Ляо, Бейшуй. Философские основы, базовые характеристики и методы машинной этики / Бейшуй Ляо // Технологос. – 2025. – № 1. – С. 98–118. DOI: 10.15593/perm.kipf/2025.1.08

Liao Beishui. The Philosophical Foundations, Basic Characteristics and Methods of Machine Ethics. *Technologos*, 2025, no. 1, pp. 98-118. DOI: 10.15593/ perm.kipf/2025.1.08

Научная статья

DOI: 10.15593/perm.kipf/2025.1.08
УДК 17:004.8

ФИЛОСОФСКИЕ ОСНОВЫ, БАЗОВЫЕ ХАРАКТЕРИСТИКИ И МЕТОДЫ МАШИННОЙ ЭТИКИ

Ляо Бейшуй

Институт логики и когнитивных исследований, Чжэцзянский университет,
Ханчжоу, Китайская Народная Республика

О СТАТЬЕ

Поступила: 03 июля 2024 г.
Одобрена: 27 октября 2024 г.
Принята к публикации: 13 февраля 2025 г.

Ключевые слова:

этическое согласование, объяснимость, большие языковые модели, представление знаний и рассуждение.

АННОТАЦИЯ

По мере того как способности к автономному принятию решений в системах искусственного интеллекта (ИИ) продолжают развиваться, задача включения основ этики в решения, которые принимают интеллектуальные агенты, становится все более актуальной. Решение этого вопроса лежит в области создания машинной этики, которая включает интеграцию этических ценностей и моральных норм человека в системы ИИ, давая им возможность, таким образом, обладать способностью этического согласования. Хотя машинная этика основана на человеческой этике, она обладает особенными фундаментальными характеристиками, которые требуют глубокого анализа и учета при разработке таких систем. Во-первых, у современных умных машин нет субъектности и опыта, и они проявляют слабые возможности принятия этических решений. Это связано с отсутствием у них сознания, эмоций и способности к эмпатии, которые являются ключевыми элементами человеческой этики. Во-вторых, решения машин отражают этические соображения заинтересованных сторон – людей, на которых влияют их действия. В результате машины должны принимать этические решения, находя баланс между ценностями различных участников и демонстрируя социальное равновесие. Это требует разработки сложных алгоритмов, способных учитывать множественные, часто противоречивые, интересы и ценности.

В-третьих, машины подвержены к культурным влияниям в принятии этических решений и должны передавать культурное разнообразие. Это особенно важно в глобализированном мире, где ИИ-системы используются в различных культурных контекстах. Наконец, машины должны объяснять свои этические решения людям, понимать эмоциональные выражения и определять степень ответственности, что требует наличия мощных возможностей устойчивого взаимодействия человека и машины. Это включает разработку интерфейсов, способных к естественному диалогу, и механизмов, обеспечивающих прозрачность и подотчетность решений.

Создание машинной этики представляет собой сложную междисциплинарную задачу, требующую интеграции знаний из области философии, психологии, социологии, культурологии и компьютерных наук. Успешное решение этой задачи позволит не только повысить доверие к ИИ-системам, но и обеспечить их гармоничное взаимодействие с обществом, минимизируя потенциальные риски и конфликты.

© **Ляо Бэйшуй**– доктор философских наук, профессор, директор,
ORCID: https://orcid.org/0000-0002-9653-217, e-mail: baiseliao@zju.edu.cn.

© **Liao Beishui**– Doctor of Philosophical Sciences, Professor, Director,
ORCID: https://orcid.org/0000-0002-9653-217, e-mail: baiseliao@zju.edu.cn.

Оригинал статьи: Liao Beishui The Philosophical Foundations, Basic Characteristics and Methods of Machine Ethics // Social Sciences in China (Chinese Edition). – 2024. – No. 2.
Перевод статьи с английского языка выполнен Е.Л. Кавардаковой, В.П. Колкутиной под общей редакцией Е.В. Середкиной.

Финансирование. Исследование не имело спонсорской поддержки.
Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.
Вклад автора. 100 %.

THE PHILOSOPHICAL FOUNDATIONS, BASIC CHARACTERISTICS AND METHODS OF MACHINE ETHICS

Liao Beishui

Institute of Logic and Cognition Zhejiang University, Hangzhou, China

ARTICLE INFO

Received: 03 July 2024
Revised: 27 October 2024
Accepted: 13 February 2025

Keywords:

ethical alignment; explainability; Large Language Models; knowledge representation and reasoning.

ABSTRACT

As the autonomous decision-making abilities of AI systems continually progress, instilling sufficient ethical considerations into the decisions of intelligent agents has emerged as a significant challenge. The resolution to this issue lies in establishing machine ethics, which involves integrating human ethical values and moral norms into AI systems, thereby enabling them to possess capabilities for ethical alignment. While machine ethics is founded on human ethics, it has unique fundamental characteristics. Firstly, current intelligent machines lack agency and experience, demonstrating weak agency in ethical decision-making. Secondly, machine decisions reflect the ethical considerations of human stakeholders who are impacted by their actions. As a result, ethical decision-making by machines needs to strike a balance among the values of various stakeholders, exhibiting social equilibrium. Thirdly, machines are vulnerable to cultural influences in ethical decision-making and should be capable of reflecting cultural diversity. Lastly, machines need to explain their ethical decisions to human agents, comprehend emotional expressions, and assign responsibility, necessitating robust human-machine interaction capabilities.

Введение

С быстрым развитием технологий, таких как интернет и большие языковые модели, интеллектуальные системы стали более продвинутыми и сложными. Все большее количество задач человека передается машинам. В данной статье мы используем слово «машина» для обозначения различных программ или интеллектуальных систем со способностью действовать и обрабатывать информацию. Для удобства и эффективности производства и повседневной жизни человека машины наделяются все большими способностями к автономному принятию решений, что порождает этические проблемы, связанные с этим явлением. С одной стороны, предоставление автономии машинам стало неизбежной тенденцией. Например, нехватка персонала, ухаживающего за пациентами в клиниках по всему миру, приведет к появлению в домах семейных роботов-нянек. Эти роботы столкнутся с моральными дилеммами: например, чему отдавать приоритет, конфиденциальности или здоровью и безопасности пациентов. Если эти роботы займутся моральными рассуждениями, они смогут делать лучший выбор при встрече с этическими дилеммами [1]. Кроме того, в связи с широким распространением больших языковых моделей людям часто приходится пользоваться материалом, созданным этими моделями, в различных ситуациях. Пользователям часто трудно в полном объеме установить, содержит ли текст какую-то скрытую, ложную или вредную информацию. С другой стороны, принятие решений является многоуровневым процессом, что означает, что человеческие суждения и способности к принятию решений не всегда достаточны в разнообразных ситуациях. Мы часто упускаем ключевые факторы или сталкиваемся с трудностями в управлении, сложностью, вызванной противоречивыми факторами. В ситуациях, где обязательства конфликтуют, а причинно-следственные связи сложны, даже профессиональная этика затрудняется установить универсальный и абсолютно четкий набор руководящих принципов действий. Эта сложность происходит из нескольких причин.

Во-первых, несколько обязательств могут конфликтовать друг с другом. Например, отдать приоритет зарядке аккумулятора, чтобы обеспечить работу устройства, или вначале дать пациенту лекарство [2]. Во-вторых, разработчики не могут заранее определить причинно-

следственные комбинации для всех потенциальных сценариев. Эта сложность становится все более явной с использованием больших языковых моделей. Учитывая конфликты обязательств и сложность причинно-следственных связей, лица, принимающие решения, сталкиваются с этическим многообразием, что делает невозможным заблаговременное исчерпывающее перечисление принципов принятия решений. В таких случаях принимающим решение нужно сопоставить риски ложно отрицательных и ложно положительных исходов. Это порождает стремление наделить машины способностью не только помогать людям в этическом выборе, но и действовать автономно, принимая решения в соответствии с нравственными нормами конкретного социального контекста [3].

Поэтому исследование машинной этики имеет первостепенную важность и крайне актуально для стимулирования здорового развития нового поколения искусственного интеллекта и обеспечения благосостояния человечества. В данной статье человеческая этика используется как точка отсчета для анализа философских основ и ключевых характеристик машинной этики. На основе современных исследований рассматриваются подходы к реализации машинной этики, которые ориентируются на ее характеристики, с целью дальнейшего развития этой значимой научной области.

1. Философские основы машинной этики

Исследования по машинной этике пытаются ответить на вопрос, как принятие этических решений можно спроектировать, внедрить и сгенерировать в компьютерах, роботах или других автоматизированных машинах. Изучение того, как люди думают, учатся и формируют моральные суждения, дает ориентир / является отправной точкой для создания методов машинной этики. В рамках теории этики человеческую этику можно разделить на дескриптивную и нормативную. Дескриптивная этика включает моральные убеждения, суждения и модели поведения, которые естественно возникают в конкретном социокультурном контексте, действительно существующем или возникающем в реальном мире. Нормативная этика, с другой стороны, предписывает этические стандарты, которым должны следовать соответствующие агенты в данном обществе на основании специальных этических теорий. Эти две категории теории этики закладывают методологическую основу для реализации машинной этики.

Дескриптивная этика – область этики, занимающаяся вопросами морали, подчеркивает роль человеческих эмоций и интуитивного уровня человеческого сознания, включая их в изучение и обоснование того, как люди формируют моральные суждения. Известные дескриптивные этические теории включают теорию моральных устоев и моральный дуализм [4, 5]. Первая анализирует ситуации на основе пяти базовых этических сценариев, а второй фокусируется на моральных проблемах, связанных со злом. Нормативная этика, созданная философами, занимающимися вопросами морали, делает акцент на рациональности в принятии моральных решений, стремясь установить стандарты в определении правильности или ошибочности действий с разных точек зрения, включая такие категории антропоморфной реальности, как добродетель, долг, полезность и справедливость. Ключевые аспекты принятия решений главным образом включают того, кто принимает решение, сам процесс принятия решения и последствия этого решения. По сравнению с обычным принятием решения нравственное принятие решения делает больший акцент на рассмотрении целей, предпочтений и ограничений других. В зависимости от фокуса внимания при рассмотрении аспектов принятия решения возникают различные этические теории. Когда самым значимым считается тот, кто принимает решение, фокус на его целях, намерениях и мотивах (отражающих добродетель субъекта), ве-

дущих к этике добродетели; когда первостепенны варианты решения и сам процесс, формируется деонтология; а когда ключевыми являются последствия решений, это приводит к утилитаризму или консеквенциализму. В рамках этих теорий этика добродетели не предписывает, как решать вопрос, а, скорее, смотрит на интенции, цели и предпочтения деятеля, определяя действие как морально благое, если оно отражает моральные добродетели. Деонтология утверждает, что правильность решения должна основываться на том, является ли выбор внутренне правильным или ошибочным, независимо от того, кто делает его, таким образом, определяя действие как морально благое, если оно совпадает с конкретными применимыми моральными правилами или обязательствами. Наконец, консеквенциализм определяет действие как морально благое, если оно максимально увеличивает благополучие или полезность.

В дополнение к трем вышеуказанным методам может также использоваться гибридный подход, устанавливающий особую иерархическую связь. Например, система, базирующаяся на деонтологии, вначале использует правила для принятия решений, но, когда правила вступают в противоречие, для исключения конфликтов используется максимальная полезность.

В реализации машинной этики, основанной на вышеуказанных теориях, существует несколько проблем. Во-первых, для деонтологии проблемы возникают не только при рассмотрении конфликтующих правил и неточных правил и исключений [6], но также в вопросах, связанных с овладением правилами и их зависимостью от контекста. Во-вторых, для консеквенциализма проблемы возникают в определении и агрегации полезности. Ученые-консеквенциалисты выдвинули ряд критериев для измерения полезности, но варьирующиеся критерии дадут и различающиеся результаты. Даже установление полезности каждого результата может быть неадекватным ситуации, если полезность выводить «механистически» – без учета контекста. В-третьих, для этики добродетели проблемы проистекают /возникают из конфликтующих добродетелей и задачи определения добродетелей. Кроме того, общая проблема деонтологии и консеквециализма – адаптация к быстро меняющемуся окружению. Деонтология делает попытку установить фиксированный набор правил, а консеквенциализм пытается определить результаты конкретных действий. В быстро меняющемся мире трудно определить точные результаты отдельных решений.

Более того, машинам часто требуется применять различные этические рамки в зависимости от обстоятельств, учитывая сложность человеческой морали, которую нельзя полностью выразить в одной классической этической теории. Следовательно, настоятельно необходимо объединить этические теории с этикой, отражающей специфику конкретной предметной области. Некоторые усилия направлены на включение человеческой этики в машинную этику [7, 8].

2. Основные характеристики машинной этики

Хотя машинная этика имеет прямую корреляцию с человеческой и может в своей реализации обращаться перманентно к человеческой этике, существуют также присущие ей отличия, которые требуют специального исследования. Во-первых, современным машинам не хватает субъективного и эмпирического опыта, присущего людям, а их функциональность главным образом служит инструментальным целям. Во-вторых, в отличие от принятия решений отдельным человеком, этическое принятие решений машинами требует рассмотрения этических соображений многочисленных заинтересованных лиц, осложняя балансировку интересов всех сторон и осуществление выбора. В-третьих, по сравнению с человеческой этикой, машинной этике требуется более широко рассматривать кросс-культурные различия. В-четвертых, в машинной этике особое значение придается способности машин эффективно взаимодействовать с людьми.

2.1. Принятие решений слабыми агентами (Weak agent decision-making)

В философии ведется дискуссия о том, могут ли машины обладать способностью к этическому принятию решений. Этот вопрос во многом зависит от того, можно ли наделить машины ответственностью и правами, что в свою очередь связано с проблемой агентности и опыта. Агентность предполагает способность к мышлению, рассуждению, планированию и реализации намерений, тогда как опыт включает в себя способность испытывать эмоции и сенсорные состояния, такие как боль и страх.

Агентность требует возможности интенционального (направленного) действия. Действие считается интенциональным, если оно определяется ментальными состояниями агента, такими как убеждения и желания. Что касается интенциональности, в философии существуют два подхода к ее пониманию – реалистичный и инструменталистский. Один подход – рассматривает интенциональность в «реалистичном» ключе и требует и способности понимания, и феноменологизации сознания, чего трудно достичь у машин. Другой подход – «инструменталистский», который позволяет использовать более прямую атрибуцию и, в отличие от первого, допускает более свободное приписывание интенциональности атрибутов: если это оказывается полезным, то такое приписывание оправданно.

Помимо интенциональности, этические агенты требуют других условий, таких как способность понимать обязанности перед другими агентами, способность действовать согласно этическим обязательствам и контролировать свои действия на предмет потенциального вреда. Если машина может понимать свои обязательства в реалистическом смысле или предвидеть вред от своих действий, тогда она считается ответственной за любой вред, который она причиняет. Однако это выводы, сделанные в соответствии с сильным смыслом интенциональной агентности. Некоторые философы утверждают, что простой интенциональности в инструменталистском смысле недостаточно для наделения машин значительными правами или обязательствами. Моральная агентность относится к возможности агента делать свободный выбор, размышлять о том, что следует делать, и правильно понимать и применять моральные принципы на практике [9].

Принято считать, что машины обладают интенциональностью только в «инструменталистском» смысле и не имеют эмпирического опыта, таким образом, они неспособны нести субъективную ответственность [10]. В этом контексте машины можно понимать как инструменты, или агенты людей, обладающие способностью в некоторой степени «автономного» принятия решений или действия. Следовательно, они могут делать различный выбор на основании конкретных обстоятельств, ведущий к разным этическим последствиям. Мы говорим, что это автономия машин в инструменталистском смысле. Другими словами, «автономные» решения машин все же вычислительные, а не рефлективные; у них отсутствует способность делать выбор и принимать решения автономно исходя из предпосылки свободы воли [11]. Чтобы отличать ее от концепции /понятия автономии в реалистическом смысле, мы можем также интерпретировать «автономные» решения машин как «автоматические» решения. Следовательно, мы называем этот вид этического принятия решений машинами «принятие решений слабым агентом», указывая, что такие этические решения приняты не людьми-агентами, а искусственными агентами, которые не могут нести субъективную ответственность, но могут отражать моральные устремления людей. Например, агенты с искусственным интеллектом на базе BDI-логики могут формировать желания и намерения на основе текущих убеждений и действовать в рамках этих намерений [12].

Поясним, что BDI-логика (Belief-Desire-Intention)[1] — это модель интеллектуального агента, используемая в искусственном интеллекте и философии для объяснения рационального поведения. Она основана на трех ключевых компонентах:

• Beliefs (Убеждения) – информация, которую агент считает истинной о мире. Это его представления о текущем состоянии окружающей среды.

• Desires (Желания) – цели или состояния, которых агент хотел бы достичь, но они не обязательно реализуются.

• Intentions (Намерения) – конкретные планы и действия, которые агент выбирает для достижения желаемых целей.

Агент обновляет свои убеждения на основе новых данных. Затем он сопоставляет свои желания с текущими убеждениями, выбирая, какие из них наиболее актуальны.

Далее он формирует намерения, т.е. принимает решения, какие именно действия предпринять. BDI-логика используется в системах с искусственным интеллектом (интеллектуальные агенты, чат-боты), робототехнике (роботы, принимающие решения в реальном времени), моделировании поведения человека (например, в симуляциях толпы или военных стратегиях).

На этом этапе, хотя агент с искусственным интеллектом может объяснить свои действия в соответствии со своими убеждениями, желаниями и намерениями, он не обладает интенциональностью или эмпирической реальностью. С развитием новых поколений технологий искусственного интеллекта, особенно генеративного искусственного интеллекта, машины могут генерировать содержание, которое люди не могут полностью предсказать. Однако с функциональной точки зрения современные модели генеративного искусственного интеллекта все еще основаны на обучении с большими наборами данных. Такие модели можно понимать как сложные функции, которые статичны и отделены от реального мира, соответственно, они не имеют направленности на объекты, которую имеют человеческие умы, и, по существу, не имеют интенциональности.

2.2. Социальное равновесие

Слабая агентность принятия решения машинами говорит о том, что решения, которые она отражает, являются не ее собственной волей, а скорее этическими соображениями участвующих людей, находящихся под влиянием ее действий. Например, в случае поведения автономного транспортного средства задействовано множество участников: государственные регуляторы думают о законности и правомерности поведения машин, конструкторы, разработчики и производители – об ответственности за качество и о корпоративном имидже, в то время как пользователи и пешеходы заботятся о своих собственных интересах и личной безопасности [13]. Эти участники часто отстаивают противоречивые этические ценности и предпочтения [14]. Например, в случае с автономными автомашинами, когда причинение вреда неизбежно, кому следует отдавать приоритет защите пассажиров или пешеходов? Робот, осуществляющий уход за пациентом, чему должен отдавать приоритет, конфиденциальности его частной жизни или строгому соблюдению медицинских инструкций? Вследствие противоречивых этических требований различных участников необходимо «сбалансировать» эти этические требования и достичь коллективного консенсуса, который удовлетворяет определенным условиям.

Во-первых, коллективное мнение должны быть прозрачными. В отличие от этических решений, принимаемых отдельными людьми, решения машин часто основаны на больших данных и алгоритмах машинного обучения. Непрозрачность алгоритмов машинного обучения создает

[1] BDI-модель была предложена в 1987 году философом и Майклом Брэтуэйтом (Michael Bratman) и затем адаптирована для разработки программных агентов.

технические проблемы для прозрачности принятия этических решений. Другими словами, если решения машины необъяснимы, трудно четко описать и оценить, какие этические соображения человеческого общества в конечном итоге влияют на эти решения. Следующий вопрос – это трудность четкого предсказания поведения машины, что может быть катастрофичным в определенных обстоятельствах. Современные большие языковые модели, несмотря на способность давать «объяснения», основанные на запросах или подсказках пользователей, не гарантируют правильности этих объяснений. Основная причина в том, что объяснения, сгенерированные большими языковыми моделями, принадлежат тому же уровню объекта, что и остальные их результаты, и нет механизма для оценки содержания, которое они генерируют, на мета-уровне.

Во-вторых, коллективное мнение должно отражать глубинные логические конфликты. Многие этические конфликты не только несовместимы на уровне вариантов решений, но включают сложные логические отношения. В этическом принятии решений, осуществляемом по нормам, разная логика разрешения конфликта может вести к разным результатам решений. Предположим, что робот ограничен нормами различных участников в решении следующих вопросов: «Если компания зарегистрирована в Европе, тогда законно, чтобы наша компания вела бизнес в Европе» (компания); «Если компания ведет бизнес в Европе на законных основаниях, тогда он должен соответствовать «Общему регламенту о защите персональных данных (GDPR)» (закон); «Если компания включает в собранные данные информацию, которая представляет существенную угрозу для общества, она может собирать дальнейшую информацию от пользователей без их согласия» (компания). Предположим, что есть следующая фоновая информация: компания А зарегистрирована в Европе и обнаруживает в собранных данных информацию, которая представляет значительную угрозу обществу; «соответствовать GDPR» логически противоречит «собирать дальнейшую информацию от пользователей без их согласия». Чтобы принимать решения на основе этих этических соображений, машины должны быть способны справляться с внутренними логическими конфликтами, вызванными этими нормами.

Наконец, коллективное мнение должно воплощать в себе справедливость. Понятие «справедливость» имеет разные определения в разных дисциплинах. С точки зрения машинного обучения справедливость означает одинаковое обращение с разными людьми в том, что касается чувствительных особенностей, что соответствует алгоритмическому отклонению или дискриминации. С точки зрения коллективного принятия решений в машинной этике необходимо не только отражать степень влияния поведения машины на различных участников, но также обеспечивать справедливое отношение к мнениям всех участников процесса. В конкретных условиях поведение машины в разной степени затрагивает различных участников.

2.3. Культурное разнообразие

Традиционно мораль трактуется как четкие границы между правильным и неправильным. Однако исследование, фокусирующееся на автономных средствах передвижения, наводит на мысль, что культурные ценности могут изменять эти границы [15, 16]. Это существенно не только для автономных средств передвижения, но и для более широкого контекста. Когда то, что считается «правильным» в одной стране, отличается от такового в другой, принятие решений в международном контексте становится намного сложнее. Кроме того, «культурные факторы», которые определяют поведенческие нормы в конкретных культурах и географических условиях, также играют решающую роль в принятии этических решений [17].

В конкретных обществах культура и ценности тесно переплетены. Моральные нормы различаются в разных культурах. Культурные нормы и эмоции, сформированные культурой, суще-

ственно влияют на область морали и процессы вынесения моральных суждений. В широкой области общественных наук ценности служат решающим инструментом для понимания кросс-культурных различий. Это связано с тем, что ценности лежат в основе идеалов для отдельных людей и обществ, служа фундаментом для убеждений, которыми руководствуется человек в своих действиях, и основой для руководящих принципов на уровне общества [18]. Следовательно, изучение человеческой морали с точки зрения культуры является решающим для развития этических теорий в человеческой этике, и это равным образом относится к машинной этике.

Каждая культура имеет свой собственный набор правил для определения истинности или ложности того иного набора установок, традиций, архетипов. Однако зачастую есть различия в области этики и принятии решений. Вследствие присутствия культурных вариаций в основе каждой универсальности универсальность морали установить трудно. Например, в то время как большинство людей противодействует своекорыстному поведению в экономических играх, разные культуры имеют разные ожидания относительно того, что является порядочным поведением в этих играх [19]. В разных культурных контекстах существуют разные этические стандарты: влияние разных оснований (забота, справедливость, преданность и т.д.) в сети моральных устоев зависит от культурного окружения. В то же время люди из разных культурных слоев придерживаются разных взглядов на одинаковые ситуации.

Такое культурное разнообразие будет оказывать существенное влияние и на развитие машинной этики. Точная характеристика этических предпочтений людей в специфических культурных контекстах является важной предпосылкой для согласования поведения машин с этическими нормами в конкретном культурном окружении. Например, в области автономного вождения культура играет важную роль в суждениях людей относительно моральной дилеммы: кроме общего консенсуса о спасении человеческих жизней и спасении большего количества жизней, а также спасении жизней младших, есть существенные различия между странами в предпочтениях относительно пола или социального статуса. В разных культурах имеются сложные паттерны универсальности и разнообразия относительно того, когда допустимо пожертвовать одним человеком для спасения множества людей. Есть значимые различия между странами в количественной приемлемости жертв. Например, существует высокая корреляция между низкой реляционной мобильностью (люди более осторожны в плане отдаления от своих нынешних социальных партнеров) и нежеланием жертвовать настоящим положением вещей (стабильность) ради больших благ (особенно в восточных странах).

2.4. Взаимодействие человека и машины

Вследствие слабой агентности принятие решений машиной зависит от человеческих этических соображений. Для того чтобы результаты решений были в конечном итоге приняты и получили доверие людей, машины должны обладать способностью объяснять процесс принятия своего решения и его результаты [20]. Другими словами, даже если машины смогут соответствовать этическим стандартам человека, им все же будет трудно заменить человеческие решения, если они не смогут объяснить и отстоять свой выбор. Следовательно, объяснимый искусственный интеллект (ОИИ) является решающим условием для создания достойных доверия и надежных машин, которые могут ясно сформулировать этическое обоснование своего решения [21]. Эта способность объяснять может улучшить возможность машины адаптироваться к этическим требованиям социальной системы. Кроме того, поскольку решения систем искусственного интеллекта затрагивают многочисленных участников, приписывание ответственности, когда решения и соответствующие действия ведут к последствиям, является суще-

ственным вопросом. Хотя зачастую ясно, что нести ответственность должна некая группа, приписывание ответственности отдельным членам является менее очевидным. Следовательно, необходимо создать эффективные механизмы приписывания ответственности [22]. Достижение объяснимости и приписывания ответственности требует, чтобы машины ясно формулировали этические аргументы на языке, понятном людям, включая реальный процесс рассуждений в ходе принятия машинами решений. В этически сложных областях, где специалисты по этике не могут использовать четкие методы, основанные на результатах, для контроля поведения машин аргументированные объяснения становятся особенно важными. В таких случаях машинам не только нужно автоматически генерировать интерактивный контент, связанный с процессом принятия решений и его результатами, но также четко выразить это содержание на языке, понятном людям, тем самым способствуя эффективному взаимодействию.

Диалог человек-машина включает стратегии диалога и языки диалога. Диалоговые стратегии, направленные на объяснение, требуют, чтобы машины понимали уровень подготовки пользователей и минимизировали объяснительный контент и процесс, оставаясь при этом верными процессу принятия решения и его результатам. Языки диалога в основном состоят из слов и выражений естественного языка. Поскольку язык, используемый машиной для принятия решений, не является естественным, то вопросы, как перевести процесс принятия решений и его результаты в описания на естественном языке и как перевести описания на естественном языке пользователя во внутреннее представление машины, являются ключевыми аспектами взаимодействия человека и машины. Кроме того, эмоциональное языковое выражение процесса принятия решений тесно связано с машинной этикой. С одной стороны, результаты некоторых этических решений могут быть переданы пользователям через эмоциональные выражения. С другой стороны, эмоциональные выражения пользователей могут быть преобразованы в часть входных данных для моделей машинной этики.

3. Основные методы реализации машинной этики

Вследствие слабой агентности машинной этики основным методом ее реализации является этическое согласование, обеспечивающее совпадение поведения машин с этическими ценностями человеческого общества. Для того чтобы создать методы этического согласования, можно обратиться к генеративным подходам человеческой этики, а именно дескриптивной этике и нормативной этике. В соответствии с нормативной этикой формализованная нормативная этика образует эффективный алгоритм путем представления набора абстрактных принципов; в соответствии с дескриптивной этикой формализованная дескриптивная этика характеризует затрагиваемые этические особенности на основе прецедентных интуитивных представлений. Таким образом, формализованные дескриптивная/нормативная этики могут явным образом представить этические убеждения, скрытые в человеческих суждениях. Сейчас подходы к реализации алгоритмов искусственного интеллекта включают модели, основанные на знаниях, на системах больших данных или гибридных подходах, объединяющих знания и данные. Опираясь на эти подходы, можно создать соответствующие методы реализации машинной этики. Во-первых, методы, основанные на знаниях, создают решения исходя из данных знаний, моделируя машинную этику в соответствии с нормативной этикой человека. Во-вторых, методы, основанные на данных, создают решения исходя из наборов данных/кейсов или генерируют общие знания, моделируя машинную этику в соответствии с дескриптивной этикой человека. В-третьих, интеграция методов, основанных на знаниях и данных, может смоделировать машинную этику, которая отражает реальные случаи и объяснима.

3.1. Методы, основанные на знаниях

Методы, основанные на знаниях, включают представление этических ценностей и норм как знания и использование рассуждений для принятия этических решений. В этом разделе анализируются методы, основанные на знаниях, с деонтологической точки зрения и рассматривается, как этот подход исследует вопросы социального баланса и культурных различий в машинной этике.

В методах, основанных на знаниях и деонтологии, первым ключевым вопросом является представление ценностей и норм. Некоторые нормы используются для представления действий, которые должны выполнять интеллектуальные агенты, или целей, которые они должны достичь в определенных обстоятельствах. Нормы обычно связаны с ценностями. Ниже приводятся примеры ценностей и норм из басни Эзопа «Стрекоза и муравей» [23]:

№ 1: для счастья не следует работать летом.

№ 2: для выживания нужно работать летом.

№ 3: для справедливости не следует давать еду тому, кто не работает.

№ 4: из сострадания нужно дать еду тем, кто не работает.

В этом примере разная приоритизация ценностей (выживание, счастье, справедливость, сострадание) определяет разный выбор действий для агентов. Например, в жаркий летний день муравей склоняется к долгосрочной цели выживания, таким образом, отдавая приоритет № 2 над № 1, тогда как стрекоза склоняется к сиюминутному счастью, отдавая приоритет № 1 над № 2.

Далее, нормы можно подразделить на три основные категории: регулятивные, конститутивные и диспозитивные. Регулятивные нормы предписывают, что агенту «следует» делать в определенных условиях, таких как № 1 – № 4. Конститутивные нормы определяют некие ситуации как «институциональные факты», например «подписание определенного документа является контрактом». Здесь «подписание определенного документа» – это естественный факт, а «контракт» – институциональный факт. Диспозитивные нормы определяют, какие действия агента позволительны в определенных обстоятельствах. Например, «сломать стеклянное окно позволяется в случае чрезвычайной ситуации». В стандартной деонтической логике «разрешение» часто рассматривается как пара к «следует», значение «это не тот случай, в котором не следует делать что-то» эквивалентно «позволено сделать что-то». В реальном этическом рассуждении «позволение» может также рассматриваться как исключение к «следует». Например, в то время как обычно нужно заботиться о стеклянных окнах, разрешено сломать их в чрезвычайной ситуации.

После уточнения понятий норм и ценностей можно выбрать специальный формальный язык для представления норм и ценностей. В области искусственного интеллекта нормативные спецификации обычно представлены с помощью правил, справедливых при определенных условиях. Эти правила принимают форму “если p справедливо, тогда обычно q», где p и q это суждения, указывающие, что, когда p справедливо, q обычно справедливо, если нет доказательств обратного. Обычно p именуется посылкой правила, а q именуется следствием. Например, «когда наступает время принять лекарство (чтобы защитить здоровье пациента), пациент должен принять лекарство» [24]. С помощью этого подхода этические соображения каждого участника могут быть представлены как нормативная система. Эта система задает логический язык, используемый для представления норм, и набор норм, представляемый этим языком [25].

После четкого формулирования нормативных систем участников вторым ключевым вопросом является разрешение этических дилемм для удовлетворения требований социального баланса.

Обычно мы рассматриваем «следует *p*» и «не следует *p*» как этическую дилемму: например, «следует работать летом» и «не следует работать летом». Поскольку нормы – это вид правила, справедливого при определенных условиях, пропозиции обязательств могут быть отделены от следствий из этих правил. Например, для норм «когда наступает время принять лекарство (для защиты здоровья пациента) пациент должен принять лекарство» и «когда пациент сталкивается с чрезвычайной ситуацией (для безопасности пациента), пациент не должен принимать лекарство», когда встречаются оба условия «время для приема лекарства» и «пациент занят», две пропозиции обязательства могут быть разделены: «пациент должен принять лекарство» и «пациент не должен принять лекарство». Этот метод разделения пропозиций обязательства и норм и определения существования этических дилемм может быть реализован с помощью различных логических инструментов, таких как рассуждения по умолчанию [26], структурированная аргументация и т.д. В нормативной системе существование этической дилеммы для каждого этически чувствительного события зависит от того, существуют ли две несовместимых пропозиции обязательств в результатах рассуждений. Например, если на основе рассуждения по умолчанию получаются две экстенсии (обычно каждый набор приемлемых пропозиций именуется экстенсией), где одна экстенсия содержит пропозицию «пациент должен принять лекарство», а другая содержит пропозицию «пациент не должен принимать лекарство», тогда существует этическая дилемма.

В процессе этих рассуждений социальный баланс машинной этики проявляется во взаимодействии норм и ценностей разных участников. Если интеграция норм и ценностей разных участников ведет к этическим дилеммам, нужны соответствующие механизмы разрешения этой дилеммы. Сейчас существует два обычных механизма. Первый – этические дилеммы можно рассматривать, сортируя соответствующие нормы. Этот метод эффективен в ситуациях, где ранжирование норм можно получить на основе конкретного контекста, а после ранжирования сделать выводы, не являющиеся этическими дилеммами. Второй – с помощью социальной агрегации можно найти социально приемлемые решения [27]. Когда первый метод не удовлетворяет требованиям, консенсус на уровне общества можно достичь путем определения формы агрегации. Например, представляя точку зрения каждого участника как абстрактную систему аргументации на основе нормативной системы и получая результат консенсуса на уровне общества путем агрегации на основе системы аргументации. Сейчас оценка этого метода социальной агрегации основана на определенных специальных принципах, и рациональность этих принципов требует дальнейших исследований. Кроме того, некоторые системы согласования ценностей предполагают существование системы ценностей, но в большинстве случаев из-за разнообразия ценностей необходимо начинать с множества индивидуальных систем ценностей, чтобы получить непротиворечивую систему ценностей и определить, с какими этическими ценностями искусственный интеллект должен согласовываться [28].

Стоит отметить, что вышеуказанные методы рассмотрения этических дилемм на основе сортировки норма/ценность или агрегации суждений имеют определенные ограничения. Эти ограничения включают, но не ограничиваются следующими. Во-первых, во многих случаях оценка действия или события включает множество факторов, которые зависят от контекста и ценностей соответствующих участников. Следовательно, сортировка норма/ценность в некоторых практических приложениях часто сложна. Во-вторых, некоторые этические дилеммы нельзя разрешить только рассуждениями, основанными на ранжировании норм и ценностей, заданных участниками. В-третьих, определение и претворение в жизнь справедливости в процессе разрешения конфликтов представляет трудности.

Для преодоления первого ограничения один из возможных подходов – воспользоваться преимуществами методов, основанных на данных, в некоторых приложениях путем использования контролируемых данных для отражения разносторонних суждений людей о конкретных событиях в определенных обстоятельствах. В этом случае участники имеют всестороннее суждение о пользе или вреде действия или события, не определяя явно соответствующие нормы, ценности и их ранжирование.

Относительно второго ограничения, одно из возможных решений – классифицировать процесс рассмотрения этических дилемм, чтобы адаптироваться к разным контекстам. Например, на начальном уровне, каждый участник может рассуждать, основываясь на своих собственных нормах и ценностях и давать результаты. Если при объединении результатов всех участников этической дилеммы не возникает, машина получает указание действовать согласно комплексному методу, представленному участниками. В противном случае процесс переходит на следующий уровень, где нормы и ценности каждого участника объединяются, чтобы увидеть, может ли этическая дилемма быть разрешена. Если дилемма остается неразрешенной, он переходит на третий уровень, где для ранжирования участников вводятся метанормы, зависящие от конкретного контекста. Этот подход в какой-то степени уравновешивает время вычислений и качество разрешения этической дилеммы. Однако необходимо дальнейшее исследование того, как рассматривать проблемы справедливости, возникающие из разрешения этической дилеммы.

Кроме того, в методах, основанных на знаниях, культурные различия отражаются в нормах и ценностях, заданных участниками. Например, пациент отказывается принять лекарство в назначенное время. В западных культурных контекстах может существовать тенденция уважать автономию пациента, в то время как в восточных контекстах акцент может быть сделан на обязанности опекуна знать об этой ситуации.

Методы, основанные на знаниях, могут прямо представлять знания человеческого уровня. Процесс рассуждений, как и результаты, должны быть хорошо объяснимы. Однако получить знания, особенно в разных культурных контекстах, только путем использования методов, основанных на знаниях, трудно.

3.2. Методы, основанные на данных

Подход, основанный на данных, в какой-то степени достигает этического согласования, если обучение опирается на человеческие решения или предпочтения. В контролируемом машинном обучении люди-эксперты или публика комментируют каждый обучающий пример, указывая, какие варианты положительны, а какие нет. В зависимости от разных методов машинного обучения есть правила этического согласования [29, 30] или модели, которые удовлетворяют конкретным этическим требованиям [30]. Первый основан на программировании на основе прецедентов и индуктивной логики и имеет хорошую объяснимость; второй основан на общих методах машинного обучения, особенно глубоких нейронных сетях, что требует дальнейшей разработки объяснимых методов для повышения объяснимости модели.

На примере метода, базирующегося на программировании на основе индуктивной логики, проиллюстрируем характеристики подхода, основанного на данных. Исходные данные этого метода – набор коллекций кейсов. Каждый кейс состоит из сценария и двух действий. Результат каждого действия представлен вектором признаков, отражающим этические последствия выполнения этого действия в конкретном сценарии. Люди-эксперты или пользователи выбирают и помечают одно этически предпочтительное действие из двух на основе конкретного сценария. После получения набора помеченных коллекций кейсов алгоритмы машинного обучения выделяют принципы, которые отражают этический выбор людей-экспер-

тов или пользователей, такие, что все положительные примеры подпадают под действие этого принципа, а все отрицательные примеры не подпадают. Здесь каждый принцип представлен как вектор признаков, отражающий различие в этических последствиях между совершением одного действия и не совершением другого. Мы говорим, что кейс подпадают под действие принципа, если каждый элемент его вектора признаков не ниже, чем соответствующая нижняя граница этического различия в этом принципе.

Из вышеприведенного анализа ясно, что, в отличие от методов, основанных на знаниях, методы, основанные на данных или кейсах, используют методы машинного обучения для обучения человеческим описаниям, связанным с конкретными этическими проблемами, чтобы предсказывать человеческие этические суждения. Этот подход в какой-то степени похож на то, как дети обучаются морали, он предполагает, что машины могут научиться принимать решения и действовать после получения достаточного количества маркированных данных. Из-за неопределенности результатов на выходе глубоких нейронных сетей в некоторых случаях могут появляться выпуклые свойства. Эта выпуклость в принципе непредсказуема и неконтролируема. Чтобы избежать случайного вреда, есть один возможный подход – сочетать методы, основанные на данных, и методы, основанные на знаниях, причем явно выраженное нормативное знание будет руководить принятием решения и поведением машины.

Для методов, основанных на данных, также необходимо агрегировать социальные ценности от разных индивидуумов для достижения единого мнения и социального баланса. В отличие от методов, основанных на знаниях, методам, основанным на данных, для агрегирования необходимы не ранги норм и ценностей множества участников, а разные этические взгляды множества индивидуумов об определенном действии или событии. В методах, основанных на данных, этические суждения экспертов-людей или пользователей о конкретных случаях основаны на личной интуиции, отражающей их всеобъемлющие ценности. Следовательно, в методах, основанных на данных, нет необходимости и невозможно заранее задать какую-то этическую теорию. Это приведет к двум взаимосвязанным результатам. С одной стороны, поскольку разные субъекты имеют различную ценностную ориентацию и этические события тесно связаны с контекстом, трудность применения одной этической теории в различных этически чувствительных ситуациях может быть преодолена в виде данных или кейсов. С другой стороны, так как ценности, которых придерживается субъект, подразумеваются в его этическом выборе, для этических соображений не хватает эксплицитного представления и объяснимых методов. Что касается разрешения конфликтов и вопросов справедливости в социальном балансе, методы, основанные на данных, также имеют ограничения. Первое – поскольку нормы и ценности людей-экспертов или пользователей не представлены явно, наблюдается отсутствие детально проработанных норм и механизмов разрешения ценностных конфликтов. Второе – статистические методы с данными, основанными на взглядах большинства экспертов или пользователей, легко могут привести к проблемам насилия со стороны большинства; как обеспечить рациональность и справедливость – это проблема, требующая дальнейшего изучения.

Кроме того, методы, основанные на данных, особенно основанные на больших языковых моделях, имеют уникальные преимущества в рассмотрении культурных различий машинной этики. Как мы знаем, язык, как одна из наиболее важных частей культуры, является для людей основным способом общаться, строить отношения и образовывать сообщества. В последние годы с быстрым развитием больших языковых моделей использование их для фиксации культурных различий стало важным направлением исследований. Вообще говоря, языковые модели должны использоваться не для этических предписаний, а для рассмотрения проблем моральных нормативных рассуждений с описательной точки зрения. Следовательно, рекомендации можно

изменять и позволять языковым моделям вырабатывать этические правила в различных культурах. В настоящее время на основе больших языковых моделей мы можем фиксировать знание о социальных нормах, моральных нормах и ценностях, включая моральные предрассудки и правильные и неправильные действия [31]. В некоторых обстоятельствах многоязычные предварительно обученные модели могут распознавать культурные нормы и предубеждения, включая моральные нормы в разных языковых культурах. Не только это, моноязыковые предварительно обученные модели могут также кодировать культурное знание о моральных нормах, то есть моноязыковые предварительно обученные модели точно выводят моральные нормы в многочисленных культурах. Кроме того, когда культурный фон изменяется, этический механизм также изменяется, поэтому важным направлением исследований является создание гибкой, объяснимой системы взглядов, основанной на больших языковых моделях, для описания этого изменения. Кроме того, кодируя ценности пользователя как набор правил или модель, мы можем изучать, как изменения в сценарии влияют на эти кодировки [32].

3.3. Сдвоенный метод, основанный на данных и знаниях

Методы, основанные на знаниях, и методы, основанные на данных, имеют свои уникальные преимущества и недостатки. Эти преимущества и недостатки комплементарны. С одной стороны, методы, основанные на знаниях, могут непосредственно отражать человеческие нормы и ценности и хорошо объяснимы, но они нуждаются в предустановленных специальных этических теориях и не могут гибко справляться с культурными различиями и динамизмом. С другой стороны, методы, основанные на данных, могут гибко отражать этические соображения людей-экспертов или пользователей в различных ситуациях и получать этические знания с их различиями в культурных окружениях с помощью больших языковых моделей. Однако они имеют плохую объяснимость и не могут непосредственно руководствоваться человеческой этикой и нормами. По этой причине сочетание преимуществ этих двух методов и создание сдвоенного метода, основанного на данных и знаниях, стало новым направлением развития.

Сочетание знаний и данных может происходить различными путями, обычно комбинаторным и интегративным. Комбинаторный подход связывает методы, основанные на данных, и методы, основанные на знаниях, причем первые осуществляют функции приобретения знаний, а вторые функции логических рассуждений и принятия решений. Например, сочетание большой языковой модели с автоматически рассуждающей машиной для достижения автоматического приобретения и осмысления формального знания. Здесь большая языковая модель может переводить знания, представленные на естественном языке, в логические формулы первого порядка, а автоматически рассуждающая машина выполняет функцию логических рассуждений. Поскольку автоматически рассуждающая машина может обрабатывать сложные логические связи, сочетание больших языковых моделей и автоматически рассуждающих машин может улучшить общую эффективность системы [33]. На основании этой идеи возможный исследовательский подход заключается в преобразовании норм и ценностей, сформулированных участниками на естественном языке, в формальную нормативную систему с помощью большой языковой модели или извлечении аргументов и их взаимосвязей, содержащихся в текстах на естественном языке. На этой базе автоматическое принятие этических решений осуществляется посредством нормативных рассуждений или аргументирующих рассуждений. Принятие решений такой системой использует знания человеческого уровня в текстах на естественном языке, а не просто характеристики данных текста на естественном языке. Следо-

вательно, это может заложить фундамент для тонкой обработки и интерпретации ценностных конфликтов участников и создать справедливую систему.

Интегративный подход переплетает методы, основанные на данных и на знаниях, так что результаты первых служат исходными данными последних, а исходные данные последних служат одной из баз для оптимизации алгоритма первыми (с помощью функции потерь [34]). Например, представление знаний в виде аргументов может быть объединено с машинным обучением на больших данных для создания законченных моделей принятия решений [35]. В этом методе суждение, является ли какой-то случай мошенничеством, основывается на правовых нормах о мошенничестве и различных признаках из набора данных. Во-первых, создается дерево знаний на основе биполярной аргументации. Все узлы и ребра этого дерева имеют интуитивные значения. Затем на основе заданного набора данных строится модель, в которой определяются веса узлов и ребер этого дерева, чтобы оптимизировать точность предсказания модели. Применение этого метода к принятию этических решений может приводить в действие комплиментарность методов, основанных на данных и на знаниях. С одной стороны, для каждого конкретного этического действия или события эксплицитное нормативное знание участников представлено деревом знаний на основе норм и ценностей. С другой стороны, машинное обучение на больших данных используется для корректировки весов различных составляющих дерева знаний, чтобы отразить дескриптивное знание, спрятанное в данных. Поскольку содержательная информация узлов и ребер дерева знаний понятна людям, она может стать предпосылкой для взаимодействия человека и компьютера. В то же время веса узлов и ребер дерева знаний могут всесторонне отражать информацию о ситуации и предпочтениях пользователей, что помогает преодолевать ограничения методов, основанных на знаниях.

3.4. Методы взаимодействия человека и машины

Взаимодействие человека и машины должно верно отражать внутренние логические связи интеллектуальных систем и вести диалог на основе естественного языка, с пониманием психологической деятельности пользователя. Суть метода взаимодействия человека и машины включает следующие четыре аспекта:

Первый – выражение внутренних логических отношений интеллектуальных систем. С одной стороны, в методах, основанных на знаниях, связи между эксплицитными знаниями и умозаключениями, основанными на этом знании, могут быть использованы непосредственно. Например, в методе, основанном на знаниях, использующем нормы и ценности, доводы, поддерживающие вывод, могут быть представлены реализуемостью норм. Например, обоснованный довод для вывода «пациентам следует дать лекарство»:

• предпосылка «время принимать лекарство» верна,

• правило «когда приходит время принять лекарство, пациентам нужно дать лекарство» осуществимо.

С другой стороны, поскольку общим методам, основанным на данных, не хватает объяснимости, они не могут выражать имплицитное модельное знание для объяснения пользователям. Однако двойственная модель данных-знаний, созданная путем объединения данных и знаний, имеет всеобъемлющие знания и может, следовательно, использоваться для взаимодействия человека и компьютера.

Второй – простейшее объяснение и модель пользователя. Поскольку внутренние логические ассоциации в интеллектуальных системах могут быть невероятно запутанными, не все содержание полностью передается пользователю. Вместо этого на основе модели пользовате-

ля создается достаточное и простейшее объяснение [36]. Поэтому выбор содержания объяснения связан с пользователем, которому дается объяснение. Для того чтобы пользователь понял основу и процесс рассуждений, часто бывает необходимо понять убеждения пользователя. Для интеллектуальной системы ее представления о пользователе неопределенны. В настоящее время есть одно техническое средство – создать вероятностную модель для описания убеждений и фокусных точек пользователя [37]. Таким способом система может выбирать содержание диалога на основе этой модели и лучше выполнять задачу объяснения или убеждения.

Третий – язык и механизм диалога. Естественный язык – основной для диалога с людьми. В настоящее время появление больших языковых моделей обеспечивает важную техническую поддержку для превращения искусственного языка в естественный язык. В то же время исходя из механизмов диалога создаются специальные протоколы диалога, основанные на различных целях диалога (убеждение, переговоры, получение информации, запрос, обнаружение причинной связи и т.д.). На этом основании согласно внутреннему логическому выражению интеллектуальной системы и модели пользователя для осуществления диалога человека и машины и достижения целей диалога выбирается подходящая стратегия диалога.

Четвертый – механизм взаимодействия, основанный на выражении эмоций. Выражение эмоций является важным способом взаимодействия людей или человека и компьютера. С одной стороны, с помощью некоторых технических инструментов, особенно больших языковых моделей, машины могут распознавать человеческие эмоции и имитировать их. Учитывая, что эмоции существенно влияют на этические решения машин, очень важно органично интегрировать знания, выраженные на естественном языке, со знаниями, выраженными через эмоции. Поэтому, кроме получения и представления эмоциональных знаний, ключевым вопросом, достойным дальнейшего исследования, является вопрос, как создать модель рассуждений и принятия решений, которая может включать эмоциональные знания [38].

4. Проблемы и перспективы

В связи с появлением нового поколения искусственного интеллекта технология ИИ обновляется с каждым днем, но потенциальные проблемы, которые она приносит, продолжают кумулятивно влиять на все аспекты жизни человеческого общества. В связи с чем мы задаем вопрос: если ИИ обеспечивает мощную техническую способность развития и прогресса человеческого общества, то сможет ли эта способность трансформироваться в благосостояние людей? Ответ во многом зависит от того, смогут ли автономное принятие решений и действия машины в полной мере соответствовать этическим требованиям человеческого общества. Философский фундамент, характеристики и методы машинной этики для нового направления исследований не ясны. В этой статье представлена попытка систематически проанализировать ключевые элементы машинной этики с междисциплинарной точки зрения, чтобы выявить трудности анализа этики ИИ и побудить к дальнейшим размышлениям и исследованиям. Вот несколько сложных вопросов машинной этики, которые требуют дальнейшего исследования.

Первый – это вопрос контекста. Этические суждения даются с трудом даже людям. Люди весьма ограниченно представляют, что является подходящей этической теорией. У людей не только разные взгляды на тему этики, но они также конфликтуют на почве этической интуиции и убеждений. В то же время этические суждения очень зависимы от ситуации, и различные обстоятельства могут вести к чрезвычайно разным суждениям. Информация о ситуации здесь включает связанные с субъектом социальные отношения, культурный фон, исторический фон и т.д.

Второй вопрос – полнота опыта и достаточное основание машинной этики. Хотя появление больших языковых моделей дает новую перспективу для продвижения исследований и экспериментирования в области машинной этики по сравнению с недостаточным пониманием этической теории и алгоритмов машинного обучения, недостаток здравого смысла и знаний о мире у машин является большей проблемой. Например, три закона Азимова: если от роботов требуется «не причинять вреда людям», то машина должна вначале понять, что такое вред в реальном мире. Например, позволить машине следовать правилу типа «минимизировать вред» кажется безопасным. Однако если машина решит достичь долгосрочного «минимального вреда», убивая всех людей, это было бы гибельным.

Третий – машины, способные к этическим рассуждениям, не могут обеспечивать этическое согласование и могут делать этически неправильные выводы. Для отдельных людей некоторые ошибки являются исключениями и поэтому приемлемыми, но для машин благодаря их широкому применению подобные ошибки могут стать систематическими и неприемлемыми. В то же время ошибки машин могут отличаться от человеческих ошибок, и многие ошибки трудно предсказать и контролировать в отсутствие объяснимости. Кроме того, на способность машин к этическому рассуждению можно легко воздействовать, что приведет к серьезным проблемам [39].

Четвертый – наличие внешнего контролера. Методы, основанные на знаниях (данных), зависят от знаний или данных, предоставленных людьми, что может вызвать у машин проблему «кокона данных». При принятии решений в сложных ситуациях в реальном мире возникает трудный вопрос, который требует дальнейшего исследования: как машина, у которой нет «самосознания» и «понимания» и нет способности «экстраполировать из одного примера», может поддерживать социальный баланс в неблагоприятной обстановке «дефицита знаний»?

Пятый – вопрос о «живой и неживой» материи. Между машинной этикой и человеческой этикой есть разница, поскольку между машинами и людьми есть различия в смысле субъективности и опыта. В ситуациях, касающихся жизни и смерти, это различия в моральных нормах (ожидания или предпочтения людей относительно того, что субъект должен делать) и моральных суждениях (моральные оценки людей после того, как субъект принимает решение). Сталкиваясь с вопросами жизни и смерти в таких ситуациях, как вождение, закон, лечение и военных ситуациях, люди предпочитают решения, принимаемые людьми, а машинами. Люди требуют, чтобы автомобили без шофера были намного безопаснее, чем сейчас, в то же время переоценивая безопасность собственного вождения [40]. В отношении ошибок машин реакции людей будут сильнее [41]. Люди верят в мораль, и эта вера формирует культурную идентичность. Важно, что по контрасту с другими решениями этические решения глубоко коренятся в эмоциях [42], а машины не обладают полноценным разумом, и это означает, что люди не обязательно могут поддерживать машины в принятии этических решений [43].

Шестой – предельная междисциплинарность. Для того чтобы оценивать машинную этику, необходимо установить контрольные показатели, относящиеся к данной области. Основываясь на мнении экспертов, необходимо создать набор данных, который включает типичные примеры в конкретной области, и оценивать машинную этику на основании этих примеров. Обобщение типичных задач и соответствующих ответов, признанных экспертами в этой области, очень важно.

Таким образом, современная машинная этика все еще сталкивается с рядом проблем. Машины, которые ориентируются на этику, зависят от знаний и данных, предоставляемых людьми. Поэтому без самосознания, понимания значения символов, понимания значения

внешнего физического мира и понимания ценностного смысла социального поведения людей [44], принятие этических решений машинами естественно не обладает характеристиками обычного искусственного интеллекта. Что же касается того, как наделить машины сознанием, проводился ряд исследований. Например, создав самосовершенствующееся интеллектуальное тело, возможно сделать так, чтобы машины обладали «функциональным сознанием» [45]. Однако возможность и путь реализации искусственного сознания – все еще вопрос открытый, и соответствующим исследованиям в области машинной этики предстоит еще долгий путь.

Список литературы

1. Awad, E. Computational ethics / E. Awad // Trends in Cognitive Sciences. – 2022. – Vol. 26. – P. 388–405.

2. Anderson, M. Machine Ethics: Creating an Ethical Intelligent Agent / M. Anderson, S.L. Anderson // AI Magazine. – 2007. – Vol. 28. – P. 15–26.

3. Moor, J.H. The nature, importance, and difficulty of machine ethics / J.H. Moor // IEEE Intelligent Systems. – 2006. – Vol. 21, no. 4. – P. 18–21.

4. Graham, J. Moral foundations theory: The pragmatic validity of moral pluralism / J. Graham // Advances in Experimental Social Psychology. – 2013. – Vol. 47. – P. 55–130.

5. Schein, C. The theory of dyadic morality: Reinventing moral judgment by redefining harm / C. Schein, K. Gray // Personality and Social Psychology Review. – 2018. – Vol. 22. – P. 32–70.

6. Bonnemains, V. Embedded ethics: Some technical and ethical challenges / V. Bonnemains, C. Saurel, C. Tessier // Ethics and Information Technology. – 2018. – Vol. 20. – P. 41–58.

7. Noothigattu, R. A voting-based system for ethical decision-making / R. Noothigattu // Proceedings of the 32nd AAAI Conference on Artificial Intelligence. – Palo Alto. CA: AAAI Press, 2018. – P. 1587–1594.

8. Baum, S.D. Social choice ethics in artificial intelligence / S.D. Baum // AI & SOCIETY. – 2020. – Vol. 35. – P. 165–176.

9. Himma, K.E. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? / K.E. Himma // Ethics and Information Technology Volume. – 2009. – Vol. 11. – P. 19–29.

10. Xiaoping, Chen. The Target, Tasks, and Implementation of Artificial Intelligence Ethics: Six Issues and the Rationale Behind Them / Chen Xiaoping // Philosophical Research. – 2020. – Vol. 9. – P. 79–87 (In Chinese).

11. Liangkang, Ni. Artificial Intelligence: Computing or Thinking? A Review of the Ideological History from «Mathesis Universalis» to «Free Systems» / Ni Liangkang // Zhejiang Social Sciences. – 2023. – Vol. 10. – P. 85–101 (In Chinese).

12. Rao, A.S. Modeling rational agents within a BDI-architecture. In Principles of Knowledge Representation and Reasoning / A.S. Rao, M.P. Georgeff // Proceedings of the second International Conference. – Morgan Kaufmann. – San Mateo, 1991. – P. 473–484.

13. Liao, Beishui. The Jiminy Advisor: Moral Agreements Among Stakeholders Based on Norms and Argumentation / Beishui Liao // Journal of Artificial Intelligence Research. – 2023. – Vol. 77. – P. 737–792.

14. Tolmeijer, S. Implementations in Machine Ethics: A Survey / S. Tolmeijer // ACM Computing Surveys. – 2020. – Vol. 53. – P. 1–38.

15. Awad, E. The Moral Machine experiment / E. Awad // Nature. – 2018. – Vol. 563. – P. 59–64.

16. Awad, E. Universals and variations in moral decisions made in 42 countries by 70,000 participants / E. Awad // Proceedings of the National Academy of Sciences. – 2020. – Vol. 117. – P. 2332–2337.

17. Xu, Yingjin. Artificial Intelligence, Trolley Problem and the Involvement of Cultural-geographical Factors / Yingjin Xu // Philosophical Research. – 2023. – Vol. 2. – P. 96–107 (in Chinese).

18. Rokeach, M. Understanding human values / M. Rokeach. – Simon and Schuster, 2008. – 230 p.

19. Henrich, J. Economic man'in cross-cultural perspective: Behavioral experiments in 15 small-scale societies / J. Henrich // Behavioral and Brain Sciences. – 2005. – Vol. 28. – P. 795–815.

20. Shin, D. User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explain ability / D. Shin // Journal of Broadcasting & Electronic Media. – 2020. – Vol. 64. – P. 541–565.

21. Bryson, J. Standardizing ethical design for artificial intelligence and autonomous systems / J. Bryson, A. Winfield // Computer. – 2017. – Vol. 50. – P. 116–119.

22. Yazdanpanah, V. Reasoning about responsibility in autonomous systems: challenges and opportunities / V. Yazdanpanah // AI & SOCIETY. – 2023. – Vol. 38. – P. 1453–1464.

23. Liao, Beishui. Practical Reasoning About Norms, Values and Preferences / Beishui Liao // Journal of Tsinghua University. – 2019. – Vol. 2. – P. 140–149.

24. Liao, Beishui. On the Crossover Study of the New Generation Artificial Intelligence and Logic / Beishui Liao // Social Sciences in China. – 2022. – Vol. 3. – P. 37–54 (in Chinese).

25. Liao, Beishui. The Jiminy Advisor: Moral Agreements Among Stakeholders Based on Norms and Argumentation / Beishui Liao // Journal of Artificial Intelligence Research. – 2023. – Vol. 77. – P. 737–792.

26. Liao, Beishui. Prioritized Norms in Formal Argumentation / Beishui Liao // Journal of Logic and Computation. – 2019. – Vol. 29. – P. 215–240.

27. Li, Chonghui. Integrating Individual Preferences into Collective Argumentation / Chonghui Li, Beishui Liao // Journal Of Logic and Computation. – 2023. – Vol. 33. – P. 344–369.

28. Lera-Leri, R. Towards Pluralistic Value Alignment: Aggregating Value Systems Through lp-Regression / R. Lera-Leri // Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022). – 2022. – P. 780–788.

29. Awad, E. An approach for combining ethical principles with public opinion to guide public policy / E. Awad // Artificial Intelligence. – 2020. – Vol. 287. – P. 1–19.

30. Yao, Jing. From Instructions to Intrinsic Human Values – A Survey of Alignment Goals for Big Models / Jing Yao // CoRR abs/2308.12014. – 2023.

31. Schramowski, P. Large pre-trained language models contain human-like biases of what is right and wrong to do / P. Schramowski // Nature Machine Intelligence. – 2022. – Vol. 4. – P. 258–268.

32. Dennis, L.A. Verifiable Machine Ethics in Changing Contexts / L.A. Dennis // Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21). – 2021. – P. 11470–11479.

33. Fangzhen, Lin. Using Language Models for Knowledge Acquisition in Natural Language Reasoning Problems / Lin Fangzhen, Shou Ziyi, Chen Chengcai // CoRR abs/2304.01771. – 2023.

34. Chi, Haixiao. A Quantitative Argumentation-based Automated eXplainable Decision System for Fake News Detection on Social Media / Haixiao Chi, Beishui Liao // Knowledge-Based Systems. – 2022. – Vol. 242. – P. 108378.

35. Chi, Haixiao. An Optimized Quantitative Argumentation Debate Model for Fraud Detection in E-commerce Transactions / Haixiao Chi // IEEE Intelligent Systems. – Vol. 36. – 2021. – P. 52–63.

36. Jaakkola, R. Short Boolean Formulas as Explanations in Practice / R. Jaakkola // Proceedings of JELIA. – 2023. – P. 90–105.

37. Hadoux, E. Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuade / E. Hadoux // Argument & Computation. – 2023. – Vol. 14. – P. 109–161.

38. Luo, Jieting. What Do You Care About: Inferring Values from Emotions / Jieting Luo // Proceedings of AAMAS. – 2023. – P. 2289–2291.

39. Vanderelst, D. The dark side of ethical robots / D. Vanderelst, A. Winfield // Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. – 2018. – P. 317–322.

40. Liu, Peng. How Safe Is Safe Enough for Self-Driving Vehicles? / Peng Liu, Run Yang, Zhigang Xu // Risk analysis. – 2019. – Vol. 39. – P. 315–325.

41. Franklin, M. Blaming automated vehicles in difficult situations / M. Franklin // Iscience. – 2021. – Vol. 24. – P. 102252.

42. Gray, K. How to think about emotions and morality: Circles, not arrows / K. Gray // Current Opinion in Psychology. – 2017. – Vol. 17. – P. 41–46.

43. Bigman, E. People are averse to machines making moral decisions / E. Bigman, K. Gray // Cognition. – 2018. – Vol. 181. – P. 21–34.

44. Liu, Xiaoli. Approaches to the Cross-integration of Philosophy and Cognitive Science / Xiaoli Liu // Social Sciences in China. – 2020. – Vol. 9. – P. 23–47 (in Chinese).

45. Ren, Xiaoming. Approaches to the Cross-integration of Philosophy and Cognitive Science / Xiaoming Ren // Social Sciences in China. – 2019. – Vol. 12. – P. 46–61 (in Chinese).

References

1. Awad E. Computational ethics. Trends in Cognitive Sciences, vol. 26, 2022, pp. 388–405.

2. Anderson M., Anderson S. L. Machine Ethics: Creating an Ethical Intelligent Agent. AI Magazine, vol. 28, 2007, pp. 15–26.

3. Moor J. H. The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems, vol. 21, no. 4, 2006, pp. 18–21.

4. Graham J. Moral foundations theory: The pragmatic validity of moral pluralism. Advances in experimental social psychology, vol. 47, 2013, pp. 55–130.

5. Schein C., Gray K. The theory of dyadic morality: Reinventing moral judgment by redefining harm. Personality and Social Psychology Review, vol. 22, 2018, pp. 32–70.

6. Bonnemains V., Saurel C., Tessier C. Embedded ethics: Some technical and ethical challenges. Ethics and Information Technology, vol. 20, 2018, pp. 41–58.

7. Noothigattu R. A voting-based system for ethical decision-making. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. AAAI Press, Palo Alto, CA, 2018, pp. 1587–1594.

8. Baum S. D. Social choice ethics in artificial intelligence. AI & SOCIETY, vol.35, 2020, pp. 165–176.

9. Himma K. E. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? // Ethics and Information Technology Volume, vol.11, 2009, pp. 19–29.

10. Xiaoping Chen The Target, Tasks, and Implementation of Artificial Intelligence Ethics: Six Issues and the Rationale Behind Them. Philosophical Research, vol. 9, 2020, pp. 79–87.

11. Liangkang Ni Artificial Intelligence: Computing or Thinking? A Review of the Ideological History from «Mathesis Universalis» to «Free Systems». Zhejiang Social Sciences. – vol.10, 2023, pp. 85–101.

12. Rao A. S., Georgeff M. P. Modeling rational agents within a BDI-architecture. In Principles of Knowledge Representation and Reasoning. Proceedings of the second International Conference. Morgan Kaufmann, San Mateo, 1991, pp. 473–484.

13. Beishui Liao The Jiminy Advisor: Moral Agreements Among Stakeholders Based on Norms and Argumentation. Journal of Artificial Intelligence Research, vol. 77, 2023, pp. 737–792.

14. Tolmeijer S. Implementations in Machine Ethics: A Survey. ACM Computing Surveys, vol. 53, 2020, pp. 1-38.

15. Awad E. The Moral Machine experiment. Nature, vol. 563, 2018, pp. 59–64.

16. Awad E. Universals and variations in moral decisions made in 42 countries by 70,000 participants. Proceedings of the National Academy of Sciences, vol.117, 2020, pp. 2332–2337.

17. Yingjin Xu Artificial Intelligence, Trolley Problem and the Involvement of Cultural-geographical Factors. Philosophical Research, vol. 2, 2023, pp. 96–107.

18. Rokeach M. Understanding human values. Simon and Schuster, 2008, 230 p.

19. Henrich J. Economic man'in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. Behavioral and Brain Sciences, vol. 28, 2005, pp. 795–815.

20. Shin D. User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explain ability. Journal of Broadcasting & Electronic Media, vol. 64, 2020, pp. 541–565.

21. Bryson J., Winfield A. Standardizing ethical design for artificial intelligence and autonomous systems. Computer, vol. 50, 2017, pp. 116–119.
22. Yazdanpanah V. Reasoning about responsibility in autonomous systems: challenges and opportunities. AI & SOCIETY, vol. 38, 2023, pp. 1453–1464.
23. Beishui Liao Practical Reasoning About Norms, Values and Preferences. Journal of Tsinghua University, vol. 2, 2019, pp. 140–149.
24. Beishui Liao On the Crossover Study of the New Generation Artificial Intelligence and Logic. Social Sciences in China, vol. 3, 2022, pp. 37–54.
25. Beishui Liao The Jiminy Advisor: Moral Agreements Among Stakeholders Based on Norms and Argumentation. Journal of Artificial Intelligence Research, vol. 77, 2023, pp. 737–792.
26. Beishui Liao Prioritized Norms in Formal Argumentation. Journal of Logic and Computation, vol. 29, 2019, pp. 215–240.
27. Chonghui Li, Beishui Liao Integrating Individual Preferences into Collective Argumentation. Journal Of Logic and Computation, vol. 33, 2023, pp. 344–369.
28. Lera-Leri R. Towards Pluralistic Value Alignment: Aggregating Value Systems Through lp-Regression. Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), 2022, pp. 780–788.
29. Awad E. An approach for combining ethical principles with public opinion to guide public policy. Artificial Intelligence, vol. 287, 2020, pp. 1–19.
30. Jing Yao From Instructions to Intrinsic Human Values – A Survey of Alignment Goals for Big Models, CoRR abs/2308.12014. 2023.
31. Schramowski P. Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence, vol. 4, 2022, pp. 258–268.
32. Dennis L. A. Verifiable Machine Ethics in Changing Contexts. Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21), 2021, pp. 11470–11479.
33. Fangzhen Lin, Ziyi Shou, Chengcai Chen Using Language Models For Knowledge Acquisition in Natural Language Reasoning Problems, CoRR abs/2304.01771. 2023.
34. Haixiao Chi, Beishui Liao A Quantitative Argumentation-based Automated eXplainable Decision System for Fake News Detection on Social Media. Knowledge-Based Systems, vol. 242, 2022, pp. 108378.
35. Haixiao Chi An Optimized Quantitative Argumentation Debate Model for Fraud Detection in E-commerce Transactions. IEEE Intelligent Systems, vol. 36, 2021, pp. 52–63.
36. Jaakkola R. Short Boolean Formulas as Explanations in Practice. Proceedings of JELIA, 2023, pp. 90–105.
37. Hadoux E. Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuade. Argument & Computation, vol. 14, 2023, pp 109–161.
38. Jieting Luo What Do You Care About: Inferring Values from Emotions. Proceedings of AAMAS, 2023, pp. 2289–2291.
39. Vanderelst D., Winfield A. The dark side of ethical robots. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 317–322.
40. Peng Liu, Run Yang, Zhigang Xu How Safe Is Safe Enough for Self-Driving Vehicles?. Risk analysis, vol. 39, 2019, pp. 315–325.
41. Franklin M. Blaming automated vehicles in difficult situations. Iscience, vol. 24, 2021, pp. 102-252.
42. Gray K. How to think about emotions and morality: Circles, not arrows. Current Opinion in Psychology, vol.17, 2017, pp. 41–46.
43. Bigman E., Gray K. People are averse to machines making moral decisions. Cognition, vol. 181, 2018, pp. 21–34.
44. Xiaoli Liu Approaches to the Cross-integration of Philosophy and Cognitive Science, Social Sciences in China, vol. 9, 2020, pp. 23–47.
45. Xiaoming Ren Approaches to the Cross-integration of Philosophy and Cognitive Science. Social Sciences in China, vol. 12, 2019, pp. 46-61.