Этическая мысль
2024. Т. 24. № 1. С. 87–100
УДК 17.0

Ethical Thought
2024, Vol. 24, No. 1, pp. 87–100
DOI: 10.21146/2074-4870-2024-24-1-87-100

ЭТИКА И ПРОБЛЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А.Г. Игнатьев

Этико-философские проблемы проектирования искусственного морального агента

Игнатьев Андрей Геннадьевич – аспирант. Национальный исследовательский университет «Высшая школа экономики». Российская Федерация, 105066, г. Москва, ул. Старая Басманная, д. 21/4, стр. 1

ORCID: 0000-0003-4256-0850
e-mail: a.ignatev@hse.ru

В статье рассмотрены объективные предпосылки и основания для создания искусственного морального агента[1] (ИМА), предложены аргументы как в пользу развития подобных проектов, так и демонстрирующие ограниченные возможности систем искусственного интеллекта (ИИ) в обладании субъектностью, необходимой для принятия моральных решений. Представлены подходы к понятию «искусственный моральный агент», рассмотрены некоторые концепции т.н. «машинной морали»[2], обозначены изначальные условия/критерии, предъявляемые к ИМА. Проблема ИМА прослеживается во взаимосвязи с теориями в области аналитической этики и философии техники. На основании такого анализа предлагаются направления исследовательской и научно-экспериментальной работы по созданию ИМА. Обосновывается тезис о несостоятельности технической реализации универсального, человекоподобного ИМА со способностью решать сложные, разноплановые моральные дилеммы и задачи, действовать в широком и разнообразном поле моральных ситуаций. В связи с этим предлагается использовать новые понятия – *вычислительный моральный квазиагент* – МКА (*computational moral quasi-agent*), а когда решения принимаются на основе человеко-машинного взаимодействия – *конвергентный, или гибридный моральный агент* – ГМА

[1] Англ. – artificial moral agent (AMA).

[2] Слова «машина», «машинный» (и др. однокоренные) – здесь и далее употребляются в значении «интеллектуальные компьютерные (информационно-вычислительные, программные) системы на основе технологии искусственного интеллекта».

(*hybrid moral agent*). *В* условиях недостаточной разработанности теоретической базы и наличия рисков внедрения ИМА предлагается сбалансированное, поэтапное развитие проектов на основе *предметно-ориентированного, контекстуального подхода*. Такой подход предполагает ответственное внедрение МКА применительно к отдельным классам систем (моделям), с конкретной прикладной задачей, в определенной физической или виртуальной среде. Целесообразность проектирования МКА или ГМА в каждом конкретном проекте должна быть оправдана, аргументирована и доказана.

Ключевые слова: искусственный интеллект, искусственный моральный агент, моральное агентство, этика в области искусственного интеллекта, сознание, разум, интеллект, субъектность, вычислительный моральный квазиагент, конвергентный (гибридный) моральный агент

Я могу объяснить многое, но не могу
объяснить, как работает мой мозг.
Никола Тесла

ИМА в матрице исследовательских проблем

Культура человека, возникшая из примитивных форм внутриродовой жизни и возделывания простейшей сельхозпродукции, подошла к рубежу создания систем, способных имитировать человекоподобные мыслительные операции и воспроизводить отдельные возможности человеческого разума или отдельные функции мышления. В процессе реализации новых амбициозных задач в этой области особую актуальность приобретает вопрос о способности подобных систем самостоятельно принимать моральные решения, при этом важно понять, чем обусловлены (детерминированы) эти решения, насколько эти решения будут приемлемы для человека и общества.

Сама идея создания *искусственного морального агента* (ИМА) и его актуальность, как правило, объясняется тем, что контроль морально-нравственных решений машин на всех этапах операций со стороны человека становится все более сложно осуществимым. Огромная размерность вычислений, объем всевозможных корреляций и скорость операций в системах ИИ не всегда позволяет человеку проверить, насколько то или иное решение, действие, прогноз или рекомендация системы соответствует этическим нормам или морально-нравственным ожиданиям индивидуума, группы, или, применительно к масштабным проектам, общества в целом. Исходя из этого, возникает соблазн частично или полностью переложить на машины функции морального агента, т.е. спроектировать автономного ИМА.

В мире активно ведутся практические разработки и теоретические исследования, связанные с конструированием ИМА на основе технологии ИИ.

В качестве попытки «компьютеризации» процесса принятия этических решений можно привести проект *MedEthEx* (*Medical Ethics Expert*) – прототип медицинского этического помощника, где используется подход Дж. Ролза к рефлексивному равновесию [Inusah, 2022] (в рамках проекта оптимизируется

процедура принятия решений при лечении больного). На основе этой модели применяется и компьютерная программа обучения медицинской этике – *MedEthEx Online* [Fleetwood, 2000].

Разнообразные теоретические концепции и гипотезы (т.н. «машинная этика», «машинная мораль», «искусственная или компьютерная мораль», «роботика» и т.п.) предлагались в рамках создания «когнитивного компьютера» [Brachman, 2002], «этического искусственного агента» [Anderson, 2007], «морального агентства в роботах» [Johansson, 2011; Podschwadek, 2017], а также в различных статьях и научных докладах, например: [Fossa, 2018; Behdadi, 2020; Martinho, 2021].

Тем не менее пока теоретическая платформа для создания ИМА выглядит достаточно мозаично, не вполне убедительно с позиций философии, нейронауки, инженерной практики, социологии, психологии и других дисциплин. Возникает определенный дисбаланс: с одной стороны, проблема ИМА требует теоретического осмысления, а с другой – избыточная и в ряде случаев оторванная от практики теоретизация отстает от быстро развивающихся систем ИИ, которые выдают решения, затрагивающие сферу морали. В этом смысле назрела необходимость выделить области, в которых ИМА – оправданная и неотложная задача, и области, где он пока не применим. Это позволит конкретизировать направления для приложения усилий.

Проблема ИМА оживила дискуссию об источниках и генезисе естественной человеческой морали (например, в рамках изучения натурализации морали, в нейронауке, техноэтике), породила новые вопросы в сфере философии сознания, потребовала прояснить взаимосвязи с такими понятиями, как «свобода воли», «справедливость», «совесть», сформулировать критерии для *автономности* систем ИИ. Создание ИМА пересекается и с задачами обеспечения доверия к системам ИИ (*trustworthiness*), снижения различных видов смещенности[3] в системах ИИ (*mitigation of bias*).

При проектировании ИМА одним из ключевых является вопрос – имитировать ли в машине человеческую способность к моральным суждениям и поступкам или следует искать иные, идеальные и универсальные механизмы и алгоритмы, разрабатывать теории, ориентированные исключительно на сложные компьютерные системы, т.е. рассматривать моральные аспекты в рамке «кибернетики третьего порядка» [Лепский, 2019] и цифровой техносферы.

В контексте разработки ИМА нельзя обойти вопросы «субъектности ИИ» (ИИ как субъект правовых отношений, в т.ч. применительно к вопросам интеллектуальных прав, регистрации патентов, возложения юридической ответственности и др.). В философии Нового времени, с характерным для него субъектоцентризмом, субъект – это прежде всего «носитель деятельности, сознания и познания» [Лекторский, 2001, 659]. Однако проблема перенесения сознания на компьютерную основу остается нерешенной, системы ИИ

[3] Смещенность (bias): систематическое различие в обработке определенных объектов, людей или групп по сравнению с другими (в технике); в когнитивной сфере используется термин предвзятость (ISO/IEC TR 24027:2021 Bias in AI systems and AI aided decision making).

не способны к самоидентификации и постановке исходных мета-целей – это аргументы, которые не позволяют говорить о создании универсального ИМА.

Становится очевидным, что проблема ИМА является настолько многоуровневой, что современная дискуссия, претерпевая разнонаправленное содержательное расширение, не находит единой опорной точки, вокруг которой можно было бы выстраивать непротиворечивые концепции. Поднимается огромный пласт теоретических и научно-практических проблем, в т.ч. затрагивающих сущность самой морали, природу человеческого разума и границы возможностей ИИ. Кроме того, возникает необходимость согласовать целый ряд понятий и терминов, поскольку новые технологические возможности требуют расширения не только инженерно-технического, но и философского словаря. Так, например, единое понимание термина «искусственный моральный агент» до сих пор отсутствует.

Таким образом, проблема моральных решений в сфере ИИ выводит нас в парадигму «сложностности» [Аршинов, 2016] и приводит к следующим наблюдениям:

– сложность и незаконченность постижения морали и ее проявлений в человеческом поведении затрудняют проектирование ИМА для неоднородной биосоциотехнической среды (отсутствие общеприемлемых моральных норм и нравственных принципов жизни самих людей; нет безупречно работающих теорий о том, что движет нравственными решениями людей; моральные ценности и принципы общества нередко расходятся с индивидуальными убеждениями индивидуума; моральные поступки человека порой самопроизвольны, спонтанны, не могут быть универсально обусловлены и алгоритмизированы);

– наблюдается отставание человечества в осмыслении и оценке всего комплекса социогуманитарного воздействия технологий ИИ на человека, общество и биосферу; без доказательного анализа такого воздействия разрабатывать моральные регуляторы для техники и конструировать ИМА представляется затруднительным (иллюстрацией могут служить системы на основе *GPT* и *LLM*, влияние которых в долгосрочной перспективе на процессы человеческого познания и на формирование представлений о мире, особенно у детей младшего школьного возраста, пока недостаточно изучено);

– на уровне технического проектирования систем ИИ существуют трудности программной загрузки в киберфизическую систему моральных правил, инструкций, различных этических установок и теорий, и полноценная «моральная прошивка» или «моральная алгоритмизация» для широкого поля задач и «моральной автономности» систем пока невозможна.

Методологические подходы

Междисциплинарность современного научного поиска сделала дискуссию вокруг ИМА еще более неоднозначной и затруднила выработку общеприемлемых и согласованных концепций. То, что раньше главным образом было уделом философов (вопросы эволюции морали, мораль и разум, свобода воли, вопросы субъектности и др.), теперь стало предметом для построения гипотез

представителями различных научных дисциплин и направлений. Разнообразие методологических приемов, используемых в различных дисциплинарных направлениях, помимо позитивного эффекта, дает противоречивые, а порой и взаимоисключающие результаты. Появляются, например, мнения о необходимости отхода от философско-теоретических объяснений феномена ИМА и рассмотрении исключительно нормативных аспектов использования ИМА (например, в публикации [Behdadi, 2020]). Это возвращает нас к известной проблеме противопоставления подходов к морали как к моральному долженствованию (прескриптивный, предписывающий подход, основанный на санкционированных общественных принципах, установках и нормах) и морали как эмоционально-чувственной, эмоционально-волевой или социально-обусловленной первопричины нравственных поступков, где решающую роль имеют индивидуальное моральное сознание, когнитивные предпосылки, эмоции и социальное окружение.

Одним из возможных решений для минимизации или устранения такого противоречия является, например, подход О.Г. Дробницкого, согласно которому «внешне взаимоисключающие положения перестают быть таковыми, когда выявляется их отношение к разным уровням теоретической абстракции (в диалектической логике – к различным ступеням восхождения от абстрактного к конкретному)» [Дробницкий, 2002, 523]. Это может быть применено к возникающему каркасу теоретических построений и многомерных усложнений в исследовании ИМА, когда сложно систематизируемая человекомерная мораль помещается в контур возможностей ИИ (рассмотрение этических аспектов в цифровой среде не может быть прямым «трафаретом» человеческой морали, но и не может ее не учитывать). Отсюда следует, что на самом начальном этапе проектирования ИМА создатели должны соединить теоретические морально-нравственные установки и цели с техническими особенностями конкретной модели ИИ в конкретной среде, предусмотреть все существующие и потенциальные воздействия системы на внешнюю среду, разобрать всю сложную парадигму внешнего влияния системы, исследовать изоморфизм законов, управляющих функционированием таких сложных объектов, как системы ИИ.

Таким образом, для всестороннего рассмотрения ИМА мы естественным путем приходим к методическим подходам, характерным для «Общей теории систем», функционирования и развития сложных систем (К.Л. фон Берталанфи, А.А. Богданов), синергетики (Г. Хакен, И. Пригожин, С.П. Курдюмов), системо-мыследеятельностной методологии (Г.П. Щедровицкий), методологии постнеклассической науки (В.С. Степин, Н.Н. Моисеев), саморазвивающихся рефлексивно-активных сред и кибернетики третьего порядка (В.Е. Лепский), энактивизма (Ф. Варела, Э. Томпсон, А. Ноэ, Е.Н. Князева), в рамках которого «сознание и его функции рассматриваются в контексте понимания сложности живого и природы сложных формообразований в мире» [Князева, 2013], теорий автопоэтических систем (У. Матурана, Ф. Варела, Н. Луман).

Вероятнее всего, дальнейшие попытки разработки ИМА методологически целесообразно вести на основе сочетания обозначенных выше и других подходов, которые позволят рассмотреть новые и порой парадоксальные проекции, возникающие в спектре «человек – техника». Любой линейный подход к ИМА,

претендующий на универсальность и всеприменимость, не может быть приложен к столь сложным по своему разнообразию и масштабам инженерно-техническим комплексам/экосистемам на основе ИИ. Проектирование ИМА неизбежно потребует анализа особенностей различных систем ИИ, классификации и типологии моделей, задач, уровня риска, среды использования, оценки нежелательной смещенности и т.д. Все эти вопросы могут быть решены путем всесторонней оценки (в т.ч. выработки соответствующих методик тестирования и измерения параметров для каждого типа системы). Внедрение же неких «универсальных» и якобы «человекоподобных», автономно действующих ИМА (вне контроля со стороны человека) в мультифункциональные цифровые экосистемы, которые будут влиять на сферу управления, социальную политику государственных институтов, на развитие медиапространства и СМИ, на формирование «молодых умов» в образовательных процессах, на работу телемедицинских систем и т.п., представляется безответственным и незрелым решением.

Критерии морального агента. Искусственный (вычислительный) моральный квазиагент и гибридный моральный агент

Проблема создания ИМА требует ясного представления о самом предмете исследований – «*моральной агентности*» и «*моральном агенте*» как субъекте морально-релевантных действий (иногда используется и термин «моральный субъект»). Как уже отмечалось, важнейшим условием моральной агентности считают наличие сознания [Gray, 2012], что обусловливает способность к намеренным моральным решениям и действиям, или, другими словами, к осознанному моральному выбору [Taylor, 2003; Bjornsson, 2020]. При этом моральный агент «воспринимается как интенциональный и ответственный» [Нарьян, 2022]. Об общей теории *интенциональной* структуры рационального человеческого действия в рамках философии сознания также рассуждает и Дж. Серль [Searle, 1988, 1995]. У некоторых авторов ключевую характеристику приобретают вопросы *ответственности* морального агента [Логинов, 2021].

В рамках понимания морального агента через роль сознания важное место отводится также способности испытывать эмоции и чувства (как первопричина, мотив и исходная почва для моральных переживаний, суждений и действий). В канве данных идей Ж. Делез, например, придавал важное значение аффектам и пристрастиям, которые управляют сознанием, это позволило ему сделать предположение о сосуществовании двух систем – рассудка и страстей, на основе которых формируются моральные теории [Делез, 2001].

Мы можем выделить следующие общие, наиболее важные требования к моральному агенту (для целей данной статьи, но далеко не исчерпывающие):

а) сознание (у человека имеется связь с эмоционально-чувственной сферой, в т.ч. за счет «телесности»);

б) обладание моральными представлениями (убеждениями) и способность совершать на их основе осознанные моральные поступки и действия.

Это, в свою очередь, предполагает и самоидентификацию агента (осознание себя и своей идентичности в окружающем мире), а также понимание собственных целей.

В результате прояснения сущности морального агента как такового становится очевидно, что сегодня даже самые совершенные системы ИИ не соответствуют этим критериям. Не так просто подобрать формулу и рецепты для интеграции моральных концепций, выработанных задолго до появления интеллектуальных компьютерных машин, с феноменом ИИ. Приведенные выше критерии морального агента не дают нам оснований приписывать даже высокосложным системам ИИ возможности, характеристики и «полномочия» целостного морального агента. Вместе с тем на данном этапе развития техники человек уже не в состоянии осуществить полноценный и всеобъемлющий контроль за системами, которые в процессе эксплуатации (в т.ч. в режиме онлайн) так или иначе оперируют в моральном поле и выдают решения, которые влияют на моральные аспекты человеческой жизни, отчасти даже воздействуют на моральную культуру человека. Такие системы в ряде случаев осуществляют операции вне зоны контроля/управления человеком, что позволяет рассматривать их де-факто как некую условную форму морального агента.

Такое положение вещей требует ввести некоторые уточнения в терминологический аппарат для этики в сфере ИИ:

1. Предлагается использовать более точную и корректную формулировку: обозначать ИМА как *вычислительный моральный квазиагент* (МКА) или как (*квази*)*моральный техноагент*. Предложенный термин изначально лишает такой объект всей полноты свойств и возможностей, присущих моральному агенту в привычной, общеупотребимой коннотации и предполагает определенные допущения, указывающие на его неполноценность и ограниченность (редуцированность).

2. На данном этапе развития техники и теоретической базы в области ИМА необходимо обеспечить в определенных интеллектуальных системах (особенно в зонах повышенной моральной ответственности) механизм контроля и сопровождения человеком процесса принятия решений машинами. Ответственные моральные действия в этом случае совершаются/управляются человеком, при этом вычислительные алгоритмы помогают выявить всю полноту корреляций и начальных условий моральной парадигмы и выдают свои промежуточные рекомендации. В этом случае удобно обозначать такие модели, как *конвергентный, или гибридный моральный агент* (ГМА). Такой подход соответствует более широкой концепции *коэволюции человека и техники*, в частности, теории развития *коэволюционного гибридного интеллекта* (*co-evolutionary hybrid intelligence*) [Krinkin, 2023].

3. Следует полностью исключить передачу моральных решений и моральной ответственности машинам при выполнении задач, где морально-нравственные аспекты имеют решающую, жизненно-важную роль для общества или индивида («красная зона»). Например, при рассмотрении определенных дел в судах, при принятии важных медицинских решений, в области применения опасных технологий, обороны и т.д. Это требование является важным аргументом для рассмотрения моральной агентности в тесной взаимосвязи

с классификацией практических кейсов и решаемых функциональных задач, порученных системам ИИ. Такая классификация может в т.ч. разрабатываться и совершенствоваться в рамках прикладной этики и отраслевых этических кодексов в сфере ИИ [Kuleshov, 2020].

Основные подходы к созданию ИМА

За последний десяток лет появилось большое количество научных публикаций, которые содержали попытки подойти к решению отдельных проблем ИМА, в т.ч. и в философском ракурсе. Наиболее распространенными в современных научно-технических статьях являются следующие подходы (указаны лишь наиболее характерные публикации, которые отражают ключевые направления дискурса) [Allen, 2005; Brundage, 2014; Farmosa, 2020; Tolmeijer, 2020]:

А. "Top-down" («сверху вниз» или «основанный на нормах»). Моральные установки, различные принципы, теории и стандарты закладываются в алгоритм системы на программном и инженерно-техническом уровнях таким образом, чтобы поведение системы (выходные результаты) не противоречили или соответствовали изначально определенным условиям или инструкциям.

Как представляется, практическая реализация данного подхода сталкивается с трудностями выбора и конкретизации предписывающих норм, которые не могут эффективно использоваться в различных средах и удовлетворять ожидания разных групп людей или сообществ. Применение в качестве таких норм библейских Заповедей Ветхого Завета, кантианской деонтологии, утилитаризма или других философских или нормативных теорий вызывает большие сомнения: как широкие, концептуальные нормы и принципы (выраженные в естественном языке) можно полноценно трансформировать в программные коды и команды? Процесс «кодирования» теоретических установок в программные алгоритмы и равноценная передача смыслов может представить значительные технические проблемы. Даже если предположить, что это возможно, остается вопрос о том, пригодны ли такие «правила» для разнообразной и меняющейся среды, всевозможных ситуаций и контекста использования систем ИИ (например, в различных культурных, религиозных или социальных средах, в ситуациях сложного морального выбора, в условиях этических дилемм и т.п.). Стоит учитывать и противоречивость (несогласованность) различных теоретических норм и принципов. Это уже подтверждено в ходе изучения этических оснований развития систем ИИ.

Б. "Bottom-up" («снизу вверх» или «метод обучения»). Постепенное целенаправленное обучение «агентов» на конкретных ситуациях, задачах или примерах для достижения определенного «автономного» морального поведения, т.е. способности принимать моральные решения.

Совершенствовать моральные решения машин путем эксперимента, проб и ошибок, поэтапной «настройки» алгоритмов – это путь, в основе которого лежит опыт, приобретаемый самим человеком, который, с детских лет познавая мир, формирует свои нравственные ориентиры и суждения. Уязвимым местом такого подхода является неспособность машин воспринимать мир

и социальную коммуникацию во всем многообразии (например, через сложные личные переживания), использовать при принятии моральных решений опыт, интуицию и эвристику, присущую человеку. С этим, вероятно, будут не согласны разработчики специальных обучающих платформ, которые имитируют различные жизненные ситуации. Действительно, сегодня наблюдаются успехи в создании социально-адаптированных компьютерных систем, обладающих отдельными элементами искусственной психики и способностью к определенному «восприятию» моральных ситуаций (для создания подобных платформ используется теория игр, различного рода симуляторы, методы ассоциативного обучения, моделирование эволюции в компьютерных системах, «моральные экосистемы» и т.п.). В качестве отдельного направления выделяются практики «воспитания» машин в виртуальной среде, в т.ч. создание виртуальных экспериментальных пространств для тестирования социальных процессов, социально-нравственной жизни и взаимодействия интеллектуальных искусственных объектов, включая моделирование моральных ситуаций. Тем не менее возможности таких машин пока далеки от имитации или воспроизводства морально-нравственной природы человека, которая складывается из чувственно-эмоциональных переживаний, разносторонней социальной коммуникации, жизненного опыта, впечатлений, колоссального объема импульсов и восприятий живой среды, поступающих через тело и органы чувств. Очевидно, что исследование морали в биологическом пространстве и реальном социуме существенно отличается от экспериментов с виртуальными моделями. Заманчивые попытки создать полную виртуальную копию мира, «оцифровать» само бытие, спроектировать некий искусственный полигон для моральной эволюции машин пока не находят убедительного технического воплощения.

В. Смешанный подход (сочетание “Top-down” и “Bottom-up”).

С одной стороны, различные сложности и противоречия подходов А и Б могут быть частично нивелированы за счет гибкого синтеза двух стратегий. С другой стороны, вероятно, в ряде случаев смешанный подход будет усиливать конфликтность способов реализации ИМА. В целом же достаточно перспективным представляется сценарий, в котором интеллектуальные системы, построенные на определенных, априорно заложенных правилах и нормах, смогут «доучиваться» и совершенствоваться на практических задачах.

Сценарии дальнейшего развития ИМА

Идеализируя будущие возможности разработчиков и гипотетически допуская развитие интеллектуальных машин до уровня, близкого к сильному ИИ[4], основные сценарии развития ИМА укладываются в рамки двух возможных направлений.

Первое из них – создание «человекоподобных», максимально антропоморфных систем ИИ, в т.ч., по Ф. Шолле, с высокой «сложностью обобщения (generalization difficulty)» [Chollet, 2019]. Это потребует новых технических

[4] Artificial general intelligence (AGI)

прорывов, новых открытий в нейрофизиологии, а также создания некоего симулякра или имитатора «воплощенного» человеческого тела, или квазибиологического субстрата, который добавит системе «квазителесность», более сложную «квазипсихику», создаст дополнительные способы интеркоммуникации и обучения систем в реальном, физическом мире, позволит производить полезные измерения существенных параметров системы, в т.ч. уровня искусственного «интеллекта». При этом, однако, даже появление систем такого уровня едва ли разрешит неоднозначные проблемы человеческой морали, не остановит поиск справедливости и социальной гармонии. Это не создаст и тот уровень доверия, который позволит бесконфликтно применять такие системы, особенно в зонах высокого риска, где моральные решения могут влиять на жизненно важные сферы человеческого бытия. Если эволюция человеческой морали и нравственности не избавила цивилизацию от несправедливости, несчастий, страданий и различных видов неравенства, едва ли машины, созданные «по образу и подобию» человека в состоянии решить эти проблемы. Появление некой удовлетворяющей всех искусственно созданной, универсальной, человекоподобной «технической моральной конструкции» в мире, далеком от совершенства, где не достигнута моральная гармония людей, выглядит неубедительно.

Второе направление, и оно представляется еще менее реалистичным, предполагает создание некоего абсолютного, превосходящего человеческий ИМА, со свойствами «морального арбитра», который гипотетически может разрешать сложные моральные противоречия и служить неким надчеловеческим ориентиром в поисках высшей формы нравственности и справедливости. Такой ИМА и соответствующие ему вычислительные процессы должны будут работать (извлекать моральные решения) на неких иных, близких к идеальным основаниях, эти процессы должны будут отличаться от того, как это делают люди, делать это лучше людей и к тому же пользоваться полным доверием общества. Подобные сценарии, как правило, входят в общую концепцию создания сверхразума [Boden, 2006; Bostrom, 2016]. В частности, предполагается, что, совершенствуя алгоритмы для автономных машин нового поколения, возможно создать и искусственную мораль, которая не будет ограничена рамками человеческого ума и мировоззрения, с присущими человеку догмами, предубеждениями, субъективизмом и различными метафизическими концепциями, порожденными человеческой сущностью. Однако такая модель предполагает, по сути, радикальное переосмысление достижений естественного разума, «мета-цифровизацию морали» до сверх- или надчеловеческих параметров и идеалов.

Продолжая данные рассуждения, мы приходим к тому, что человек, осознавая свое несовершенство, якобы способен создать нечто более идеальное, в данном случае, применительно к области морали, – создать ИМА, превосходящий существующие в человеческом обществе моральные регуляторы и системы поведения. Здесь возникают по меньшей мере две проблемы. Кто сможет провести объективную оценку такому ИМА, измерить его выходные параметры и дать «лицензию» на совершенство? Способен ли человек, находясь в своем измерении мира, искусственно произвести нечто большее по размеру

и значению, чем он сам (особенно в сфере морали, которая, по сути, лежит в сердцевине человеческой сущности и природы, которая является именно продуктом разума и культуры человеческой цивилизации)? Если вопрос об эффективности утилитарно-функционального, вычислительно-прикладного применения интеллектуальных машин в деятельности человека можно отнести к дискуссионным, а в ряде ситуаций признать превосходство именно вычислительных способностей ИИ, то в области морали и нравственности гипотезы конструирования близкой к абсолютному идеалу «моральной машины» выглядят неубедительными. В проблеме ИМА и в целом в изучении возможностей человека и машины, вероятно, можно проследить определенный «провал в объяснении» [Микиртумов, 2017], т.е. несводимости внешнего описания реакций человека (например, снятия доступных параметров головного мозга) и «внутреннего описания актов сознания» в рамках философской интуиции.

Приведенный выше анализ не исключает научно обоснованного применения некоторых ИМА (в скорректированной нами формулировке – МКА) определенной контекстуальной модальности для систем ИИ со строго определенными прикладными функциями и в конкретной среде эксплуатации. Вполне оправданным представляется также дальнейший поиск возможностей более широкого внедрения МКА для решения строго определенных задач, где максимально снижены риски и экспериментально подтвержден позитивный эффект подобного внедрения.

Выводы

1. С учетом разнообразия концепций и теорий в моральной философии, философии сознания, подходов к проблеме субъектности, а также, принимая во внимание ограниченные возможности машин (не являются исходными носителями моральных убеждений, целей и интенций), на данном этапе реализация полноценного, универсального ИМА, способного самостоятельно действовать в неструктурированной физической или виртуальной среде и применительно ко всему спектру самых различных моральных аспектов и ситуаций, представляется нереальной.

Вместе с тем для некоторых систем ИИ, которые оперируют в моральном измерении, для удобства понимания предлагается ввести и использовать новые термины, подразумевающие, что такие системы условно могут выполнять функции *вычислительного морального квазиагента* (МКА) или *гибридного морального агента* (ГМА). Однако такие объекты не обладают всем комплексом характеристик и внутренней природой полноценного *морального агента*.

Действительно, системы ИИ в той или иной степени влияют на итоговые моральные решения (либо на основе исходных предписаний человека, либо в пределах определенной технической *автономности*, которая заложена в модели. В ряде случаев такая автономность является непрозрачной и необъяснимой или же способ обработки системой человеческих инструкций в области морали не поддается объяснениям). Однако этого недостаточно, чтобы говорить о *моральной агентности* таких систем.

2. Для продвижения теоретических и прикладных исследований в области МКА/ГМА целесообразно применять *предметно-ориентированный, контекстуальный метод*. При таком методе разработка и проектирование моделей ИИ реализуется с учетом конкретной прикладной задачи и конкретной операционной среды. Определенной системе в определенных условиях эксплуатации может передаваться лишь такая часть полномочий, которая гарантирует безрисковое участие МКА в морально-значимых решениях. Экспериментальное внедрение МКА/ГМА целесообразно начинать в областях, где решения/действия машин не способны нанести моральный ущерб. Для этого необходима обоснованная классификация прикладных задач и соответствующих рисков применения МКА. При этом даже такое участие машин в морально-значимых решениях должно быть обратимым, т.е. предусмотрены возможности их пересмотра человеком.

Ethical and Philosophical Problems of Designing Artificial Moral Agent

Andrei G. Ignatev

School of Philosophy and Cultural Studies, HSE University. 21/4, p. 1 Staraya Basmannaya Str., Moscow, 105066, Russian Federation.

ORCID: 0000-0003-4256-0850
e-mail: a.ignatev@hse.ru

The article considers objective preconditions and reasons for the creation of an *artificial moral agent* (AMA), offers arguments both in favour of the development of such projects and demonstrating the limited capabilities of artificial intelligence systems in possessing the subjectness necessary for making moral decisions. First of all, disclosing the problem in terms of moral philosophy, approaches to the notion of "artificial moral agent" are presented, some concepts that hypothetically could provide the realisation of the so-called "machine morality" are discussed, and the initial conditions/criteria for AMA are outlined. The IMA problem is traced in relationship to theories in the fields of analytical ethics and philosophy of technology. On the basis of such an analysis the directions of research and scientific-experimental work on the creation of AMA are proposed. The thesis is justified about the failure of technical realisation of a universal, human-like AMA with the ability to solve complex, diverse moral dilemmas and tasks, to act in a wide and diverse field of moral situations. In this regard, it is proposed to use new terms – *computational moral quasi-agent* (MQA), and when decisions are made on the basis of human-machine interaction – *convergent or hybrid moral agent* (HMA). In the conditions of insufficiently developed theoretical basis and the presence of risks of MQA implementation, the author proposes a balanced, phased development of projects based on a *subject-oriented, contextual approach* (SOCA). Such an approach implies responsible implementation of MQA for specific types of systems (models) with a specific application task, in a specific physical or virtual environment (domain). In this case, the feasibility of designing MQA or HMA in a particular project should be explained, justified and proved.

Keywords: artificial intelligence, artificial moral agent, moral agency, ethics in artificial intelligence, consciousness, mind, intelligence, subjectness, computational moral quasi-agent, convergent hybrid moral agent

Литература / References

Аршинов В.И., Буданов В.Г. Парадигма сложностности и социогуманитарные проекции конвергентных технологий // Вопросы философии. 2016. № 1. С. 59–70.

Arshinov, V.I., Budanov, V.G. “Paradigma slozhnostnosti i sociogumanitarnye proekcii konvergentnyh tehnologij” [Paradigm of Complexity and Socio-humanitarian Projection of Convergent Technologies], *Voprosy filosofii*, 2016, No. 1, pp. 59–70. (In Russian)

Делёз Ж. «Эмпиризм и субъективность: опыт о человеческой природе по Юму» // Критическая философия Канта: учение о способностях. Бергсонизм. Спиноза / Пер. с фр. Я.И. Свирского. М.: ПЭР СЭ, 2001.

Deleuze, G. “Empirizm i sub”ektivnost’: opyt o chelovecheskoj prirode po Yumu” [Empirisism and Subjectivity. An Essay on Hume’s Theory of Human Nature], in: G. Deleuze, *Kriticheskaya filosofiya Kanta: uchenie o sposobnostyah. Bergsonizm. Spinoza* [Kant’s Critical Philosophy: the Teaching of the Faculties. Bergsonism], transl. by Ya.I. Svirskii. Moscow: PER SE Publ., 2001. (In Russian)

Дробницкий О.Г. Моральная философия: Избр. труды / Сост. Р.Г. Апресян*. М.: Гардарики, 2002.

Drobnickii, O.G. *Moral’naya filosofiya: Izbr. Trudy* [Moral Philosophy: Selected Works], ed. by. R.G. Apresyan. Moscow: Gardariki Publ., 2002. (In Russian)

Лекторский В.А. Субъект // Новая философская энциклопедия / Под ред. В.С. Степина, А.А. Гусейнова, Г.Ю. Семигина, А.П. Огурцова. Т. 3. М.: Мысль, 2001. С. 659–660.

Lektorskii, V.A. “Sub’ekt” [Agent], *Novaya filosofskaya ehntsiklopediya* [New Philosophical Encyclopedia], eds. V.S. Stepin, A.A. Guseynov, G.Yu. Semigin, A.P. Ogurtzov, Vol. 3. Moscow: Mysl’ Publ., 2001, pp. 659–660.

Лепский В.Е. Вызовы будущего и кибернетика третьего порядка // Проектирование будущего. Проблемы цифровой реальности: труды 2-й Межд. конф., 2019, Москва. М.: ИПМ им. М.В. Келдыша, 2019. С. 64–70.

Lepskiy, V.E. “Vyzovy buduschego i kibernetika tret’ego poryadka” [Challenges of the Future and Third-order Cybernetics], *Proceedings of the Second International Conference “Futurity Designing. Digital Reality Problems”*. Moscow: Keldysh Institute of Applied Mathematics, 2019, pp. 64–70. (In Russian)

Князева Е.Н. Энактивизм: концептуальный поворот в эпистемологии // Вопросы философии. 2013. № 10. С. 91–104.

Knyazeva, E.N. “Enaktivizm: konceptual’nyj povorot v epistemologii” [Enactivism: A Conceptual Turn in Epistemology], *Voprosy filosofii*, 2013, No. 10, pp. 91–104. (In Russian)

Логинов Е.В., Гаврилов М.В., Мерцалов А.В., Юнусов А.Т. Этика и метафизика моральной ответственности // Этич. мысль / Ethical Thought. 2021. Т. 21. № 2. С. 5–17.

Loginov, E.V., Gavrilov, M.V., Mercalov, A.V., Yunusov, A.T. “Etika i metafizika moral’noj otvetstvennosti” [Ethics and Metaphysics of Moral Responsibility], *Eticheskaya mysl’ / Ethical Thought*, 2021, Vol. 21, No. 2. pp. 5–17. (In Russian)

Микиртумов И.Б. Философская логика: трилемма номологии, «социологии» и «физиологии» // Философские науки. 2017. № 1. С. 85–94.

Mikirtumov, I.B. “Filosofskaya logika: trilemma nomologii, ‘sociologii’ i ‘fiziologii’ [Philosophical Logic: the Trilemma of Nomology, “Sociology” and “Physiology”], *Filosofskie nauki*, 2017, No. 1, pp. 85–94. (In Russian)

Нарьян С.К., Быков А.В. Проблема моральной агентности акторов: перспективы социологического подхода в контексте теории «моральной диады» // Социологический журнал. 2022. Т. 28. № 1. С. 8–23.

Naryan, S.K., Bykov, A.V. “Problema moral’noj agentnosti aktorov: perspektivy sociologicheskogo podhoda v kontekste teorii ‘moral’noj diady’” [The Problem of Moral Agency

of Actors: Perspectives of Sociological Approach in the Context of 'Moral Dyad' Theory], *Sociologicheskij zhurnal*, 2022, Vol. 28, No. 1, pp. 8–23. (In Russian)

Anderson, M., Anderson, S.L. "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine*, 2007, No. 28 (4), pp. 15–26.

Allen, C., Smit, I., Wallach, W. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches", *Ethics and Information Technology*, 2005, No. 7, pp. 149–155.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford UP, 2016.

Boden, M. *Mind as Machine: A History of Cognitive Science*. Oxford: Oxford UP, 2006.

Brundage, M. "Limitations and Risks of Machine Ethics", *Journal of Experimental and Theoretical Artificial Intelligence*, 2014, No. 26 (3), pp. 355–372.

Behdadi, D., Munthe, C. "A Normative Approach to Artificial Moral Agency", *Minds and Machines*, 2020, No. 30, pp. 195–218.

Bjornsson, G., Shepherd, J. "Determinism and Attributions of Consciousness", *Philosophical Psychology*, 2020, Vol. 33, No. 4, pp. 549–568.

Brachman, R.J. "Systems That Know What They're Doing", *IEEE Intelligent Systems*, 2002, No. 17 (6), pp. 67–71.

Chollet, F., *On the Measure of Intelligence* [https://arxiv.org/pdf/1911.01547.pdf, accessed on 05.02.2024].

Fleetwood, J., Vaught, W., Feldman, D., "MedEthEx Online: A Computer-Based Learning Program in Medical Ethics and Communication Skills", *Teaching and Learning in Medicine*, 2000, pp. 96–104.

Formosa, M. Ryan. *Making Moral Machines: Why we need Artificial Moral Agents* [https://philarchive.org/rec/FORMMM, accessed on 05.01.2024].

Fossa, F. "Artificial Moral Agents: Moral Mentors or Sensible Tools?", *Ethics and Information Technology*, 2018, Vol. 20, pp. 115–126.

Gray, K., Young, L., Waytz, A. "Mind Perception is the Essence of Morality", *Psychological Inquiry*, 2012, Vol. 23. No. 2, pp. 101–124.

Inusah, H., Quansah, P.K. "Rawls' Reflective Equilibrium as a Method of Justifying Moral Beliefs", *Axiomathes*, 2022, No. 32 (Suppl 2), pp. 629–645.

Johansson, L. "Robots and Moral Agency", *Theses in Philosophy from the Royal Institute of Technology*, Stockholm, 2011.

Krinkin, K., Shichkina, Y., Ignatyev, A. "Co-evolutionary Hybrid Intelligence is a Key Concept for the World Intellectualization", *Kybernetes*, 2023, Vol. 52, No. 9, pp. 2907–2923.

Kuleshov, A., Ignatiev, A., Abramova, A., Marshalko, G. "Addressing AI Ethics Through Codification", *Proceedings of the International Conference Engineering Technologies and Computer Science (EnT)*. Moscow, 2020, pp. 24–30.

Martinho, A., Poulsen, A., Kroesen, M., Chorus, C. "Perspectives About Artificial Moral Agents", *AI and Ethics*, 2021, No. 1, pp. 477–490.

Podschwadek, F. "Do Androids Dream of Normative Endorsement? On the Fallibility of Artificial Moral Agents", *Artificial Intelligence and Law*, 2017, No. 25 (3), pp. 325–339.

Searle, J. "The Construction of Social Reality", in: J. Searle, *Rationality in Action*. New York: The Free Press, 1995.

Searle J. *Internationality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge UP, 1988.

Taylor, A. *Animals and Ethics: An Overview of the Philosophical Debate*. New York: Broadview Press, 2003.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A. "Implementations in Machine Ethics: A Survey", *ACM Computing Surveys*, 2020, No. 53 (6), pp. 1–38.