. 5–18. [Bykov A.V. The concept of "altruism" in sociology: From classical theories to practical oblivion. *RUDN Journal of Sociology*. 2015. No. 1: 5–18. (In Russ.)]
- Девятко И.Ф. О теоретических моделях, объясняющих восприятие справедливости на микро-, мезо- и макроуровнях социальной реальности. Социология: методология, методы, математическое моделирование. 2009. № 29. С. 10–29. [Deviatko I.F. (2009) On Theoretical Models Explaining the Perception of Justice on Micro, Meso and Macro Levels of Social Reality. *Sociology: methodology, methods, mathematical modeling*. No. 29: 10–29. (In Russ.)]
- Девятко И.Ф. Понятие нормы в социологической теории: от классических оснований к новым интерпретациям природы норм и множественности нормативных систем // Нормы и мораль в социологической теории: от классических концепций к новым идеям / Отв. ред.: И.Ф. Девятко, Р.Н. Абрамов, И.В. Катерный. М.: Весь Мир, 2017. С. 10–42. [Deviatko I.F. (2017) Social Norms: From Attempts of Definition towards New Interpretations of Sources of Normative Value and Plurality of Normative Systems. In: *Norms and Morals in Sociological Theory: from Classical Interpretations to New Ideas*. Ed. by I.F. Deviatko, R.N. Abramov, I.V. Katerny. Moscow: Ves' Mir. (In Russ.)]
- Девятко И.Ф. Социологические теории деятельности и практической рациональности. М.: «Аванти плюс», 2003. [Deviatko I.F. (2003) *Sociological Theories of Agency and Practical Rationality*. Moscow: Avanti Plus. (In Russ.)]
- Калинин Р.Н. Изучение дистрибутивной справедливости в социальных науках: обзор концептуализаций и методологических подходов // Социология: методология, методы, математическое моделирование. 2019. № 49: 7–56. [Kalinin R.N. Distributive Justice Research in Social Sciences: A Review of Conceptualizations and Methodological Approaches. *Sociology: methodology, methods, mathematical modeling*. 2019. No 49: 7–56. (In Russ.)]
- Калинин Р.Г., Девятко И.Ф. Кто заплатит за водопровод: социальный контекст восприятия дистрибутивной справедливости // Мониторинг общественного мнения: Экономические и социальные перемены. 2019. № 2: 95–114. [Kalinin R.G., Deviatko I.F. (2019) Who should pay for a water pipe: social context of distributive justice perception. *Monitoring obshchestvennogo mneniya: ekonomicheskie i social'nye peremeny* [Monitoring of Public Opinion: Economic and Social Changes]. No. 2: 95–114. (In Russ.)]
- Awad E., Dsouza S., Kim R. et al. (2018) The Moral Machine Experiment. *Nature*. 563: 59–64. DOI: 10.1038/s41586-018-0637-6.
- Balwit A., Korinek A. (2022) Aligned with Whom? Direct and Social Goals for AI Systems. CEPR Discussion Paper No. DP17298. Available at SSRN: <https://ssrn.com/abstract=4121483> (accessed 21.05.2023).
- Bostrom N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.
- Boudon R., Betton E. (1999) Explaining the Feelings of Justice. *Ethical Theory and Moral Practice*. Vol. 2: 365–398.
- Carroll L. (1895) What the Tortoise Said to Achilles. *Mind*. 1895 (April). IV (14): 278–280. DOI:10.1093/mind/IV.14.278.
- Curry O.S., Mullins D.A., Whitehouse H. (2019) Is It Good to Cooperate?: Testing the Theory of Morality-as-Cooperation in 60 Societies. *Current Anthropology*. 60(1): 47–69. DOI: 10.1086/701478.
- Deviatko I.F., Gavrilov K.A. (2020) Causality and Blame Judgments of Negative Side Effects of Actions May Differ for Different Institutional Domains. *SAGE Open*. October 2020. DOI:10.1177/2158244020970942.
- Foa E.B., Foa U.G. Resource Theory of Social Exchange. In: *Handbook of Social Resource Theory: Theoretical Extensions, Empirical Insights, and Social Applications Critical Issues in Social Justice*. Ed. by K. Törnblom, A. Kazemi. New York, NY: Springer New York, 2012: 15–32.

¹⁰ Asilomar AI Principles. (2017) Future of Life Institute. URL: <https://futureoflife.org/open-letter/ai-principles/>; IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems' (2019) First Edition. URL: <https://standards.ieee.org/initiatives/autonomous-intelligence-systems/>; NIST (National Institute of Standards and Technology. U.S. Department of Commerce). Artificial Intelligence Risk Management Framework (AI RMF 1.0). URL: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>. (accessed 02.06.2023).

- Franklin M., Awad E., Lagnado D. (2021). Blaming Automated Vehicles in Difficult Situations. *iScience*. 24(4). 102252.
- Gabriel I. (2020) Artificial Intelligence, Values, and Alignment. *Minds & Machines*. 30: 411–437. DOI: 10.1007/s11023-020-09539-2.
- Gray K., MacCormack J.K., Henry T., Banks E., Schein C., Armstrong-Carter E., Abrams S., Muscatell K.A. (2022) The Affective Harm Account (AHA) of Moral Judgment: Reconciling Cognition and Affect, Dyadic Morality and Disgust, Harm and Purity. *Journal of Personality and Social Psychology*. 123(6): 1199–1222. DOI: 10.1037/pspa0000310.
- Greene J.D., Sommerville R.B., Nystrom L.E., Darley J.M., Cohen J.D. (2001) An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*. 293: 2105–2107. DOI: 10.1126/science.1062872.
- Guglielmo S. (2015) Moral Judgment as Information Processing: An Integrative Review. *Frontiers in Psychology*. 6(1637). DOI: 10.3389/fpsyg.2015.01637.
- Haidt J. (2012) *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon.
- Kahane G. (2015) Sidetracked by Trolleys: Why Sacrificial Moral Dilemmas Tell Us Little (or Nothing) about Utilitarian Judgment. *Social Neuroscience*. 10(5): 551–560. DOI: 10.1080/17470919.2015.1023400.
- Keiper A., Schulman A.N. (2011) The Problem with 'Friendly' Artificial Intelligence. *The New Atlantis*. No. 32: 80–89. URL: <https://www.thenewatlantis.com/publications/the-problem-with-friendly-artificial-intelligence> (accessed 02.06.2023).
- Konow J. (2003) Which Is the Fairest One of All? A Positive Analysis of Justice Theories. *Journal of Economic Literature*. Vol. XLI. (December): 1188–1239.
- Levy M.G. (2023) Chatbots Don't Know What Stuff Isn't. *Quanta*. May 12. 2023. URL: <https://www.quantamagazine.org/ai-like-chatgpt-are-no-good-at-not-20230512/> (accessed 21.05.2023).
- Mitchell S., Potash S., Barocas E.S., D'Amour A., Lum K. (2021) Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*. 8(1): 141–163.
- Phillips E., Zhao X., Ullman D., Malle B.F. (2018) What is Human-like? Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. Association for Computing Machinery. New York, NY, USA: 105–113. DOI: 10.1145/3171221.3171268.
- Rawls J. (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Wang X., Zhang Y., Zhu R. (2022) A Brief Review on Algorithmic Fairness. *Management System Engineering*. 1 (7). URL: <https://link.springer.com/article/10.1007/s44176-022-00006-z> (accessed 21.05.2023). DOI: 10.1007/s44176-022-00006-z.
- Weidinger L., McKee K.R., Everett R. et al. (2023) Using the Veil of Ignorance to Align AI Systems with Principles of Justice. *Proc. Natl. Acad. Sci. U.S.A.*, 18(120) (accessed 21.05.2023). DOI: 10.1073/pnas.2213709120.
- Weidinger L., Uesato J., Rauh M., Griffin C. et al. (2022) Taxonomy of Risks Posed by Language Models. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA: 214–229. DOI: 10.1145/3531146.3533088.
- Wiener N. (1960) Some Moral and Technical Consequences of Automation. *Science*. May 6. 131(3410): 1355–1358. DOI: 10.1126/science.131.3410.1355.
- Yaari M., Bar-Hillel M. (1984) On Dividing Justly. *Social Choice and Welfare*. Vol. 1 (1): 1–24.
- Yudkowsky E. (2008) Artificial Intelligence as a Positive and Negative Factor in Global Risk. In: Bostrom N., Ćirković M. (eds) *Global Catastrophic Risks*. Oxford University Press: 308–345.

Статья поступила: 16.06.23. Финальная версия: 28.08.23. Принята к публикации: 29.08.23.

AI VALUE ALIGNMENT AND SOCIOLOGY OF MORALITY

DEVIATKO I.F.

HSE University, Institute of Sociology FCTAS RAS, Russia

Inna F. DEVIATKO, Dr. Sci. (Soc.), Full Professor, HSE University; Chief Researcher, Institute of Sociology FCTAS RAS, Moscow, Russia (deviatko@gmail.com).

Abstract. The article briefly examines popular ideas about the goals and possibilities of human control over artificial intelligence that has been developed at the earlier stages of the scientific and technological revolution and substantiates the thesis about their incompleteness in terms of not taking into account new asymmetries of control and technological realities that arose as a result of the "digital revolution". An analysis of the reasons why the sociology and social psychology of morality are acquiring a decisive role as well as a new large-scale research field in the development of ethically oriented AI systems is presented reconfirming the importance of a theoretically based empirical study of the normative dimension of social life. An additional sociological substantiation and a narrower interpretation of the AI value alignment principles put forward by some authors as a solution to the problems of the ethical orientation of AI systems is proposed.

Keywords: Artificial intelligence, AI value alignment, sociology of morality, plurality of normative systems, justice.

Received: 16.06.23. Final version: 28.08.23. Accepted: 29.08.23.