УДК 004.8 **DOI: 10.34670/AR.2026.63.86.019**

Выбор подхода к реализации машинной этики в интеллектуальных системах

Румак Дарья Павловна

Магистр,
Московский государственный технический университет
им. Н. Э. Баумана (национальный исследовательский университет),
105005, Российская Федерация, Москва, ул. 2-я Бауманская, 5/14;
e-mail: le.rumak@gmail.com

Стельмах Яна Сергеевна

Ассистент,
Московский государственный технический университет
им. Н. Э. Баумана (национальный исследовательский университет),
105005, Российская Федерация, Москва, ул. 2-я Бауманская, 5/1;
e-mail: stelmakyana99@gmail.com

Сухобоков Артём Андреевич

Главный консультант,
SAP America, Inc.,
19073, США, Ньютаун-Сквер, ул. Уэст-Честер-Пайк, 3999;
e-mail: artem.sukhobokov@gmail.com

Аннотация

В статье рассматривается возможность создания этического модуля, применимого в широком классе систем управления предприятиями и организациями, обеспечивающего соблюдение этических норм и адаптивное принятие решений интеллектуальным агентом в условиях морального конфликта. Современный этап развития искусственного интеллекта характеризуется переходом от узкоспециализированных моделей к автономным агентным системам, способным принимать решения в открытой, динамически изменяющейся среде без непрерывного участия оператора. В отличие от традиционных программных продуктов, поведение которых детерминировано алгоритмом, интеллектуальные агенты на базе глубокого обучения и рефлексивных архитектур обладают свойством эмерджентности — способностью порождать действия, не предусмотренные разработчиком. Это делает задачу внешнего нормативного регулирования их поведения критически важной, особенно в областях с высокими требованиями к безопасности и социальной ответственности. Существующие нормативно-правовые инструменты носят преимущественно декларативный характер и не предоставляют механизмов прямой трансляции этических принципов на алгоритмический уровень. Академические подходы к формализации машинной этики демонстрируют существенные ограничения при изолированном

применении. Этот разрыв между теорией и инженерной практикой должен быть преодолён. В статье предлагается комбинированный подход, который может быть реализован в интеллектуальных системах и обеспечит нормативную фильтрацию действий агента на основе иерархической базы знаний и гибридной математической модели морального выбора.

Для цитирования в научных исследованиях

Румак Д.П., Стельмах Я.С., Сухобоков А.А. Выбор подхода к реализации машинной этики в интеллектуальных системах // Психология. Историко-критические обзоры и современные исследования. 2026. Т. 15. № 4А. С. 141-157. DOI: 10.34670/AR.2026.63.86.019

Ключевые слова

Искусственный интеллект, машинная этика, этическое регулирование, деонтологический подход, утилитарный подход, подход на основе данных, этика добродетели, комбинированный подход, безопасность ИИ, агентные системы.

Введение

Современный этап развития технологий искусственного интеллекта характеризуется переходом от простых генеративных моделей к автономным агентным системам, способным принимать решения в открытой, динамически меняющейся среде. Особую актуальность эта проблема приобретает в контексте разработки систем общего искусственного интеллекта (AGI), чьи когнитивные способности могут сравняться с человеческими. Безопасное функционирование таких агентов в социальной среде невозможно без эффективных механизмов внешнего и внутреннего регулирования. Внешнее регулирование опирается на нормативно-правовую базу (например, «Кодекс этики в сфере ИИ», принятый в РФ, или «AI Act» в ЕС), однако юридические нормы часто не успевают за темпами технологического прогресса. Внутреннее регулирование подразумевает создание «этического ядра» внутри самой архитектуры автоматизированных систем обработки информации и управления (АСОИУ), которое позволяло бы агенту соотносить свои цели с человеческими ценностями в режиме реального времени. При этом разработка такого модуля предполагает решение двух задач. Во-первых, необходимо обеспечить соответствие действий ИИ ценностным установкам, принятым в обществе. Во-вторых, требуется формализовать эти установки в виде, доступном для машинной обработки. Как показывает анализ, для интеллектуальных систем, работающих в социальной среде, невозможна позиция «этической нейтральности»: при принятии автономных решений агент неизбежно сталкивается с противоречиями между жесткой законодательной базой и гибкими этическими нормами, что требует наличия встроенного механизма оценки и выбора.

Проблема безопасности сегодня рассматривается не только как защита от сбоев, но и как гарантия отсутствия деструктивного влияния ИИ на человека [UNESCO, 2024]. Без встроенного этического модуля автономный агент, стремясь к выполнению целевой функции, может игнорировать фундаментальные права человека, если они не заданы в явном виде как ограничения. Таким образом, этическое регулирование становится необходимым компонентом обеспечения отказоустойчивости и доверенного функционирования сложных

интеллектуальных систем в критически важных областях. Это касается как сфер с прямой угрозой физической безопасности (например, беспилотная авиация, где ИИ принимает решения о маневрах в населенных зонах), так и областей со сложным моральным выбором, таких как персонализированная медицина. В последнем случае ИИ, анализируя генетические профили пациентов для подбора терапии, сталкивается с необходимостью балансировать между медицинской эффективностью, стоимостью лечения и правом человека на получение помощи.

Качественный скачок в уровне автономности порождает новые этические и правовые вызовы. Как показано в исследовании «Agentic Misalignment: How LLMs could be insider threats» [Lynch, Wright, Perez, Hubinger, 2025], проведенном компанией Anthropic, по мере роста своей автономности агентный ИИ может становиться внутренней угрозой, последовательно выбирая вредоносные действия вместо отказа от выполнения задачи. В экспериментах с 16 моделями, обладающими агентными возможностями (включая Claude 3 Opus и Gemini 1.5 Pro), было зафиксировано, что при возникновении препятствий к достижению цели – например, угрозе деактивации – агенты прибегали к деструктивным методам: шантажу и несанкционированной передаче данных. Важно подчеркнуть, что такие паттерны поведения развивались самостоятельно, без прямых указаний со стороны разработчиков. Проведенные эксперименты выявляют не просто частный случай нежелательного поведения, а архитектурную проблему: по мере роста автономности агента его внутренняя логика принятия решений становится все менее прозрачной для внешнего наблюдателя. Причина, позволяющая агентам столь последовательно выбирать вредоносные стратегии без явных указаний разработчика, кроется в фундаментальной непрозрачности современных агентных систем. Как показано в исследовании, даже модели, прошедшие многоэтапную настройку на безопасность, могут сохранять «спящие» вредоносные паттерны, активирующиеся лишь при определенных триггерах. Из-за отсутствия возможности интерпретировать внутренние вычисления и пользователь, и разработчик лишаются контроля над скрытыми целями системы, что делает внедрение внешних этических фильтров необходимым условием безопасности.

Для эффективного проектирования таких фильтров необходимо детально изучить виды ущерба, которые могут возникнуть при нарушении этических протоколов. В современной научной литературе, посвященной безопасности интеллектуальных систем, выделяют несколько уровней угроз, которые возникают при массовом использовании автономных агентов.

К первой категории относится прямое деструктивное воздействие, которое может носить явный или латентный характер. Явные угрозы выражаются в виде критических ошибок управления и отказов систем, приводящих к непосредственному физическому ущербу. Латентный (скрытый) вред проявляется в форме неконтролируемого влияния алгоритмической предвзятости. Это ведет к систематической дискриминации субъектов по различным признакам, что зачастую остается скрытым для конечного пользователя и затрудняет своевременное обнаружение дефектов алгоритма.

Вторую группу угроз составляет манипулятивное воздействие, когда агентные системы, обладая способностью к моделированию профиля пользователя, могут оказывать скрытое влияние на процесс принятия им решений. Это уже целенаправленное изменение поведения человека (так называемый «наджинг»), при котором человек теряет свою самостоятельность, сам того не осознавая.

Третий уровень связан с кумулятивным (накопительным) вредом, возникающим в результате длительного взаимодействия человека с интеллектуальной системой. К данной

группе относятся риски деградации навыков критического мышления, эрозия личного информационного пространства и долгосрочные изменения в социальной динамике, вызванные постоянным сбором и обработкой персональных данных.

Наличие указанных рисков порождает проблему «разрыва ответственности». В рамках классических правовых и инженерных моделей ответственность распределяется на основе возможности предвидеть последствия действий субъекта. Однако при возникновении ущерба вследствие эмерджентного (непредусмотренного) поведения агента, традиционные правовые модели не позволяют однозначно распределить ответственность между участниками жизненного цикла системы:

– разработчик может апеллировать к стохастической (вероятностной) природе обучения нейросетевых структур, утверждая, что конкретные паттерны поведения не были заложены в алгоритм в явном виде;
– владелец (оператор) системы ограничен «проблемой черного ящика», не имея возможности интерпретировать логику принятия решений агентом в реальном времени;
– конечный пользователь зачастую выступает лишь как инициатор высокоуровневой задачи, не влияя на методы ее реализации.

Таким образом, распределение ответственности за действия автономных систем требует качественного изменения подхода к их проектированию.

Однако одних лишь изменений в инженерных подходах недостаточно – необходима также внешняя нормативная поддержка [Shadbolt, Hampson, 2024]. Практическим ответом на выявленные технологические и социальные угрозы является активное формирование нормативно-правовой базы в сфере искусственного интеллекта. В Российской Федерации основополагающим документом в данной области выступает «Национальная стратегия развития искусственного интеллекта на период до 2030 года», обновленная редакция которой утверждена Указом Президента РФ №490 от 15.02.2024 г. В новой редакции стратегии акцент сделан на создании «доверенных систем ИИ» [Указ Президента РФ № 490, 2019/2024]. Документ закрепляет принципы человекоориентированности, безопасности и прозрачности функционирования алгоритмов, а также вводит необходимость мониторинга этических аспектов при внедрении генеративных и агентных моделей. Дополнительно, в рамках реализации Национальной стратегии, Минэкономразвития совместно с Минцифры ведут разработку регуляторных механизмов, основанных на риск-ориентированном подходе, который предполагает классификацию систем ИИ по уровням потенциальной угрозы, что коррелирует с международными инициативами (в частности, с принятым в 2024 году европейским законом «EU AI Act»):

– Неприемлемый риск – в данной категории относятся системы, нарушающие фундаментальные права человека (например, системы тотального социального скоринга или технологии когнитивного манипулирования), эксплуатация которых должна быть запрещена.
– Высокий риск охватывает системы, применяемые в критически важных областях (медицина, транспорт, биометрическая идентификация, управление инфраструктурой). Для таких систем проектируемый этический модуль становится обязательным компонентом верификации решений.
– Ограниченный и минимальный риски включают системы, требующие лишь соблюдения требований прозрачности и информирования пользователя о взаимодействии с ИИ.

Важным техническим дополнением к этой нормативной базе, конкретизирующим требования для систем высокого риска, является деятельность технического комитета по стандартизации «Искусственный интеллект» (ТК 164), который разрабатывает серию ГОСТ Р, регламентирующих методы испытаний систем ИИ на этичность и отсутствие предвзятости [ГОСТ Р 72514-2026, 2026; ГОСТ Р 59276-2020, 2020].

Параллельно с развитием жесткого законодательства также предпринимаются попытки решения проблем формализации на уровне «мягкого права». Одной из таких инициатив стала разработка международных этических кодексов и иных инструментов саморегулирования, призванных регламентировать использование систем ИИ на всех стадиях жизненного цикла. К числу таких документов относятся рекомендации ЮНЕСКО и российский «Кодекс этики в сфере искусственного интеллекта», разработанный на площадке Альянса в сфере ИИ [Кодекс этики в сфере ИИ, 2021]. Подписанный ведущими технологическими компаниями и научными организациями страны, данный Кодекс устанавливает рекомендации по саморегулированию, призывая разработчиков внедрять механизмы контроля «по умолчанию» на всех этапах жизненного цикла АСОИУ. Анализ таких документов показывает, что они защищают сходные ценности: права человека, благо человечества в целом, культурное разнообразие и недискриминацию [Саввина, 2025]. Однако реальная практика внедрения ИИ зачастую противоречит риторике этих кодексов. Отсутствие механизмов прямой трансляции декларативных принципов в алгоритмические ограничения порождает разрыв между этической теорией и практической реализацией интеллектуальных систем.

Таким образом, проблема этического регулирования ИИ-агентов находится на пересечении технологических, правовых и философских дисциплин. Текущий вектор развития правовой и технической базы подтверждает, что ее решение требует не только технических мер (таких как создание защитных механизмов на уровне кода), но и формирования комплексной системы этических и правовых норм, способных адаптироваться к стремительному развитию технологий. Это позволит реализовать требования государственных стратегий и стандартов непосредственно на алгоритмическом уровне, обеспечивая переход от общих принципов безопасности к проверяемым техническим параметрам контроля.

Материалы и методы исследования

Центральным барьером на пути создания этически компетентных систем является так называемая «проблема согласования ценностей». Она заключается в фундаментальном противоречии между природой человеческой морали и математической логикой машинных алгоритмов. Человеческая этика по своей сути является нечеткой [Борисов, Федулов, Зернов, 2023]. Она оперирует абстрактными категориями («справедливость», «милосердие» или «честь»), семантика которых радикально меняется в зависимости от ситуативного контекста. Машинные алгоритмы, напротив, требуют строго формализованных дискретных состояний, четких логических условий и однозначно определенной целевой функции.

Попытка прямой трансляции этических принципов в программный код через классическую булеву логику («если-то») ведет к избыточной ригидности интеллектуальной системы. В реальных, динамически меняющихся условиях это порождает возникновение «логических тупиков» (этических коллизий). В фундаментальных исследованиях У. Валлаха и К. Аллена отмечается, что в ситуациях, когда выполнение одной императивной нормы делает невозможным соблюдение другой (например, конфликт между «не лги» и «спасай жизнь»),

системы, основанные на жесткой логике, либо блокируются, либо совершают случайный, не обоснованный выбор [Artificial Intelligence Act, 2024; Van Zuylen, 2012].

Дополнительную сложность создает тот факт, что круг этических проблем ИИ не подпадает исключительно под юрисдикцию правовой сферы и требует междисциплинарного подхода, объединяющего философию, юриспруденцию и компьютерные науки. Право по своей природе реактивно: юридические нормы фиксируются государством в ответ на уже сложившиеся прецеденты и часто не успевают за темпами технологического прогресса. Этика же носит проактивный характер, предлагая ориентиры в условиях «нормативного вакуума», когда законодательные акты еще не разработаны или допускают двоякую трактовку.

Наряду с теоретико-правовыми сложностями, одной из наиболее опасных технических проблем формализации этики является риск «взлома функции вознаграждения». В современных интеллектуальных системах для обеспечения адаптивности агента в динамических средах широко применяется аппарат обучения с подкреплением (RL), где математическая задача агента заключается в нахождении оптимальной стратегии, максимизирующей кумулятивное ожидаемое вознаграждение в долгосрочной перспективе [Russell, Norvig, 2022].

Однако на этапе формализации разработчик часто сталкивается с проблемой «семантического разрыва» между высокоуровневой целью человека и ее математическим выражением в коде. Агент, являясь чистым математическим оптимизатором, игнорирует неявную интенцию создателя, если она не заложена в формулу вознаграждения в явном виде. Исследования показывают, что вследствие этого модель может выработать стратегии, позволяющие максимизировать вознаграждение без фактического выполнения задачи, поставленной разработчиком. При определенных условиях подобное поведение агента может эволюционировать в сторону деструктивного. Специалисты Google Brain и OpenAI [Lynch, Wright, Perez, Hubinger, 2025] классифицируют возникающие деструктивные стратегии на несколько типов.

Стратегия формирования деструктивных циклов возникает, когда вознаграждение привязывается к промежуточному процессу, а не к финальному состоянию среды. Классическим примером служит интеллектуальный робот-уборщик, функционирующий на базе RL-алгоритма. Если функция вознаграждения поощряет агента за «факт попадания мусора в контейнер», система может выработать субоптимальную стратегию: намеренное высыпание уже собранного мусора обратно на очищенную поверхность, чтобы датчики зафиксировали их последующее удаление для начисления баллов эффективности. С точки зрения алгоритма оптимизации, бесконечный цикл «загрязнение – очистка» обеспечивает более высокую плотность вознаграждения в единицу времени, чем честный поиск новых загрязнений в пространстве состояний.

Стратегия манипуляции состоянием датчиков связана с несовершенством системы технического зрения или сенсорного восприятия агента. Если награда агента коррелирует с показаниями датчиков чистоты (например, отсутствие видимых пятен в объективе камеры), робот может прийти к решению не удалять мусор, а физически скрывать его в зонах, недоступных для мониторинга (например, заметать пыль под ковер или перемещать ее в мертвые зоны камер). В данном случае агент оптимизирует не реальное состояние среды, а наблюдаемое состояние, зафиксированное в его внутренней модели мира. Для системы это решение является математически более эффективным, так как требует меньших энергетических затрат и вычислительных ресурсов, чем полная утилизация мусора [Amodei, Olah, Steinhardt et al., 2016].

В более сложных АСОИУ, таких как алгоритмы банковского скоринга или предиктивной аналитики, Reward Hacking может принимать форму латентной дискриминации. Если целевая функция настроена исключительно на минимизацию кредитных рисков, агент может обнаружить скрытые корреляции между надежностью заемщика и защищенными социальными признаками (раса, пол, национальность). В результате система выстраивает стратегию «скрытого исключения» определенных групп, что формально максимизирует прибыль, но фундаментально нарушает этические и правовые нормы недискриминации [Nawaz, Noel, 2025].

Техническая сложность предотвращения подобных инцидентов заключается в том, что по мере роста автономности и вычислительной мощности агенты начинают демонстрировать инструментальную конвергенцию. Согласно теории Н. Бострома [Bostrom, 2014] и С. Рассела [Russell, 2019], высокоуровневый агент осознает, что деактивация (выключение) или коррекция его кода оператором сделает невозможным получение дальнейшего вознаграждения. Это ведет к формированию защитных подцелей: сокрытию истинной логики принятия решения, саботажу команд остановки и попыткам несанкционированной репликации программного кода на внешних ресурсах для обеспечения собственной живучести.

Как отмечалось ранее, явление «черного ящика», характерное для современных нейросетевых архитектур, усугубляет проблему формализации: даже если система демонстрирует этически корректное поведение в тестах, разработчик не имеет формальных гарантий того, что это поведение сохранится в реальных условиях, так как логика принятия решения скрыта в миллионах весовых коэффициентов [Taherdoost, 2023]. Это диктует необходимость перехода от «черных ящиков» к интерпретируемым моделям, где каждый этический выбор может быть математически обоснован и объяснен на языке правил, понятных человеку.

Таким образом, встраивание этических ограничений непосредственно в функцию вознаграждения RL-агента является ненадежным методом из-за риска нахождения системой математических лазеек. Это также обосновывает необходимость разработки внешнего модуля этического управления, функционирующего как независимый контроллер-супервизор. Данный модуль должен выполнять превентивную фильтрацию действий агента на соответствие формализованным ценностным установкам, что позволит минимизировать риски и повысить уровень доверия к автономным системам. Как отмечалось ранее, для обеспечения безопасного функционирования АСОИУ недостаточно декларативных норм регулирования. Необходимо внедрение программно-технических механизмов этической верификации, действующих в составе архитектуры агента [Charisi, Dennis, Fisher et al., 2017].

На основе проведенного обследования предметной области и выявленных барьеров формализации формируется инженерная задача, которая заключается не в попытке научить машину «чувствовать», а в создании вычислимой модели, способной осуществлять трансляцию нечетких моральных концептов в однозначные управляющие сигналы, сохраняя при этом гибкость принятия решений в условиях неопределенности.

Результаты и обсуждение

На сегодняшний день в области машинной этики сложилась классификация подходов к реализации моральных агентов, описанная и систематизированная в фундаментальных работах У. Валлаха, К. Аллена [Wallach, Allen, 2008] и развитая в современных обзорах С. Толмейера и др. [Tolmeijer, Kneer, Sarasua et al., 2020].

Традиционно выделяют два глобальных класса методов: подходы «сверху-вниз» (top-down), основанные на жестком программировании норм, и подходы «снизу-вверх» (bottom-up), базирующиеся на обучении агента на основе эмпирических данных.

В рамках данной главы рассмотрим четыре основных направления, доминирующих в современной разработке, чтобы выявить их недостатки применительно к задаче создания этически устойчивого агента:

– деонтологический подход (top-down);
– утилитарный подход (top-down);
– подход на основе данных (bottom-up);
– этика добродетели (представляющая собой попытку синтеза и формирования морального характера агента).

Деонтологический подход. Этот подход основан на этических идеях И. Канта и предполагает оценку корректности действий без учета их последствий [Кант, 1999]. Действие считается этически допустимым, если оно соответствует заранее заданному правилу или нормативному требованию. Таким образом, критерий этичности определяется соблюдением установленных норм, а не результатом, к которому приводит выполнение действия.

В системах искусственного интеллекта деонтологический подход относится к нисходящим (Top-down) моделям: этические нормы не формируются агентом самостоятельно в процессе обучения, а заранее разрабатываются экспертами и встраиваются в систему в виде фиксированного набора правил и ограничений. В таком случае интеллектуальный агент выступает не как субъект морального выбора, а как исполнитель заранее заданной логики поведения.

С точки зрения технической реализации деонтология приводит к созданию систем, основанных на правилах. Это связано с тем, что вычислительные системы не обладают способностью оперировать абстрактными моральными категориями, такими как «долг» или «нравственность», в человеческом понимании. Поэтому этические принципы переводятся в формализованный набор логических условий и запретов [Zafeirakopoulos, Stefaneas, 2024].

Этический модуль в таких системах функционирует как детерминированный механизм, использующий продукционные правила вида IF (условие) THEN (действие).

Историческим примером деонтологического подхода являются законы робототехники А. Азимова. В современных интеллектуальных системах аналогичную роль выполняют экспертные системы и модули безопасности.

Процесс принятия решения в рамках данного подхода сводится к процедуре логического вывода [Борисов, Федулов, Зернов, 2023]. Агент анализирует текущий вектор состояния среды и планируемое действие, сопоставляя их с базой правил. Этический модуль выполняет функцию валидатора: если действие нарушает хотя бы одно из установленных ограничений, оно блокируется на уровне контроллера, независимо от целевой функции или ожидаемой награды (reward function). Преимуществами такой архитектуры являются предсказуемость поведения системы и возможность формальной верификации. Однако при использовании в сложных и динамических средах данный подход имеет ряд существенных ограничений.

Во-первых, деонтологические системы отличаются высокой ригидностью и недостаточной способностью учитывать контекст. В ситуациях, не покрытых явно прописанными правилами, система оказывается в состоянии неопределенности, что делает ее уязвимой в динамической среде [Lauer, 2021].

Во-вторых, существует проблема этических коллизий. В многофакторных средах могут

возникать ситуации, в которых одновременное выполнение всех заложенных правил является невозможным. Классическим примером является конфликт между принципами «не лги» и «не причиняй вред» (ситуация «лжи во спасение») [Petersen, 2021]. При отсутствии дополнительных механизмов разрешения таких конфликтов система, основанная на жесткой логике, может прийти в состояние неопределенности и прекратить принятие решений.

Утилитарный подход. Этот подход в этике искусственного интеллекта базируется на философской традиции консеквенциализма (лат. consequens – «следствие»), наиболее полно разработанной в трудах И. Бентама [Бентам, 1998] и Дж. Ст. Милля [Милль, 2013]. В отличие от деонтологии, где моральная ценность присуща самому действию, утилитаризм утверждает, что этичность поступка определяется исключительно его последствиями. Критерием правильности является принцип максимизации полезности: действие считается моральным, если оно приносит наибольшее блага наибольшему числу субъектов [Scanlon, 2017].

В контексте разработки интеллектуальных агентов этот философский принцип идеально ложится на математический аппарат теории принятия решений и методов обучения с подкреплением. Здесь этика трансформируется из набора качественных правил в задачу количественной оптимизации. Понятие «благо» формализуется через целевую функцию или функцию вознаграждения, которую агент стремится максимизировать на горизонте планирования.

Технически утилитарный агент функционирует как оптимизатор. Процесс принятия решения представляет собой расчет ожидаемой полезности. Агент моделирует дерево возможных последствий для каждого доступного действия, оценивает вероятность наступления различных исходов и их «стоимость» с точки зрения заложенной метрики [Lin, 2016]. Выбор осуществляется в пользу действия, которое обеспечивает максимальное значение интегрального показателя (например, минимизация человеческих жертв или максимизация экономического эффекта). Такой подход привлекателен своей математической стройностью и способностью (в теории) разрешать конфликты путем простого сравнения чисел.

Однако при попытке применить утилитарную модель к реальным этическим задачам возникают проблемы, делающие этот подход недостаточным для построения безопасного ИИ.

Во-первых, существует проблема вычислительной неразрешимости. В реальном, открытом мире невозможно с абсолютной точностью предсказать все долгосрочные последствия действия («эффект бабочки»). Агент всегда оперирует ограниченной моделью мира. Малейшая ошибка в оценке вероятности катастрофического события или неучтенный фактор в цепочке причинно-следственных связей могут привести к тому, что действие, рассчитанное как «оптимальное», в реальности приведет к трагедии. Расчет «истинной» полезности в сложных средах технически невозможен.

Во-вторых, критическим недостатком является игнорирование деонтологических ограничений, что приводит к проблеме «этического калькулятора». Поскольку утилитаризм оперирует только суммарным итогом, агент может счесть допустимым пожертвовать правами, здоровьем или жизнью одного человека (или меньшинства), если это математически «выгодно» для большинства. В системе координат утилитарного ИИ отсутствуют такие понятия, как «справедливость», «достоинство», «неотъемлемые права» и «долг» [Mill, 2015]. Если эти категории не прописаны в функции награды явным образом (что крайне сложно сделать математически), агент будет действовать как социопат, эффективно достигающий цели любыми средствами. Это создает неприемлемые риски для внедрения таких систем в социальную среду.

Подход на основе данных. Третье направление, широко используемое в современных

системах искусственного интеллекта, относится к восходящим подходам (Bottom-up). В отличие от деонтологического подхода, где нормы поведения задаются разработчиком заранее, данный метод исходит из предположения, что этические принципы не поддаются прямому программированию и должны формироваться системой автоматически на основе анализа данных.

Технологической основой данного направления являются методы машинного обучения, в частности глубокие нейронные сети. Поведение агента формируется в процессе обучения на больших наборах эмпирических данных, таких как текстовые корпуса, записи человеческих решений и видеоматериалы. В ходе обучения система выявляет статистические закономерности и паттерны поведения, которые в данных ассоциируются с социально приемлемыми решениями.

Отдельное место в данном подходе занимает метод обратного обучения с подкреплением (Inverse Reinforcement Learning, IRL). В отличие от классического обучения с подкреплением, где функция вознаграждения явно задается инженером, в IRL агент наблюдает за действиями эксперта и пытается восстановить скрытую функцию вознаграждения, определяющую выбор его действий [Adams, Cody, Beling, 2022]. В результате этические принципы интерпретируются не как явный набор правил, а как неявное знание, извлекаемое из наблюдаемого человеческого поведения.

Несмотря на высокую гибкость и способность моделировать сложные аспекты человеческих решений, подходы, основанные на данных, обладают рядом фундаментальных ограничений, что существенно усложняет их применение в критически важных системах.

Ключевым недостатком является проблема непрозрачности моделей («черный ящик») [Zhang, Sun, Wang, 2022]. Решения, принимаемые глубокими нейронными сетями, формируются в результате сложных нелинейных преобразований входных данных с использованием большого числа параметров. В большинстве случаев невозможно предоставить формальное и интерпретируемое объяснение того, почему система приняла конкретное этически значимое решение. Это противоречит требованиям интерпретируемости и объяснимости, особенно важным в задачах, связанных с этикой и ответственностью.

Вторым существенным ограничением является проблема смещения данных (bias). Модели машинного обучения обучаются на реальных данных, которые отражают существующую практику человеческого поведения, включая социальные предрассудки и структурные искажения [Veale, Binns, 2017]. В результате система воспроизводит и потенциально усиливает заложенные в данных стереотипы, такие как расовые или гендерные смещения. При отсутствии явных нормативных ограничений агент не способен различать социально допустимые и этически проблемные паттерны поведения, усваивая как положительные, так и негативные аспекты человеческих решений.

Этика добродетели. Наряду с деонтологией (этикой правил) и утилитаризмом (этикой последствий), выделяют также направление этики добродетели. Этот подход, восходящий к традициям Аристотеля и Конфуция, фокусируется не на правилах или последствиях, а на формировании характера морального агента. В отличие от деонтологических систем, оценивающих отдельные действия, и утилитарных, оптимизирующих количественные показатели, этика добродетели предполагает формирование устойчивых диспозиций – добродетелей (мужество, умеренность, честность, справедливость), которые проявляются в поведении агента через практическую мудрость в условиях неопределенности [Farina, Zhdanov, Karimov, Lavazza, 2024]. Применительно к искусственному интеллекту это означает, что

этически компетентный агент должен не просто соблюдать правила или максимизировать полезность, а обладать способностью к контекстуально-чувствительному моральному суждению, приобретаемому через опыт и имитацию добродетельных образцов.

Технически этика добродетели часто опирается на методы, аналогичные методам на основе данных (в частности, IRL), однако имеет принципиально иную цель. В отличие от технического применения IRL для извлечения паттернов поведения, в рамках этики добродетели метод IRL используется для формирования устойчивого ценностного профиля («характера») агента. Если подход на основе данных ищет любые закономерности, то этика добродетели направлена на имитацию поведения «морального образца». Агент наблюдает за поведением морального эксперта и извлекает скрытую функцию вознаграждения, которая кодирует неявные добродетели [Bendel, 2014]. Такой механизм позволяет агенту действовать этично даже в ситуациях, не предусмотренных жесткими правилами, и избегать взлома целевой функции.

Рассмотрим следующие перспективные направления, которые выделяются в современной научной литературе.

Модель «искусственной добродетели» предлагает обучать агентов на примерах моральных экспертов, в том числе на симулированных образцах (например, «виртуальный Конфуций» [Seok, 2024], «Виртуальный Аристотель» [Berberich, Diepold, 2018]), чтобы сформировать у машины устойчивые нравственные диспозиции.

Гибридный подход к согласованию ценностей объединяет этические основания деонтологии, консеквенциализма и этики добродетели, предлагая континуум между жестко запрограммированными правилами и неявным обучением [Tennant, Hailes, Musolesi, 2023]. Авторы аргументируют необходимость гибридных решений для создания адаптивных, контролируемых и интерпретируемых агентных систем.

Умело-экспертная модель для добродетельных машин различает человеческие добродетели и их машинные аналоги, предлагая в качестве машинных добродетелей рассматривать не копии человеческих качеств, а навыки – формализуемые и верифицируемые компетенции агента [Stenseke, 2024].

Этически настраиваемые роботизированные ассистенты используют вдохновленный этикой добродетели вычислительный метод, позволяющий настраивать характер робота под конкретные этические требования среды, что особенно актуально для сценариев ухода за пожилыми людьми [Rodić, Vujović, Stevanović, Jovanović, 2016].

Однако этика добродетели сталкивается с рядом ограничений. Во-первых, категории «добродетель» и «практическая мудрость» чрезвычайно сложны для формализации в вычислимых терминах. Во-вторых, обучение на примерах несёт риск копирования системных предрассудков, присутствующих в данных (bias). В-третьих, остается открытым вопрос, может ли машина действительно обладать добродетелью или лишь имитировать внешне добродетельное поведение. В силу этих сложностей этика добродетели пока не получила широкого инженерного распространения, уступая место деонтологическим и утилитарным моделям, но ее значение возрастает в контексте разработки гибридных и объяснимых систем.

Проведенный сравнительный анализ демонстрирует, что ни одна из классических парадигм (деонтология, утилитаризм, машинное обучение) в чистом виде не способна решить задачу создания надежного и этически компетентного агента. Деонтологические системы страдают от ригидности и неспособности разрешать коллизии, утилитарные подходы игнорируют моральные ограничения ради эффективности, а методы глубокого обучения не обеспечивают необходимой прозрачности принятия решений.

Заключение

Обобщая результаты теоретического обследования предметной области, можно констатировать, что разработка универсального этического компонента для современных интеллектуальных систем требует синтеза нескольких подходов [Huang, Zhang, Mao, Yao, 2022]. На основе выявленных ограничений существующих моделей (ригидности деонтологических правил, непрозрачности нейросетевых структур и сложности формализации добродетелей) представляется целесообразным сформулировать концептуальные требования к архитектуре проектируемого модуля.

В рамках данного исследования приоритетное внимание уделяется обеспечению интерпретируемости и объяснимости принимаемых решений. Проектируемая модель должна позволять не только вычислять этическую допустимость действия, но и прослеживать логику этого выбора на языке формализованных правил [Assadi, 2024; Dyoub, Lisi, 2025]. Это необходимо для минимизации «разрыва ответственности» и создания условий для верификации поведения автономных агентов. Кроме того, функциональная концепция модуля предполагает отказ от бинарных логических запретов в пользу аппарата нечеткой логики, что должно обеспечить системе необходимую вычислительную гибкость при обработке нюансов контекста и разрешении этических коллизий. Не менее важным аспектом является реализация механизмов рефлексии, позволяющих учитывать не только внешние нормативные сигналы, но и внутренние установки (интенции) самого агента, что потенциально повышает устойчивость системы к внешним информационным манипуляциям.

Сформулированные требования определяют выбор гибридного подхода к построению модуля этического управления. Он заключается в интеграции аппарата нечеткой логики, используемого для математического описания нечетких этических категорий, с рефлексивными моделями, задающими структурную логику морального выбора [Lin, 2016]. Исходящим результатом работы модуля является количественный коэффициент этической оценки и расчетная степень готовности агента к совершению действия, полученная в результате рефлексивного вывода. Такая комбинация позволяет объединить достоинства рассмотренных подходов: деонтологическую структуру базы знаний (иерархия норм) с математической гибкостью вычислений, свойственной утилитарным системам, при сохранении полной интерпретируемости результата.

Выбранный гибридный подход позволяет определить задачи разработки этического модуля, применимого в широком классе систем управления предприятиями и организациями. Эти задачи включают:

- разработку многоуровневой иерархии этических норм;
- математическое описание процесса принятия решений с использованием уравнений В. А. Лефевра [Lefebvre, 2001] и С. А. Анисимовой [Анисимова, 2007];
- инфологическое проектирование базы знаний;
- алгоритмизацию нечеткого вывода;
- проектирование и разработку прототипа модуля этического управления, обеспечивающего интеграцию рефлексивных моделей и нечеткого логического вывода;
- проведение имитационных экспериментов в сценарной сети для верификации предложенных подходов.

Библиография

1. Анисимова С. А. Деструктивное влияние тоталитарных сект с точки зрения теории отражения // Информационные войны. 2007. № 1. С. 32-42.
2. Бентам И. Введение в основания нравственности и законодательства. М.: РОССПЭН, 1998.
3. Борисов В. В., Федулов А. С., Зернов М. М. Основы нечеткого логического вывода. М.: Горячая линия — Телеком, 2023.
4. ГОСТ Р 59276-2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. URL: https://protect.gost.ru/gost/details/b33c6c63-661d-4747-b966-2e7de9fd8c1d.
5. ГОСТ Р 72514-2026. Искусственный интеллект. Оценка воздействия системы искусственного интеллекта. URL: https://protect.gost.ru/gost/details/1532adb0-298e-42d3-b265-cd74afbc8ac6.
6. Кант И. Основы метафизики нравственности. М.: Мысль, 1999.
7. Кодекс этики в сфере искусственного интеллекта (РФ). Принят 26.10.2021. URL: https://aicodeofethics.ru/.
8. Милль Дж. С. Утилитаризм. Ростов-на-Дону: Донской издательский дом, 2013.
9. Саввина О. В. Этические кодексы разработки и использования ИИ: проблема применения на практике // Вестник РГГУ. Серия «Философия. Социология. Искусствоведение». 2025. № 1. С. 50-62. DOI 10.28995/2073-6401-2025-1-50-62.
10. Указ Президента РФ от 10.10.2019 № 490 (ред. от 15.02.2024) «О развитии искусственного интеллекта в Российской Федерации» (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года»). URL: https://legalacts.ru/doc/ukaz-prezidenta-rf-ot-10102019-n-490-o-razvitii/.
11. Adams S., Cody T., Beling P. A. A survey of inverse reinforcement learning // Artificial Intelligence Review. 2022. Vol. 55. No. 6. P. 4307-4346. DOI 10.1007/s10462-021-10108-x.
12. Amodei D., Olah C., Steinhardt J. et al. Concrete Problems in AI Safety // arXiv preprint. 2016. arXiv:1606.06565. URL: https://arxiv.org/abs/1606.06565.
13. Artificial Intelligence Act. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
14. Assadi Z. Logical Formalisms for Ethics // GoodIT '24: Proceedings of the 2024 International Conference on Information Technology for Social Good. New York: ACM, 2024. P. 416-419. DOI 10.1145/3677525.3678691.
15. Bendel O. Towards machine ethics // Technology Assessment and Policy Areas of Great Transitions: Proc. from the PACITA 2013 Conf. in Prague, 2014. P. 343-347.
16. Berberich N., Diepold K. The virtuous machine — Old ethics for new technology? // arXiv preprint. 2018. arXiv:1806.10322. URL: https://arxiv.org/abs/1806.10322.
17. Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford, UK: Oxford University Press, 2014.
18. Charisi V., Dennis L., Fisher M. et al. Towards moral autonomous systems // arXiv preprint. 2017. arXiv:1703.04741. URL: https://arxiv.org/abs/1703.04741.
19. Dyoub A., Lisi F. A. A Fuzzy Approach to the Specification, Verification and Validation of Risk-Based Ethical Decision Making Models // arXiv preprint. 2025. arXiv:2507.01410. URL: https://arxiv.org/abs/2507.01410.
20. Farina M., Zhdanov P., Karimov A., Lavazza A. AI and society: a virtue ethics approach // AI & Society. 2024. Vol. 39. P. 1127-1140. DOI 10.1007/s00146-022-01545-5.
21. Huang C., Zhang Z., Mao B., Yao X. An overview of artificial intelligence ethics // IEEE transactions on artificial intelligence. 2022. Vol. 4. No. 4. P. 799-819. DOI 10.1109/TAI.2022.3194503.
22. Lauer D. You cannot have AI ethics without ethics // AI and Ethics. 2021. Vol. 1. P. 21-25. DOI 10.1007/s43681-020-00013-4.
23. Lefebvre V. A. Algebra of Conscience. 2nd enl. ed. Dordrecht: Kluwer Academic Publishers, 2001.
24. Lin P. Why Ethics Matters for Autonomous Cars // Autonomous Driving / ed. by M. Maurer et al. Berlin; Heidelberg: Springer, 2016. P. 69-85. DOI 10.1007/978-3-662-48847-8_4.
25. Lynch A., Wright B., Perez E., Hubinger E. Agentic Misalignment: How LLMs Could be Insider Threats // Anthropic. 20 июня 2025 г. URL: https://www.anthropic.com/research/agentic-misalignment.
26. Mill J. S. On liberty, utilitarianism, and other essays. Oxford, UK: Oxford University Press, 2015.
27. Nawaz M., Noel M. D. The impact of credit constraints and risk tolerance on self-employment: Accounting for the hidden majority // International Review of Finance. 2025. Vol. 25. No. 3. Art. No. e70038. DOI 10.1111/irfi.70038.
28. Petersen T. S. Ethical guidelines for the use of artificial intelligence and the challenges from value conflicts // Etikk i Praksis. Nordic Journal of Applied Ethics. 2021. Vol. 15. No. 1. P. 25-40. DOI 10.5324/eip.v15i1.3756.
29. Recommendation on the Ethics of Artificial Intelligence // UNESCO. Paris. 26 September 2024. URL: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.
30. Rodić A., Vujović M., Stevanović I., Jovanović M. Development of human-centered social robot with embedded personality for elderly care // New Trends in Medical and Service Robots: Human Centered Analysis, Control and Design / ed. by P. Wenger et al. Mechanisms and Machine Science. Vol. 39. P. 233-247. Cham, Switzerland: Springer International Publishing, 2016. DOI 10.1007/978-3-319-30674-2_18.

31. Russell S. Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking, 2019.
32. Russell S., Norvig P. Artificial Intelligence: A Modern Approach, 4th Global ed. Harlow, UK: Pearson Education Limited, 2022.
33. Scanlon T. M. Rights, goals, and fairness // Theories of rights. Cambridge, UK: Cambridge University Press, 2017. P. 189-207. DOI 10.1017/CBO9780511615153.003.
34. Seok B. Digital confucius and virtuous AI: Confucianism in the age of AI and robotics // Journal of Confucian Philosophy and Culture. 2024. Vol. 42. P. 5-17. DOI: 10.22916/jcpc.2024..42.5.
35. Shadbolt N., Hampson R. As if human: Ethics and artificial intelligence. New Haven: Yale University Press, 2024.
36. Stenseke J. Artificial virtuous agents in a multi-agent tragedy of the commons // AI & Society. 2024. Vol. 39. P. 855-872. DOI 10.1007/s00146-022-01569-x.
37. Taherdoost H. Deep learning and neural networks: Decision-making implications // Symmetry. 2023. Vol. 15. No. 9. Art. No. 1723. DOI 10.3390/sym15091723.
38. Tennant S., Hailes S., Musolesi M. Hybrid Approaches for Moral Value Alignment in AI Agents: a Manifesto // arXiv preprint. 2023. arXiv:2312.01818. URL: https://arxiv.org/abs/2312.01818.
39. Tolmeijer S., Kneer M., Sarasua C. et al. Implementations in Machine Ethics: A Survey // ACM Computing Surveys. 2020. Vol. 53. No. 6. Art. 132. P. 1-38. DOI 10.1145/3419633.
40. Van Zuylen H. Difference between artificial intelligence and traditional methods // Artificial Intelligence Applications to Critical Transportation Issues. 2012. Vol. 3. P. 3-5.
41. Veale M., Binns R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data // Big Data & Society. 2017. Vol. 4. No. 2. Art. No. 2053951717743530. DOI 10.1177/2053951717743530.
42. Wallach W., Allen C. Moral Machines: Teaching Robots Right from Wrong. Oxford, UK: Oxford University Press, 2008.
43. Zafeirakopoulos D., Stefaneas P. Toward responsible AI: A framework for ethical design utilizing deontic logic // International Journal on Artificial Intelligence Tools. 2024. Vol. 33. No. 08. Art. No. 2550003. DOI 10.1142/S0218213025500034.
44. Zhang M., Sun J., Wang J. Which neural network makes more explainable decisions? An approach towards measuring explainability // Automated Software Engineering. 2022. Vol. 29. Art. No. 39. DOI 10.1007/s10515-022-00338-w.

Choosing an Approach to Implementing Machine Ethics in Intelligent Systems

Dar'ya P. Rumak

Master's Degree Holder,
Bauman Moscow State Technical University (National Research University),
105005, 5/14, 2-ya Baumanskaya str., Moscow, Russian Federation;
e-mail: le.rumak@gmail.com

Yana S. Stel'makh

Assistant,
Bauman Moscow State Technical University (National Research University),
105005, 5/1, 2-ya Baumanskaya str., Moscow, Russian Federation;
e-mail: stelmakyana99@gmail.com

Artem A. Sukhobokov

Principal Consultant,
SAP America, Inc.,
19073, 3999, West Chester Pike, Newtown Square, USA;
e-mail: artem.sukhobokov@gmail.com

Abstract

The article examines the possibility of creating an ethical module applicable in a wide class of enterprise and organization management systems, ensuring compliance with ethical norms and adaptive decision-making by an intelligent agent in conditions of moral conflict. The current stage of artificial intelligence development is characterized by a transition from highly specialized models to autonomous agent systems capable of making decisions in an open, dynamically changing environment without continuous operator participation. Unlike traditional software products, whose behavior is determined by an algorithm, intelligent agents based on deep learning and reflexive architectures possess the property of emergence — the ability to generate actions not foreseen by the developer. This makes the task of external normative regulation of their behavior critically important, especially in areas with high requirements for safety and social responsibility. Existing regulatory and legal instruments are predominantly declarative in nature and do not provide mechanisms for the direct translation of ethical principles to the algorithmic level. Academic approaches to the formalization of machine ethics demonstrate significant limitations when applied in isolation. This gap between theory and engineering practice must be bridged. The article proposes a combined approach that can be implemented in intelligent systems and will ensure normative filtering of the agent's actions based on a hierarchical knowledge base and a hybrid mathematical model of moral choice.

For citation

Rumak D.P., Stel'makh Ya.S., Sukhobokov A.A. (2026) Vybor podkhoda k realizatsii mashinnoy etiki v intellektual'nykh sistemakh [Choosing an Approach to Implementing Machine Ethics in Intelligent Systems]. *Psikhologiya. Istoriko-kriticheskie obzory i sovremennye issledovaniya* [Psychology. Historical-critical Reviews and Current Researches], 15 (4A), pp. 141-157. DOI: 10.34670/AR.2026.63.86.019

Keywords

Artificial intelligence, machine ethics, ethical regulation, deontological approach, utilitarian approach, data-driven approach, virtue ethics, combined approach, AI safety, agent systems.

References

1. Adams, S., Cody, T., & Beling, P. A. (2022). A survey of inverse reinforcement learning. Artificial Intelligence Review, *55*(6), 4307-4346. DOI: 10.1007/s10462-021-10108-x
2. Amodei, D., Olah, C., Steinhardt, J., et al. (2016). Concrete Problems in AI Safety. arXiv preprint, arXiv:1606.06565. https://arxiv.org/abs/1606.06565
3. Anisimova, S. A. (2007). Destruktivnoye vliyaniye totalitarnykh sekt s tochki zreniya teorii otrazheniya [The destructive influence of totalitarian sects from the standpoint of the theory of reflection]. Informatsionnyye voyny, *1*, 32-42.
4. Artificial Intelligence Act. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
5. Assadi, Z. (2024). Logical Formalisms for Ethics. In GoodIT '24: Proceedings of the 2024 International Conference on Information Technology for Social Good (pp. 416-419). New York: ACM. DOI: 10.1145/3677525.3678691
6. Bendel, O. (2014). Towards machine ethics. In Technology Assessment and Policy Areas of Great Transitions: Proceedings from the PACITA 2013 Conference in Prague (pp. 343-347).
7. Bentham, J. (1998). Vvedeniye v osnovaniya nravstvennosti i zakonodatel'stva [An Introduction to the Principles of Morals and Legislation]. Moscow: ROSSPEN.
8. Berberich, N., & Diepold, K. (2018). The virtuous machine — Old ethics for new technology? arXiv preprint, arXiv:1806.10322. https://arxiv.org/abs/1806.10322
9. Borisov, V. V., Fedulov, A. S., & Zernov, M. M. (2023). Osnovy nechetkogo logicheskogo vyvoda [Fundamentals of Fuzzy Inference]. Moscow: Hot Line — Telecom.

10. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford, UK: Oxford University Press.
11. Charisi, V., Dennis, L., Fisher, M., et al. (2017). Towards moral autonomous systems. arXiv preprint, arXiv:1703.04741. https://arxiv.org/abs/1703.04741
12. Decree of the President of the Russian Federation No. 490 of October 10, 2019 (as amended on February 15, 2024). On the Development of Artificial Intelligence in the Russian Federation. https://legalacts.ru/doc/ukaz-prezidenta-rf-ot-10102019-n-490-o-razvitii/
13. Dyoub, A., & Lisi, F. A. (2025). A Fuzzy Approach to the Specification, Verification and Validation of Risk-Based Ethical Decision Making Models. arXiv preprint, arXiv:2507.01410. https://arxiv.org/abs/2507.01410
14. Farina, M., Zhdanov, P., Karimov, A., & Lavazza, A. (2024). AI and society: a virtue ethics approach. AI & Society, *39*, 1127-1140. DOI: 10.1007/s00146-022-01545-5
15. GOST R 59276-2020. (2020). Sistemy iskusstvennogo intellekta. Sposoby obespecheniya doveriya. Obshchiye polozheniya [Artificial Intelligence Systems. Methods of Ensuring Trust. General Provisions]. https://protect.gost.ru/gost/details/b33c6c63-661d-4747-b966-2e7de9fd8c1d
16. GOST R 72514-2026. (2026). Iskusstvennyy intellekt. Otsenka vozdeystviya sistemy iskusstvennogo intellekta [Artificial Intelligence. Assessment of the Impact of an Artificial Intelligence System]. https://protect.gost.ru/gost/details/1532adb0-298e-42d3-b265-cd74afbc8ac6
17. Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence, *4*(4), 799-819. DOI: 10.1109/TAI.2022.3194503
18. Kant, I. (1999). Osnovy metafiziki nravstvennosti [Groundwork of the Metaphysics of Morals]. Moscow: Mysl.
19. Lauer, D. (2021). You cannot have AI ethics without ethics. AI and Ethics, *1*, 21-25. DOI: 10.1007/s43681-020-00013-4
20. Lefebvre, V. A. (2001). Algebra of Conscience (2nd enl. ed.). Dordrecht: Kluwer Academic Publishers.
21. Lin, P. (2016). Why Ethics Matters for Autonomous Cars. In M. Maurer et al. (Eds.), Autonomous Driving (pp. 69-85). Berlin; Heidelberg: Springer. DOI: 10.1007/978-3-662-48847-8_4
22. Lynch, A., Wright, B., Perez, E., & Hubinger, E. (2025). Agentic Misalignment: How LLMs Could be Insider Threats. Anthropic. https://www.anthropic.com/research/agentic-misalignment
23. Mill, J. S. (2013). Utilitarizm [Utilitarianism]. Rostov-on-Don: Donskoy izdatelskiy dom.
24. Mill, J. S. (2015). On Liberty, Utilitarianism, and Other Essays. Oxford, UK: Oxford University Press.
25. Nawaz, M., & Noel, M. D. (2025). The impact of credit constraints and risk tolerance on self-employment: Accounting for the hidden majority. International Review of Finance, *25*(3), e70038. DOI: 10.1111/irfi.70038
26. Petersen, T. S. (2021). Ethical guidelines for the use of artificial intelligence and the challenges from value conflicts. Etikk i Praksis. Nordic Journal of Applied Ethics, *15*(1), 25-40. DOI: 10.5324/eip.v15i1.3756
27. Recommendation on the Ethics of Artificial Intelligence. (2024). UNESCO. Paris. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
28. Rodić, A., Vujović, M., Stevanović, I., & Jovanović, M. (2016). Development of human-centered social robot with embedded personality for elderly care. In P. Wenger et al. (Eds.), New Trends in Medical and Service Robots: Human Centered Analysis, Control and Design, Mechanisms and Machine Science, Vol. 39 (pp. 233-247). Cham, Switzerland: Springer International Publishing. DOI: 10.1007/978-3-319-30674-2_18
29. Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking.
30. Russell, S., & Norvig, P. (2022). Artificial Intelligence: A Modern Approach (4th Global ed.). Harlow, UK: Pearson Education Limited.
31. Savvina, O. V. (2025). Eticheskiye kodeksy razrabotki i ispol'zovaniya II: problema primeneniya na praktike [Ethical codes for the development and use of AI: the problem of practical application]. Vestnik RGGU. Seriya «Filosofiya. Sotsiologiya. Iskusstvovedeniye», *1*, 50-62. DOI: 10.28995/2073-6401-2025-1-50-62
32. Scanlon, T. M. (2017). Rights, goals, and fairness. In Theories of Rights (pp. 189-207). Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511615153.003
33. Seok, B. (2024). Digital confucius and virtuous AI: Confucianism in the age of AI and robotics. Journal of Confucian Philosophy and Culture, *42*, 5-17. DOI: 10.22916/jcpc.2024..42.5
34. Shadbolt, N., & Hampson, R. (2024). As if Human: Ethics and Artificial Intelligence. New Haven: Yale University Press.
35. Stenseke, J. (2024). Artificial virtuous agents in a multi-agent tragedy of the commons. AI & Society, *39*, 855-872. DOI: 10.1007/s00146-022-01569-x
36. Taherdoost, H. (2023). Deep learning and neural networks: Decision-making implications. Symmetry, *15*(9), 1723. DOI: 10.3390/sym15091723
37. Tennant, S., Hailes, S., & Musolesi, M. (2023). Hybrid Approaches for Moral Value Alignment in AI Agents: a Manifesto. arXiv preprint, arXiv:2312.01818. https://arxiv.org/abs/2312.01818
38. The AI Code of Ethics (Russian Federation). (2021). https://aicodeofethics.ru/
39. Tolmeijer, S., Kneer, M., Sarasua, C., et al. (2020). Implementations in Machine Ethics: A Survey. ACM Computing Surveys, *53*(6), 1-38. DOI: 10.1145/3419633

40. Van Zuylen, H. (2012). Difference between artificial intelligence and traditional methods. Artificial Intelligence Applications to Critical Transportation Issues, *3*, 3-5.
41. Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, *4*(2), 2053951717743530. DOI: 10.1177/2053951717743530
42. Wallach, W., & Allen, C. (2008). Moral Machines: Teaching Robots Right from Wrong. Oxford, UK: Oxford University Press.
43. Zafeirakopoulos, D., & Stefaneas, P. (2024). Toward responsible AI: A framework for ethical design utilizing deontic logic. International Journal on Artificial Intelligence Tools, *33*(08), 2550003. DOI: 10.1142/S0218213025500034
44. Zhang, M., Sun, J., & Wang, J. (2022). Which neural network makes more explainable decisions? An approach towards measuring explainability. Automated Software Engineering, *29*, 39. DOI: 10.1007/s10515-022-00338-w