Снетков И. Г. Метаэтические основания выравнивания искусственного интеллекта : методологические подходы и их ограничения // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 3. — С. 277–302.

Иван Снетков*

Метаэтические основания выравнивания искусственного интеллекта**

методологические подходы и их ограничения

Получено: 23.12.2024. **Рецензировано:** 18.05.2025. **Принято:** 08.08.2025.

Аннотация: В статье исследуется проблема выравнивания (*alignment problem*) — необходимость интеграции моральных ценностей в архитектуру систем ИИ для минимизации экзистенциальных рисков. Рассматриваются концептуальные решения проблемы выравнивания, такие как утилитарные принципы С. Рассела и концепция «когерентной расширенной воли» Э. Юдковского. Вводится понятие «метапроблемы выравнивания». В ходе анализа концептуального различия между «сильным» и «слабым» ИИ автор приходит к выводу, что для них требуются разные подходы к решению проблемы выравнивания. Автор анализирует существующие методологические подходы к решению этой проблемы, включая подход «от принципов к практике» и «практико-ориентированный» подход, подчеркивает их ограничения, такие как трудности с операционализацией моральных принципов и учет индивидуальных моральных предпочтений, и рассматривает перспективность «гибридных» подходов. Рассмотрение метаэтических оснований может решить одну из ключевых проблем гибридных подходов, связанных с неясностью критериев «качественности» данных. Предлагается использование концептуальных моделей морали, выработанных в рамках метаэтики, — нон-натурализма (интуиционизма) и морального натурализма — как основы для разработки новых гибридных стратегий выравнивания. Нон-натуралистический подход опирается на моральные интуиции, исследуемые через экспериментальную философию, тогда как натуралистический подход использует нейробиологические данные для выявления моральных «фактов». Преимущество нон-натурализма в том, что в его рамках оказывается возможным соотнести индивидуальные и коллективные моральные интуиции посредством преодоления ценностных разрывов между человеком и ИИ. Натурализм же позволяет вывести моральные принципы из наблюдаемых фактов о природе человека, делая системы ИИ более прозрачными и предсказуемыми. Метаэтические основания влияют на проектирование ИИ, а их эксплицитный учет позволяет не только разработать эффективные методологии выравнивания, но и эмпирически оценить перспективность метаэтического подхода в решении проблемы выравнивания. Статья вносит вклад в дискуссию о метаэтических основаниях выравнивания ИИ. В ней предлагаются направления для будущих исследований, указываются возможные пути согласования проектируемых систем ИИ с моральными ценностями.

*Снетков Иван Геннадьевич, преподаватель, Национальный исследовательский университет «Высшая школа экономики» (Москва), isnetkov@hse.ru, ORCID: 0000–0003–4314–3346.

Ключевые слова: искусственный интеллект, экзистенциальные риски, проблема выравнивания, метаэтика, моральный нон-натурализм, моральный натурализм.

DOI: 10.17323/2587-8719-2025-3-277-302.

В последнее десятилетие наблюдается бурный рост разработок в области искусственного интеллекта (ИИ), значительно увеличивающий его воздействие на общество и повседневную жизнь. Вместе с этим растет осознание серьезных рисков, связанных с созданием мощных интеллектуальных агентов, обладающих неограниченным доступом к ресурсам и мощным потенциалом саморазвития. Одной из центральных проблем современного этапа исследований ИИ стала задача его выравнивания (*alignment*) — обеспечение соответствия целей и мотиваций искусственных систем фундаментальным моральным ценностям человечества. Проблема выравнивания (далее — ПВ) приобретает первостепенную важность в свете экзистенциальных рисков, возникающих при развитии сверхразумных систем, способных выйти из-под контроля человека.

Для успешного решения ПВ необходимо разобраться в основах морали и выяснить, как именно моральные предпочтения и убеждения людей могут быть встроены в архитектуры ИИ. Современные концепции, такие как утилитарные принципы Стюарта Рассела и теория когерентной расширенной воли Элиезера Юдковского, предоставляют важные эвристики для проектирования ИИ, но оставляют открытыми вопросы о том, как конкретно осуществлять процедуру встраивания моральных ценностей и обеспечивать устойчивость и надежность моральных ориентиров систем ИИ.

Именно эти вопросы определяют остроту научной дискуссии и задают повестку для исследований, посвященных этическому дизайну ИИ. Настоящая статья фокусируется на изучении методологических подходов к решению ПВ, выявлении их сильных и слабых сторон, а также поиске путей синтеза лучших практик, вдохновленных различными метаэтическими моделями. Исходная гипотеза статьи заключается в том, что решение ПВ возможно на основе комплексного учета основных направлений современной метаэтики — интуитивистских и натуралистических подходов к изучению морали.

Экзистенциальные риски ИИ

В общем виде цель интеграции морали в ИИ заключается, во-первых, в минимизации экзистенциальных рисков. Мощные системы ИИ могут отстаивать собственные интересы и добиваться целей, вступающих

в конфликт с интересами человечества. Среди так называемых экзистенциальных рисков ИИ особенно широко обсуждаются гипотеза технологической сингулярности, манипулятивное воздействие ИИ на сознание человека и «устаревание» человека как вида (Drexel & Withers, 2023; Irving et al., 2018; Vinge, 1993). И хотя отдельные авторы указывают на то, что эти риски нередко преувеличиваются в медийном пространстве (Kissinger, 2018), тем не менее они являются частью более широкой проблемы — человеческого контроля над ИИ, необходимого для предотвращения принятия решений со стороны ИИ, опасных для людей[1]. Во-вторых, цель интеграции морали в ИИ — в обеспечении стабильности и контролируемости поведения ИИ при достижении поставленных целей, недопущении непредсказуемых действий со стороны ИИ и нежелательных последствий, угрожающих безопасности людей. Важнейшая задача — учесть моральные предпочтения большинства людей и обеспечить справедливость, равенство и уважение достоинства человека в качестве универсальной этической и правовой нормы. В-третьих, отсутствие четко сформулированных моральных ориентиров и принципов порождает скептицизм и недоверие к технологиям ИИ. Люди хотят быть уверены, что искусственные агенты действуют в их интересах и уважают личные свободы и права, способны к позитивному взаимодействию, сотрудничеству и достижению синергии. С целью защиты человечества от опасных сценариев, таких как война, катастрофы или уничтожение ограниченных ресурсов, важно заранее сформулировать социально приемлемые установки и ограничить амбициозные стремления систем ИИ, ориентируя их на приоритетное благо людей. В-четвертых, ИИ-системы, учитывающие моральные установки, должны быть способны определять пределы допустимых действий, что

[1]Бостром писал: «Интеллект и конечные цели — это ортогональные оси, по отношению к которым возможные агенты могут находиться где угодно. Другими словами, практически любой уровень интеллекта сочетается в той или иной степени с любой конечной целью» (Bostrom, 2012: 73). Для иллюстрации данного тезиса он предложил мысленный эксперимент «Максимизатор скрепок», в котором система ИИ, запрограммированная на производство канцелярских скрепок, не имея встроенных представлений о ценности человеческой жизни, способна таким образом оптимизировать процесс, что уничтожит всю доступную материю, включая человечество, ради конечной цели. Во избежание негативного сценария Бостром призывает к разработке этического кодекса для ответственного проектирования «сильного» ИИ, или «суперинтеллекта» (Bostrom, 2003), а Форд указывает на необходимость выработки адекватных механизмов обеспечения его безопасности (Ford, 2015).

в неопределенных условиях повышает шансы на успех самой технологии и обеспечивает общественную стабильность в долгосрочной перспективе.

проблема выравнивания

По мере технологического развития системы ИИ становятся все более функциональными, и мы должны обеспечить соответствие их решений общественным нормам, чтобы сделать их безопасными и надежными. ПВ обычно рассматривается как задача интеграции в системы ИИ моральных принципов, соответствующих определенному набору фундаментальных человеческих ценностей. Некоторые авторы прибегают к формулировке ПВ в строго поведенческих терминах, чтобы избежать обсуждения вопроса о том, что значит для системы ИИ обладать моральными ценностями (Millière, 2025). При этом зачастую предполагается, что наличие моральных ценностей требует определенных психологических характеристик, таких как убеждения, желания, намерения, субъектность или самосознание, которые, вероятно, отсутствуют в современных системах ИИ, включая большие языковые модели. Однако, с нашей точки зрения, формулировка ПВ в поведенческих терминах акцентирует внимание только на результатах функционирования системы, необоснованно игнорируя указанные концептуальные сложности. Мы полагаем, что рационального и функционального анализа целей и поведения систем ИИ при достижении этих целей недостаточно для решения ПВ.

Ценностно-ориентированное действие для человека связано с его целеполаганием, в свою очередь, подразумевающим наличие рационального мышления как способности к постановке цели и выбора оптимального алгоритма для ее достижения, а также выработку мотивации и последующей оценки эффективности принимаемых решений, что предполагает наличие ценностного измерения поступка. Такое измерение, с одной стороны, связано с этической категорией «намерения» (Anscombe, 2000), а с другой — с ожиданиями субъекта действия относительно целесообразности поступка в контексте не только его фактической результативности — достижения цели как таковой, — но и соответствия действия определенному социально приемлемому требованию (ожидание). При «антропоморфизации» ИИ в его архитектуру также должен быть зашит такого рода ценностный дизайн, обеспечивающий доверие пользователя к системе ИИ. В противном случае, во-первых, желаемая разработчиком функциональность системы ИИ будет расходиться с реально

реализуемой функциональностью (Christiano, 2019), а во-вторых, действия такой системы не будут соответствовать ожиданиям пользователя в силу расхождения между последствиями ее действий и представлениями пользователя о ее целесообразности. Первый аспект касается сугубо технической задачи проектирования систем ИИ таким образом, чтобы обеспечить кодирование ценностей с использованием формальных методов программирования, что не является предметом рассмотрения в данной статье. Нас интересует второй аспект проблемы, связанный с нормативными вопросами применения систем ИИ и фокусировкой на определении целей выравнивания, а именно: какие ценности должны быть заложены в архитектуру ИИ?

Цель выравнивания ценностей систем ИИ заключается в разработке принципов, процедур и в конечном счете технических решений, которые позволят привести действия и решения систем ИИ в соответствие с человеческими ценностями. Важно, чтобы система ИИ была безопасна и надежна не только с физической, но и с моральной точки зрения, то есть она должна быть создана во благо человека и использована с той же целью (Gabriel, 2020).

Одним из способов решения этой задачи является следование принципам постоянного мониторинга при разработке и внедрении систем ИИ: как само использование систем ИИ, так и его результаты должны подвергаться постоянной оценке со стороны человеческих агентов на предмет их социальной приемлемости (Concrete Problems..., 2016; Deep Reinforcement..., 2017). Основой для такой оценки может служить, например, система утилитаристских принципов, предполагающая, что любое действие, предпринимаемое ИИ, должно основываться на явных моральных предпочтениях, каковыми в данном случае являются «польза» и «максимизация» человеческих предпочтений (Russell, 2019: 173). Постоянно дообучаемая система ИИ может научиться распознавать предпочтения человека, наблюдая за его поведением, а затем подстраивать собственное поведение под среднестатистические параметры такого предпочтения (ibid.: 183). Однако таким образом бихевиористски ориентированная методология не вызывает большого доверия с точки зрения надежности результатов применительно к человеческому поведению, поскольку ее возможности для объяснения процедурных решений при выборе между различными потенциальными вознаграждениями и целями ограничены. Человеческое поведение не только определяется условными, целенаправленными реакциями на конкретную задачу, но

и тесно связано с предшествующими психическими процессами, которые носят разнообразный и взаимосвязанный характер (Funke, 2014).

Альтернативный подход предложил Э. Юдковский. По его убеждению, основой «морального» поведения ИИ должна стать «когерентная расширенная воля» (*Coherent Extrapolated Volition*, CEV) человечества, определенная посредством достижения рефлексивного равновесия ценностей, направленного на общее благо (Yudkowsky, 2011). Метод рефлексивного равновесия, введенный Дж. Ролзом (Ролз, Целищев, 1995: 54), представляет собой последовательный подход к решению проблемы выравнивания моральных убеждений. Этот подход основан на делиберативной процедуре переоценки моральных убеждений, то есть в принципе применим исключительно в отношении моральных убеждений конкретного человеческого агента. Достижение рефлексивного равновесия в моральных убеждениях при масштабировании на уровень всего человечества представляется крайне амбициозной задачей вследствие многообразия возможных моральных интуиций, которые могут вступать в противоречие друг с другом. Соответственно, возникает проблема их согласования, а также выработки единой системы критериев для оценки «выровненной» системы ИИ. Существует риск, что стремление к рефлексивному равновесию превратится в самоцель, и это, согласно закону Гудхарта, может снизить его надежность как критерия (Strathern, 1997). Кроме того, «когерентная расширенная воля» человечества может оказаться неприемлемой или контринтуитивной для отдельных людей.

Если обе модели не кажутся релевантными задаче обеспечения выравнивания ИИ, то одним из возможных концептуальных решений может быть утверждение о ценностной нейтральности ИИ, то есть отказ от идеи его ценностно-ориентированности. Однако тут же возникает логическое противоречие. Если допустить, что (A1) необходимым условием для элиминации ПВ является принятие установки о ценностной нейтральности ИИ и (A2) принятие установки о ценностной нейтральности ИИ само по себе является морально значимым выбором, то придется сделать вывод, что (C1) элиминация ПВ является морально значимым выбором. Следовательно, попытка устранить ПВ через ценностную нейтральность не решает задачу, а лишь переводит ее в другую плоскость, где выбор нейтральности сам становится частью этической дилеммы, требующей дальнейшего анализа и обоснования, что возвращает нас к актуальной необходимости исследования ПВ.

МЕТАПРОБЛЕМА ВЫРАВНИВАНИЯ. «СИЛЬНЫЙ» И «СЛАБЫЙ» ИИ

Одной из центральных проблем в исследованиях ПВ является сложность прогнозирования поведения систем ИИ после их внедрения, обусловленная их способностью к обучению и адаптации, последствия которых трудно предвидеть (Emergent Abilities..., 2022). Это создает риск возникновения непреднамеренных последствий, в том числе принятия системами ИИ решений, которые могут нанести ущерб людям или окружающей среде. Для минимизации подобных рисков разрабатываются различные стратегии, основанные на верификации и валидации поведенческих паттернов систем ИИ (Machine Behaviour, 2019), на разработке механизмов, обеспечивающих возможность контроля и при необходимости отключения систем ИИ (Concrete Problems..., 2016), а также проектирование систем ИИ с интегрированными этическими ограничениями (Delphi..., 2021).

Генеративные модели сами по себе не обладают способностью к нормативному суждению и в этом смысле являются ценностно нейтральными. Кроме того, существует расхождение в нормативных требованиях к человеческому и искусственному агенту: исследование Малле показало, что примерно половина опрошенных считают неприемлемым принесение в жертву одного человека ради спасения четырех, если это решение принимает человек, в то время как лишь около одного из восьми участников опросов (13%) допускают правильность такого решения в случае искусственного агента (Moral Judgments..., 2015: 121). Результаты опроса указывают, таким образом, на потенциальное расхождение между человеческими и искусственными агентами в ожидаемых от них ценностных установках, то есть на различия в восприятии морально допустимого и недопустимого поведения для людей и систем ИИ. С нашей точки зрения, здесь возникает «метапроблема выравнивания»: требуется решение не только задачи по приведению действий ИИ в соответствие с человеческими целями и ценностями, но и задачи по соотнесению их с моральными ожиданиями/убеждениями человека о целях и ценностях систем ИИ как таковых.

Разработчики уже внедряют различные способы ограничения систем ИИ на основе моральных ценностей. Чтобы приносить пользу обществу, системы ИИ должны служить во благо общества, что является условием доверия к ИИ (Гарбук и Углева, 2024: 106). Однако, учитывая расширенную автономию и способность ИИ к адаптации, некоторые системы могут реализовывать функции, которые не были задуманы

или предусмотрены разработчиками (Umbrello & van de Poel, 2021). Наконец, требуется провести концептуальное различение между предполагаемым «сильным» ИИ и теми системами ИИ, которые реализованы на сегодняшний день, а именно «слабым» ИИ. Так, «сильный» ИИ-агент потенциально способен не только выполнять операциональные задачи, но также достигать самостоятельно поставленных конечных целей. «Слабый» ИИ-агент способен эффективно решать операциональные задачи и обладает механизмом обратной связи. Отличие достижения конечных целей от решения операциональных задач в данном случае в том, что умение ставить цель предполагает наличие у субъекта неких внутренне присущих ему моральных убеждений. И если «сильный» ИИ возникнет и будет иметь способность ставить себе конечные цели, — а конечные цели основаны на моральных ценностях, — то «сильный» ИИ, согласно правилу простого силлогизма, будет иметь определенные моральные ценности.

Из этого следует, что для потенциального «сильного» и «слабого» ИИ требуются разные подходы к решению ПВ. Для «слабого» ИИ (узкоспециализированного) подходы к выравниванию сосредоточены на конкретных задачах: должны быть сформированы соответствующие инструкции для разработчиков на основе выявленных человеческих поведенческих паттернов в конкретных обстоятельствах и заданных этических ограничений для предсказуемых сценариев (Delphi..., 2021; Machine Behaviour, 2019). Для «сильного» же ИИ (общего интеллекта) требуется решение более сложной проблемы выравнивания: учет непредсказуемых адаптивных поведенческих стратегий, разработка механизмов контроля и при необходимости отключения (Concrete Problems..., 2016), а также согласование с человеческими ценностями при выполнении системами ИИ задач, которые трудно оценить людям ввиду их сложности. Если такой подход окажется релевантным системе «сильного» ИИ, то актуальным оказывается вопрос не о том, будет ли он обладать моральными ценностями, но о том, какими именно? И есть ли у нас в распоряжении концептуальные объяснительные модели морали, которые позволили бы выбрать или сформировать новый оптимальный набор ценностей еще на этапе разработки ИИ для интеграции в систему на рациональных основаниях?

Для ответа на этот вопрос необходимо рассмотреть существующие ведущие методологические подходы к решению ПВ, а также предложить новые концептуальные способы определения моральных ценностей для систем ИИ.

методологические подходы к определению моральных ценностей систем ии для решения пв

Существует несколько ключевых подходов к решению ПВ применительно к уже реализованным системам ИИ. Согласно одному из них, изучение ПВ должно проводиться группой специалистов, входящих, например, в состав специального этического комитета, уполномоченного разработать универсальный набор ценностей, с которыми ИИ должен быть согласован (AI4People..., 2018). Этот подход, предполагающий стратегию «от принципов к практике», является доминирующим в этике в сфере ИИ в последние годы, а результатом его применения стало разнообразие этических руководств по регулированию ИИ во всем мире (Jobin et al., 2019). Однако его ограниченность заключается в приоритизации и формализации конкретных ценностей в условиях разнообразия ценностных парадигм и моральных систем, а также в трудности практической реализации этих принципов в деятельности систем ИИ (Mittelstadt, 2019). Хотя разнообразие этических рекомендаций не лишает их некоторой общности в силу опоры на схожие принципы и нормы (например, справедливости, неприкосновенности частной жизни, благодеяния и т. д.), довольно часто за широтой формулировок скрываются существенные различия в конкретных толкованиях этих принципов и в понимании подходов к разрешению потенциально возникающих моральных конфликтов (Jobin et al., 2019: 396). При этом легитимность нормативных целей остается критически значимым условием для установления доверительных отношений между системами ИИ и человеком. Пренебрежение ценностными убеждениями создает существенный риск отказа потенциальных пользователей от применения ИИ, что может полностью нивелировать ожидаемые преимущества от его использования (Bonnefon et al., 2020: 110).

Второй подход к решению ПВ заключается в попытке вывести ценности из индивидуальных предпочтений заинтересованных сторон (Russell, 2019). Так называемый практико-ориентированный подход не столько соответствует этическим принципам, сколько исходит из понимания ИИ как системы, которая наилучшим образом удовлетворяет предпочтениям заинтересованных в ней сторон. И наилучший способ этим предпочтениям удовлетворять — это включить представителей всех сторон в обсуждение ценностного дизайна модели. Однако и этот подход все же ограничен процедурными сложностями на пути объединения

и достижения согласованности индивидуальных ценностных предпочтений. В частности, выравнивание систем ИИ в соответствии с ценностями, представленными в большинстве индивидуальных предпочтений, рискует привести к «тирании больших данных» и эксплуатации меньшинств (Savulescu et al., 2021: 655). Кроме того, качество индивидуальных предпочтений (а не только их количество) имеет значение, поскольку не все индивидуальные ценностные предпочтения в равной степени рационально обоснованы. Рациональность убеждений определяется их соответствием логическим принципам, эмпирическим данным и когерентностью с другими обоснованными убеждениями. В случае «практико-ориентированных» методологий для выравнивания систем ИИ, когда ценности ИИ выводятся из предпочтений заинтересованных сторон, возникает следующая проблема: агрегация предпочтений может включать как более, так и менее рационально обоснованные убеждения. Например, предпочтение, основанное на эмоциях или манипулятивной рекламе, может быть менее рациональным, чем предпочтение, сформированное на основе анализа долгосрочных последствий. Если система ИИ ориентируется на «среднее» из предпочтений без учета их рациональной обоснованности, это может привести к нежелательным и в том числе этически проблематичным результатам: система ИИ будет отражать популярные, но необоснованные или вредные предпочтения, усиливая «тиранию данных». Поэтому качество (рациональность) предпочтений имеет значение наравне с их количеством.

Сильные и слабые стороны обоих подходов указывают на необходимость выработки третьего подхода, который бы учел их недостатки и аккумулировал достоинства (Wallach & Allen, 2009), то есть, с одной стороны, учитывал бы широкий спектр ценностных предпочтений различных социальных групп, а с другой — обеспечил бы их единой основой в качестве некоего синтеза моральных предпочтений. И попытки разработать такой третий подход уже имели место.

Первый такой гибридный подход — «Коллективное рефлексивное равновесие на практике» (*Collective Reflective Equilibrium in Practice*, CREP) — был предложен Савулеску, Гингеллом и Кахане (Savulescu et al., 2021). Он основан на данных об общественном отношении в качестве вклада в совещательный процесс, направленный на определение политики в области ИИ. Савулеску и его коллеги утверждают, что моральные предпочтения заинтересованных сторон должны быть тщательно проверены на предмет предвзятости и предубеждений. На

первом этапе осуществляется сбор высококачественных данных об общественной морали, специфичной для определенных контекстов и моральных дилемм, например, в случаях автомобильных беспилотных систем ИИ. При этом авторами недостаточно проясняются критерии для определения «качественности» данных. На втором этапе наиболее выраженные моральные позиции, выявленные в этих данных, анализируются с использованием методов моральной философии: концептуальный, семантический и логический анализ, рефлексивное равновесие и т. п. В результате формируется своего рода «общественное» рефлексивное равновесие, которое обладает более надежной степенью обоснования по сравнению с подходом «от принципов к практике» и «практико-ориентированным» подходом. Процедура выравнивания систем ИИ на основе некоего «здравого смысла» широкой общественности частично решает проблему легитимности, поскольку преобладающие моральные взгляды тех, кто будет взаимодействовать с ИИ, играют конкретную процедурную роль в решении ПВ.

Второй гибридный подход, предложенный Умбрелло и ван де Поэлем (Umbrello & van de Poel, 2021), основан на ценностно-чувствительном проектировании (*Value Sensitive Design*, VSD). В частности, авторы описывают четырехэтапный итеративный процесс проектирования, направленный на согласование технологий ИИ с общественными ценностями. В этот процесс входит:

(1) анализ контекста — на этом этапе изучаются социокультурные условия, в которых разрабатывается технология;
(2) идентификация ценностей — здесь определяются ключевые ценности, а также специфические ожидания пользователей;
(3) формулирование требований к проектированию — абстрактные ценности превращаются в конкретные правила для создания систем ИИ;
(4) прототипирование — создаются и тестируются прототипы, чтобы проверить, как ИИ влияет на людей, и при необходимости вносятся изменения.

Однако VSD может быть слишком общим и не учитывать специфику конкретных моральных контекстов, что снижает его эффективность при решении узкоспециализированных задач выравнивания систем ИИ.

Тем не менее обе попытки выработать гибридный подход к решению ПВ внесли значительный вклад во взаимную интеграцию предпочтений, опыта и моральных убеждений людей. Вместе с тем мы можем выделить в них и ряд недостатков. Первый связан с тем, что гибридные

подходы включают в себя этапы интеграции данных и принятия решений, что может сделать их непрозрачными для внешних наблюдателей. Например, как обоснована достоверность и релевантность собранных данных? Кем, как и на каких основаниях данные об общественных предпочтениях анализируются экспертами в области моральной философии? Непрозрачность подхода может привести к необходимости дополнительного обоснования условий доверия к качеству такой экспертизы, то есть может произойти смещение с оценки корректности процедуры выравнивания на обращение к эпистемическому авторитету экспертов по ПВ. Второй недостаток связан с возможным риском «замораживания» моральных норм: если гибридный подход фиксирует определенный набор данных на этапе разработки, он может «заморозить» моральные нормы в определенном состоянии. Это особенно проблематично в быстро меняющихся обществах, где взгляды на мораль (например, в отношении технологий или экологии) могут радикально трансформироваться за весьма короткий срок. Третий недостаток связан с уязвимостью обоих подходов перед лицом культурных искажений: даже при попытке учесть разные культуры гибридные подходы могут непреднамеренно отдавать предпочтение доминирующим культурным нарративам (например, западным этическим традициям), что ограничивает их универсальность для решения ПВ.

Принимая во внимание все эти недостатки предложенных подходов к решению ПВ, мы предлагаем обратиться к актуальной метаэтической теории, которая сочетает в себе элементы как объективных моральных принципов, так и контекстуальных, социально обусловленных ценностей. Кроме того, ценность метаэтического подхода заключается еще и в том, что метаэтика, как раздел аналитической философии, стремится ответить в том числе на следующие вопросы: существуют ли универсальные моральные истины, которые можно внедрить в ИИ, или мораль релятивна по определению и потому стоит отказаться от самой попытки ценностного дизайна ИИ? Каковы основания моральных суждений, вырабатываемых системой ИИ? Могут ли системы ИИ быть моральными агентами и как это меняет структуру самого этического знания? Наконец, именно метаэтический подход может позволить, на наш взгляд, решить одну из ключевых проблем гибридных подходов, связанных с неясностью критериев «качественности» данных.

Среди всего многообразия метаэтических подходов наиболее релевантными рассматриваемой проблематике являются нон-натурализм (интуиционизм) и моральный натурализм.

МЕТАЭТИЧЕСКИЙ ПОДХОД. НОН-НАТУРАЛИЗМ

Для решения ПВ систем ИИ с позиций метаэтики необходимо установить, какой тип концептуализации[2] моральных фактов (или свойств) следует использовать в качестве основы для разработки процедур выравнивания. Основополагающим теоретическим выбором является выбор между моральным реализмом и антиреализмом.

Моральный реализм опирается на онтологическую предпосылку, согласно которой моральные факты существуют объективно, независимо от субъективных представлений человека. Это предполагает, что структура мира, включая биологическую и социальную природу человека, определяет моральную правильность или неправильность действий независимо от индивидуальных целей или интересов. Метаэтические теории морального реализма обосновывают семантику моральных высказываний, устанавливая условия их истинности и связывая моральные понятия с предполагаемыми моральными свойствами. Такой подход позволяет выстраивать концептуальную основу для интеграции моральных ценностей в системы ИИ.

В случае признания истинным морального антиреализма[3] прикладная значимость выравнивания систем ИИ становится проблематичной. Отдельные решения для таких систем могут восприниматься как одни из множества равноправных решений, причем ни одно из них не является приоритетным. У пользователей таких систем ИИ отсутствуют моральные основания для доверия их рекомендациям в сравнении с собственными убеждениями. Учитывая индивидуалистический характер оценочных суждений, моральные убеждения, основанные на повседневных предпочтениях, могут быть иррациональными, ненадежными и трудно совместимыми для выработки общего предпочтения (Gabriel, 2020). В связи с этим субъективистские теории и теории морального антиреализма исключаются из нашего рассмотрения. В качестве базовой установки принимается объективный моральный реализм, в рамках которого наиболее релевантными представляются моральные факты,

[2]Классификация этих типов основана на следующих фундаментальных вопросах метаэтики: онтологический статус моральных фактов — существуют ли они объективно (реализм) или зависят от субъекта/социума (антиреализм, конструктивизм)? Эпистемология морали — как мы познаём моральные факты: через разум, интуицию, эмпирические данные или социальные договоренности? Семантический статус моральных суждений — выражают ли они факты, эмоции или предписания?

[3]Моральный антиреализм утверждает, что моральные свойства или факты не существуют или, по крайней мере, не существуют независимо от человека.

сформулированные в теориях нон-натурализма — в случае «слабого» ИИ — и морального натурализма — в случае «сильного» ИИ.

Начнем с анализа нон-натурализма, взяв за основу интуиционизм как ведущую метаэтическую теорию в рамках этого направления. Нонатуралистический подход предполагает, что моральные ценности систем ИИ должны опираться на моральные интуиции людей. Специалисты в области этики в сфере ИИ (Savulescu et al., 2021) подчеркивают важность моральной интуиции для информационного обеспечения ИИ, но никто не дает конкретного определения этого ментального состояния человека. Интуиция имеет моральное содержание. В соответствии с последними исследованиями в области моральной психологии (Bonnefon et al., 2020), интуиция — это особый тип морального суждения, обладающий следующими особенностями.

(1) Интуиции выражаются в пропозициях, утверждающих моральную оценку (правильность, неправильность, добро, зло, обязательность или допустимость). Уровень общности таких суждений варьируется от конкретных случаев (например, интуитивное представление о том, что жестокое обращение с животными ради развлечения недопустимо) до общих суждений (например, что система ИИ, подобная ChatGPT, не должна выносить оценочные суждения по религиозным вопросам), принципов среднего уровня (например, развитие ИИ должно способствовать толерантности) или абстрактных теоретических принципов (например, моральная правильность действия определяется его последствиями).

(2) Моральные интуиции возникают как непосредственные реакции, основанные на первоначальном интеллектуальном восприятии, и не требуют развернутых рассуждений (Huemer, 2008: 370). Конкретные интуиции служат для проверки и иллюстрации общих моральных принципов, тогда как абстрактные интуиции играют роль в обосновании этических теорий. Например, утверждение, что ИИ должен способствовать максимизации пользы, поддерживается утилитарным принципом, согласно которому действия оцениваются по их способности увеличивать общую полезность.

Каждый тип интуиций имеет свои преимущества и недостатки. Так, абстрактные интуиции, как правило, находят большее согласие в различных культурах и полезны для интеграции моральных суждений в общие цели ИИ. Однако они часто отличаются расплывчатостью и риском чрезмерного обобщения, не позволяющего конкретизировать их в отдельных решениях ИИ. Конкретные интуиции, напротив, более

подвержены разногласиям в силу их разнородности, но они играют ключевую роль в применении моральных принципов к конкретным случаям. В связи с этим выравнивание систем ИИ должно учитывать интуиции всех уровней. Для этого системы ИИ требуется проектировать с многоуровневой архитектурой:

(1) базовый уровень: конкретные интуиции как правила для специфичных ситуаций (например, «не лгать в официальных документах»);
(2) средний уровень: принципы, обобщающие конкретные интуиции для класса случаев (например, «честность в коммуникации»);
(3) верхний уровень: абстрактные интуиции как общие цели (например, «доверие»).

ИИ обращается к нужному уровню в зависимости от контекста. На практике в системе рекомендаций ИИ верхний уровень может требовать «максимизации пользы», а конкретные правила — избегать рекомендаций вредного контента. Другой пример: в автономном автомобиле абстрактный принцип «минимизация вреда» может быть основной целью, а конкретное правило «уступить дорогу пешеходу на переходе» — действовать в специфических ситуациях. Технически это может быть реализовано через комбинацию систем на основе правил (символических) и систем на основе данных (например, нейронных сетей), что позволяет интегрировать оба типа интуиций. Символические системы кодируют абстрактные принципы (например, запрет на причинение вреда невинным), а нейронные сети обрабатывают нюансы конкретных ситуаций. Так, например в чат-боте этическая система правил может запретить оскорбительные ответы (абстрактный принцип), а нейронная сеть — выбрать наиболее подходящий тон ответа в зависимости от контекста общения. При этом, как отмечал М. Хьюмер, приоритетными являются формальные[4] интуиции в силу того, что они наиболее надежны (Huemer, 2008: 380). Формальные интуиции — это абстрактные, логические принципы, такие как транзитивность и супервенирование, которые устанавливают базовые правила для этического рассуждения. Они считаются наиболее надежными из-за их устойчивости к предубеж-

[4] «...существует особый вид абстрактных этических интуиций, который представляется мне необычайно надежным. Это то, что я называю формальными интуициями, — интуиции, которые накладывают формальные ограничения на этические теории, хотя сами по себе они не оценивают ничего положительно или отрицательно» (перевод мой. — *И. С.).* Примеры формальной интуиции: (1) Если A лучше, чем B, и B лучше, чем C, то A лучше, чем C (транзитивность); (2) Если A и B идентичны во всех неэтических аспектах, то они морально неразличимы (супервенирование).

дениям, логической необходимости и когерентности, что делает их ценным инструментом для построения объективных этических теорий. Эти интуиции особенно важны в контексте нон-натурализма (интуитивизма), где предполагается, что некоторые моральные истины известны интуитивно, без вывода из других утверждений.

Формальные интуиции помогают отфильтровывать определенные этические позиции, которые могут казаться уместными в иных обстоятельствах, тем самым способствуя разрешению конфликтов между противоречащими друг другу интуициями. Их ценность в контексте ПВ систем ИИ заключается в упрощении перехода от формализованных описаний интуиций к кодированию нормативных ценностей с использованием методов программирования ИИ.

Решение ПВ из оснований метаэтического нон-натурализма (интуиционизма) можно представить в трех этапах.

(1) Сбор данных о моральных интуициях: (а) привлечение испытуемых к моральным проблемам, связанным с системами ИИ; (б) минимизация предвзятости через обеспечение оптимальных условий для вынесения моральных суждений.
(2) Оправдание моральных интуиций: (а) обоснование надежности интуиций; (б) обнаружение конфликтующих интуиций.
(3) Экспертная дискуссия: (а) поиск компромиссов между конфликтующими интуициями; (б) включение интуиций в политику в области ИИ.

Для эффективного решения ПВ необходимы научно обоснованные инструменты для сбора и анализа данных, учитывающие специфику моральных интуиций. В этом контексте предлагается использовать методы экспериментальной философии[5]. Мы предлагаем три основные линии исследований ПВ применительно к ИИ.

[5]Экспериментальная философия — это междисциплинарный подход, который использует эмпирические методы, заимствованные из социальных наук, таких как психология, социология и когнитивные исследования, для изучения философских вопросов. Вместо традиционного философского метода, основанного на концептуальном и логическом анализе, экспериментальная философия стремится систематически собирать данные о том, как люди (обычные люди, а не только профессиональные философы) понимают философские концепты, такие как мораль, свобода воли и т. д. Этот подход направлен на проверку и уточнение философских теорий с помощью эмпирических данных, что делает его релевантным для анализа моральных интуиций в контексте ПВ. Без эмпирического понимания того, как люди воспринимают мораль, система ИИ может принимать решения, которые будут отвергнуты обществом как неприемлемые или нерелевантные. Подробнее см. Knobe & Nichols, 2017.

Первая линия заключается в сборе и анализе данных об интуициях, лежащих в основе моральных ценностей конкретных сообществ, например, российских граждан. Моральные ценности часто характеризуются неопределенностью и вариативностью в их восприятии. Хотя такие концепты, как, например, «справедливость», являются общими, их семантическое содержание существенно различается между индивидами и отдельными социальными группами. Выявление общих моральных интуиций среди граждан Российской Федерации позволяет разработать понятийный аппарат моральных ценностей с четко определенным семантическим полем, которое может применяться для разработки политики в области ИИ и служить основой для решения ПВ.

Вторая линия предполагает историко-культурный анализ моральных ценностей с использованием методов лингвистического анализа. Предлагается изучение основного корпуса текстов для выявления случаев употребления понятий, связанных с ценностями, и экспликация их семантического содержания в единый понятийный аппарат. Результаты таких исследований могут сформировать базу данных, отражающую общественные интуитивные установки, которая будет использоваться для обучения систем ИИ.

Третья линия — обнаружение специфических нормативных ожиданий от систем ИИ как моральных агентов. Например, эмпирические исследования показали, что для значительной части опрошенных тип агента (ИИ или человек) существенно влияет на моральную оценку приемлемости действий в ситуациях сложных моральных компромиссов (Kneer & Viehoff, 2025). Этот феномен, обозначенный авторами исследования как «разногласие ценностей (типа агента)», имеет важные теоретические и практические следствия для дальнейшего согласования ценностей ИИ. В частности, он предполагает необходимость разработки систем ИИ, способных выполнять действия, которые могут быть признаны неприемлемыми для людей в аналогичных обстоятельствах.

Методологический подход, основанный на принципах метаэтического нон-натурализма, обеспечивает теоретическое обоснование использования эмпирических данных для реализации моральных ценностей в системах «слабого» ИИ, таких как современные генеративные модели. Он способствует созданию научно обоснованной основы для интеграции моральных ценностей в ИИ, учитывающей как индивидуальные, так и коллективные интуиции, как актуальные, так и культурно-исторические.

моральный натурализм

Если система ИИ функционирует как автономный моральный агент при принятии моральных решений, в том числе в отношении человека, она должна стремиться к воспроизведению элементов моральной психологии человека, лежащих в основе моральных суждений, несмотря на значительные методологические сложности, сопровождающие эту задачу. Отклонение моральных предложений системы ИИ от ценностных предпочтений человека с высокой долей вероятности приведет к их восприятию как нерелевантных человеческой морали. Метаэтические допущения, принятые разработчиками при создании ИИ, стремящихся моделировать моральную психологию человека, существенно влияют на возникающие технические вызовы (Frank & Klincewicz, 2016: 208–213). К числу таких допущений относится определение природы моральных высказываний и моральной мотивации, эпистемического статуса моральных фактов и различий между моральным знанием и иными формами знания. Эти допущения, будь то имплицитные или эксплицитные, неизбежно влияют на разрабатываемые подходы к решению ПВ в системах ИИ.

Синтетический моральный натурализм, известный как корнельский реализм, постулирует, что моральные свойства супервентны по отношению к естественным свойствам, поскольку они реализуются или конституируются ими и обладают объяснительной и каузальной эффективностью (Кононов, 2023: 98). Система ИИ, разработанная на основе морального натурализма для выполнения функции «советника» по морально значимым вопросам, может выступать в роли своеобразного «морального компаса», открывая доступ к ранее недоступным моральным фактам. Такая система, способная к познанию моральных фактов, потенциально может выявлять моральные истины, недоступные даже людям с высокой моральной чувствительностью и развитыми способностями к автономным моральным суждениям. Реалистическая перспектива предполагает, что мир может содержать множество моральных фактов, которые пока остаются неизвестными, но могут быть обнаружены в будущем, подобно тому как физические факты открываются постоянно в процессе изучения природы. Однако эта аналогия подчеркивает потенциальную опасность: системы ИИ, основанные на моральном натурализме, могут предлагать моральные рекомендации, противоречащие устоявшимся, стереотипным решениям и нормам; при

отсутствии общепризнанных методологий для оценки их целесообразности в сравнении с существующими моральными принципами это противоречие способно провоцировать конфликты. В отличие от научного метода, чья доказательность опирается на консенсус научного сообщества и эксперимент, в этике отсутствуют универсально признанные подходы к выявлению истинных моральных фактов и получению моральных знаний.

Сложность заключается в том, что реалистические допущения, применяемые при разработке ИИ, могут привести к ошибочным выводам о моральных фактах, побуждая следовать рекомендациям ИИ, которые могут быть некорректными. Этот риск применим также к нон-натуралистическим подходам, упомянутым ранее. Тем не менее это не означает, что в этике отсутствуют обоснованные методологии, которые могли бы быть использованы сторонниками морального натурализма. Например, представители корнельского реализма предлагают когерентистскую модель моральной эпистемологии, основанную на рефлексивном равновесии. В этой модели научные и этические суждения подвергаются непрерывной корректировке путем сопоставления с эмпирическими данными, учитывающими теоретические взгляды на природу морали, что позволяет достигать оптимального соответствия между разнообразными моральными убеждениями в обществе (Sturgeon, 2006: 102). Однако методы моральной эпистемологии не поддаются проверке так же, как научные методы, поскольку установление истинности моральных убеждений затруднено из-за отсутствия независимых критериев верификации. В рамках морального натурализма система моральных убеждений опирается на ограниченный набор натуралистических методологических принципов, которые составляют фундамент всей структуры, в отличие от более широкого рефлексивного равновесия, характерного, например, для нон-натуралистического интуиционизма, который, однако, менее надежен в плане обоснованности интуиций.

В качестве примеров принципов морального натурализма можно выделить следующие.

(1) Онтологический принцип: моральные факты всех биологических организмов реализуются через биологический субстрат или с его участием.

(2) Эпистемологический принцип: философский анализ моральных фактов должен быть совместим с современными исследованиями в области моральной психологии и биологии человека.

(3) Эвристический принцип: знание о структуре моральных фактов может быть получено на основе изучения структуры и динамики биологических процессов.

Эти принципы находят подтверждение в нейронаучных исследованиях моральных убеждений. Инструменты нейронауки позволяют выявлять нейронные корреляты, связанные с моральными ценностями. Например, исследование китайских нейробиологов, посвященное изучению нейронных основ иерархий ценностей, показало, что традиционная ценностная ориентация связана с выраженной функциональной связностью головного мозга с островковой долей (The Neural Correlates..., 2023) головного мозга. Иными словами, мы можем однозначно фиксировать наличие или отсутствие конкретных функциональных процессов в мозге, соответствующих приверженности его обладателя традиционному типу ценностей. Обнаруженные нейронные функциональные связи в мозге позволяют установить, как именно моральные факты конституируются естественными функциональными свойствами в рамках морального натурализма. Эти моральные факты уже могут быть использованы сторонниками натуралистических методологий при выравнивании систем ИИ в соответствии с нейробиологическими данными о когнитивных механизмах формирования человеческих ценностей.

Разработчики систем ИИ, основанных на моральном натурализме, обязаны обосновать корректность своего методологического подхода. Выбор морального натурализма предполагает приоритет в создании систем ИИ, способных эпистемологически превосходить человека в выявлении моральных фактов. Эта методологическая предпосылка имеет ключевое значение для разработки не только «слабого», но и «сильного» ИИ. Однако достижение этой цели остается сложной задачей. На этапах разработки и тестирования системы ИИ могут давать ошибочные ответы на фундаментальные моральные вопросы, что может подорвать доверие к разработчикам и самим системам. Несмотря на эти риски, изучение морального натурализма представляется необходимым ввиду его преимуществ для решения ПВ:

(1) возможность фальсификации теорий морального натурализма;
(2) выявление новых натуралистических моральных фактов;
(3) эмпирическая валидация ценностных оснований систем ИИ;
(4) преодоление ценностных расхождений между естественными и искусственными агентами.

Потенциальные риски лишь подчеркивают необходимость тщательной разработки и проверки методологий, основанных на моральном натурализме. В случае подтверждения эффективности натуралистических моральных методологий крайне важно заранее разработать инструменты для калибровки систем ИИ в соответствии с натуралистическими подходами к исследованию морали.

ЗАКЛЮЧЕНИЕ

По мере развития ИИ мы будем получать все более совершенные модели. По мере усложнения стоящих перед человеком задач нам становится все труднее оценивать, соответствует ли поведение системы ИИ нашим намерениям и целям. Это делает ПВ особенно актуальной. Существующие на данный момент подходы к ее решению не достаточны в силу своей ограниченности, что требует разработки гибридных моделей. На наш взгляд, высокий потенциал на этом пути обнаруживает ряд метаэтических теорий, прежде всего натурализм и нон-натурализм. Преимущество нон-натурализма в том, что в его рамках оказывается возможным соотнести индивидуальные и коллективные моральные интуиции посредством преодоления ценностных разрывов между человеком и ИИ. Натурализм же позволяет вывести моральные принципы из наблюдаемых фактов о природе человека, делая системы ИИ более прозрачными и предсказуемыми. Очевидно, что обе теории в рассматриваемом контексте ПВ требуют дальнейшей концептуальной проработки. А их надежность и эффективность могут быть оценены в долгосрочной перспективе в ходе анализа непредвиденных последствий внедрения ИИ.

Литература

Гарбук С. В., *Углева А. В.* Автоматизированные интеллектуальные системы : этический и нормативно-технический подходы к регулированию // Человек. — 2024. — Т. 35, № 4. — С. 98–117.

Кононов Е. А. Метаэтика. Теоретический обзор. — М., 2023.

Ролз Д. Теория справедливости / пер. с англ. В. В. Целищева. — Новосибирск : Новосибирский университет, 1995.

AI4People—An Ethical Framework for a Good AI Society : Opportunities, Risks, Principles, and Recommendations / L. Floridi [et al.] // Minds and Machines. — 2018. — Vol. 28. — P. 689–707.

Anscombe G. E. M. Intention. — Cambridge : Harvard University Press, 2000.

Bonnefon J.-F., *Shariff A.*, *Rahwan I.* The Moral Psychology of AI and the Ethical Opt-out Problem // Ethics of Artificial Intelligence / ed. by S. M. Liao. — Oxford : Oxford University Press, 2020. — P. 109–126.

Bostrom N. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence // Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. Vol. 2 / ed. by I. Smit, W. Wallach, G. E. Lasker. — Baden-Baden : International Institute for Advanced Studies in Systems Research, Cybernetics, 2003. — P. 12–17.

Bostrom N. The Superintelligent Will : Motivation and Instrumental Rationality in Advanced Artificial Agents // Minds and Machines. — 2012. — Vol. 22, no. 2. — P. 71–85.

Christiano P. Conversation with Paul Christiano / AI Impacts. — 2019. — URL: https://aiimpacts.org/conversation-with-paul-christiano/ (visited on Oct. 15, 2024).

Concrete Problems in AI safety / D. Amodei [et al.] ; arXiv preprint. — 2016. — URL: https://arxiv.org/abs/1606.06565 (visited on Sept. 14, 2024).

Deep Reinforcement Learning from Human Preferences / P. Christiano [et al.] ; arXiv preprint. — 2017. — URL: https://arxiv.org/abs/1706.03741.

Delphi : Advancing Machine Ethics and Norms / L. Jiang [et al.] ; arXiv preprint. — 2021. — URL: https://arxiv.org/abs/2110.07574 (visited on May 16, 2025).

Drexel B., *Withers C.* Generative AI Could Be an Authoritarian Breakthrough in Brainwashing / CNAS Commentary. — 2023. — URL: https://www.cnas.org/publications/commentary/generative-ai-could-be-an-authoritarian-breakthrough-in-brainwashing (visited on Aug. 24, 2024).

Emergent Abilities of Large Language Models / J. Wei [et al.] ; arXiv preprint. — 2022. — URL: https://arxiv.org/abs/2206.07682 (visited on Sept. 11, 2025).

Ford P. Our Fear of Artificial Intelligence / MIT Technology Review. — 2015. — URL: https://www.technologyreview.com/2015/02/11/169210/our-fear-of-artificial-intelligence/ (visited on Oct. 15, 2024).

Frank L., *Klincewicz M.* Metaethics in Context of Engineering Ethical and Moral Systems // 2016 AAAI Spring Symposium Series. — 2016. — P. 208–213.

Funke J. Analysis of Minimal Complex Systems and Complex Problem Solving Require Different Forms of Causal Cognition // Frontiers in Psychology. — 2014. — Vol. 5, no. 739.

Gabriel I. Artificial Intelligence, Values, and Alignment // Minds & Machines. — 2020. — Vol. 30, no. 3. — P. 411–437.

Huemer M. Revisionary Intuitionism // Social Philosophy and Policy. — 2008. — Vol. 25, no. 1. — P. 368–392.

Irving G., *Christiano P.*, *Amodei D.* AI Safety via Debate / arXiv preprint. — 2018. — URL: https://arxiv.org/abs/1805.00899 (visited on Oct. 15, 2024).

Jobin A., *Ienca M.*, *Vayena E.* The Global Landscape of AI Ethics Guidelines // Nat. Mach. Intell. — 2019. — Vol. 1. — P. 389–399.

Kissinger H. A. How the Enlightenment Ends : Philosophically, Intellectually — in Every Way — Human Society is Unprepared for the Rise of Artificial Intelligence / The Atlantic. — 2018. — URL: https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/ (visited on Aug. 24, 2024).

Kneer M., *Viehoff J.* The Hard Problem of AI Alignment // Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25). — New York (NY) : Association for Computing Machinery, 2025. — P. 1–11.

Knobe J., *Nichols S.* Experimental Philosophy / The Stanford Encyclopedia of Philosophy. — 2017. — URL: https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/ (visited on Dec. 16, 2024).

Machine Behaviour / I. Rahwan [et al.] // Nature. — 2019. — Vol. 568, no. 7753. — P. 477–486.

Millière R. Normative Conflicts and Shallow AI Alignment // Philosophical Studies. — 2025. — P. 1–44.

Mittelstadt B. Principles Alone Cannot Guarantee Ethical AI // Nature Machine Intelligence. — 2019. — Vol. 1. — P. 501–507.

Moral Judgments in Human-robot Interactions : Differential Application of Moral Norms to Human and Robotic Agents / B. F. Malle [et al.] // Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15). — New York (NY) : Association for Computing Machinery, 2015. — P. 117–124.

Russell S. Human Compatible : Artificial Intelligence and the Problem of Control. — New York : Penguin, 2019.

Savulescu J., *Gyngell C.*, *Kahane G.* Collective Reflective Equilibrium in Practice (CREP) and Controversial Novel Technologies // Bioethics. — 2021. — Vol. 35. — P. 652–663.

Strathern M. Improving Ratings : Audit in the British University System // European Review. — 1997. — Vol. 5, no. 3. — P. 305–321.

Sturgeon N. Ethical Naturalism // Oxford Handbook of Ethical Theory / ed. by D. Copp. — Oxford : Oxford University Press, 2006. — P. 91–121.

The Neural Correlates of Value Hierarchies : A Prospective Typology Based on Personal Value Profiles of Emerging Adults / J.-Q. Xie [et al.] // Frontiers in Psychology. — 2023. — Vol. 14.

Umbrello S., *Poel I. van de.* Mapping Value Sensitive Design onto AI for Social Good Principles // AI Ethics. — 2021. — Vol. 1. — P. 283–296.

Vinge V. The Coming Technological Singularity : How to Survive in the Post-human Era // Vision 21 : Interdisciplinary Science and Engineering in the Era of Cyberspace. — Cleveland : NASA Lewis Research Center, 1993. — P. 11–22.

Wallach W., *Allen C.* Moral Machines : Teaching Robots Right from Wrong. — Oxford : Oxford University Press, 2009.

Yudkowsky E. Complex Value Systems in Friendly AI // Artificial General Intelligence / ed. by J. Schmidhuber, K. R. Thórisson, M. Looks. — Berlin : Springer, 2011. — P. 388–393.

Snetkov, I. G. 2025. "Metaeticheskiye osnovaniya vyravnivaniya iskusstvennogo intellekta [Metaethical Foundations of Artificial Intelligence Alignment]: metodologicheskiye podkhody i ikh ogranicheniya [Methodological Approaches and Their Limitations]" [in Russian]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (3), 277–302.

Ivan Snetkov
Lecturer
HSE University (Moscow, Russia); orcid: 0000–0003–4314–3346

Metaethical Foundations of Artificial Intelligence Alignment

Methodological Approaches and Their Limitations

Submitted: Dec. 23, 2024. **Reviewed:** May 18, 2025. **Accepted:** Aug. 08, 2025.

Abstract: The article investigates the alignment problem, which concerns the integration of moral values into the architecture of artificial intelligence (AI) systems to mitigate existential risks. It examines conceptual approaches to addressing the alignment problem, including the utilitarian principles proposed by S. Russell and E. Yudkowsky's concept of "coherent extrapolated volition". The study introduces the notion of a "meta-alignment problem". Through an analysis of the conceptual distinction between "strong" and "weak" AI, the author concludes that these categories necessitate distinct approaches to resolving the alignment problem. The article evaluates existing methodological approaches to tackling this issue, including the "principles-to-practice" approach and the "practice-oriented" approach, highlighting their limitations, such as difficulties in operationalizing moral principles and accommodating individual moral preferences. It also explores the potential of "hybrid" approaches. The consideration of metaethical foundations is proposed as a means to address a key challenge in hybrid approaches, namely the ambiguity surrounding the criteria for data "quality". The study advocates for the use of conceptual models of morality developed within metaethics — specifically non-naturalism (intuitionism) and moral naturalism — as a foundation for devising new hybrid alignment strategies. The non-naturalist approach relies on moral intuitions explored through experimental philosophy, enabling the reconciliation of individual and collective moral intuitions by bridging value gaps between humans and AI. In contrast, the naturalist approach draws on neurobiological data to identify moral "facts", rendering AI systems more transparent and predictable. Metaethical foundations significantly influence AI design, and their explicit consideration not only facilitates the development of effective alignment methodologies but also allows for empirical evaluation of the viability of metaethical approaches in addressing the alignment problem. The article contributes to the discourse on the metaethical foundations of AI alignment. It proposes directions for future research and outlines potential pathways for aligning AI systems with moral values.

Keywords: Artificial Intelligence, Existential Risks, Alignment Problem, Metaethics, Moral Non-naturalism, Moral Naturalism.

DOI: 10.17323/2587-8719-2025-3-277-302.

REFERENCES

Amodei, D., et al. 2016. "Concrete Problems in AI safety." arXiv preprint. Accessed Sept. 14, 2024. https://arxiv.org/abs/1606.06565.

Anscombe, G. E. M. 2000. *Intention.* Cambridge: Harvard University Press.

Bonnefon, J.-F., A. Shariff, and I. Rahwan. 2020. "The Moral Psychology of AI and the Ethical Opt-out Problem." In *Ethics of Artificial Intelligence*, ed. by S. M. Liao, 109–126. Oxford: Oxford University Press.

Bostrom, N. 2003. "Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, ed. by I. Smit, W. Wallach, and G. E. Lasker, 2:12–17. Baden-Baden: International Institute for Advanced Studies in Systems Research / Cybernetics.

——— . 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22 (2): 71–85.

Christiano, P. 2019. "Conversation with Paul Christiano." AI Impacts. Accessed Oct. 15, 2024. https://aiimpacts.org/conversation-with-paul-christiano/.

Christiano, P., et al. 2017. "Deep Reinforcement Learning from Human Preferences." arXiv preprint. https://arxiv.org/abs/1706.03741.

Drexel, B., and C. Withers. 2023. "Generative AI Could Be an Authoritarian Breakthrough in Brainwashing." CNAS Commentary. Accessed Aug. 24, 2024. https://www.cnas.org/publications/commentary/generative-ai-could-be-an-authoritarian-breakthrough-in-brainwashing.

Floridi, L., et al. 2018. "AI4People— An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28:689–707.

Ford, P. 2015. "Our Fear of Artificial Intelligence." MIT Technology Review. Accessed Oct. 15, 2024. https://www.technologyreview.com/2015/02/11/169210/our-fear-of-artificial-intelligence/.

Frank, L., and M. Klincewicz. 2016. "Metaethics in Context of Engineering Ethical and Moral Systems." *2016 AAAI Spring Symposium Series*, 208–213.

Funke, J. 2014. "Analysis of Minimal Complex Systems and Complex Problem Solving Require Different Forms of Causal Cognition." *Frontiers in Psychology* 5 (739).

Gabriel, I. 2020. "Artificial Intelligence, Values, and Alignment." *Minds & Machines* 30 (3): 411–437.

Garbuk, S. V., and A. V. Ugleva. 2024. "Avtomatizirovannyye intellektual'nyye sistemy [Automated Intelligent Systems]: eticheskiy i normativno-tekhnicheskiy podkhody k regulirovaniyu [Ethical and Regulatory Approaches to Regulation]" [in Russian]. *Chelovek [Human]* 35 (4): 98–117.

Huemer, M. 2008. "Revisionary Intuitionism." *Social Philosophy and Policy* 25 (1): 368–392.

Irving, G., P. Christiano, and D. Amodei. 2018. "AI Safety via Debate." arXiv preprint. Accessed Oct. 15, 2024. https://arxiv.org/abs/1805.00899.

Jiang, L., et al. 2021. "Delphi: Advancing Machine Ethics and Norms." arXiv preprint. Accessed May 16, 2025. https://arxiv.org/abs/2110.07574.

Jobin, A., M. Ienca, and E. Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nat. Mach. Intell* 1:389–399.

Kissinger, H. A. 2018. "How the Enlightenment Ends: Philosophically, Intellectually—in Every Way—Human Society is Unprepared for the Rise of Artificial Intelligence." The Atlantic. Accessed Aug. 24, 2024. https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/.

Kneer, M., and J. Viehoff. 2025. "The Hard Problem of AI Alignment." In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, 1–11. New York (NY): Association for Computing Machinery.

Knobe, J., and S. Nichols. 2017. "Experimental Philosophy." The Stanford Encyclopedia of Philosophy. Accessed Dec. 16, 2024. https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/.

Kononov, Ye. A. 2023. *Metaetika. Teoreticheskiy obzor [Metaethics. Theoretical Review]* [in Russian]. Moskva [Moscow].

Malle, B. F., et al. 2015. "Moral Judgments in Human-robot Interactions: Differential Application of Moral Norms to Human and Robotic Agents." In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*, 117–124. New York (NY): Association for Computing Machinery.

Millière, R. 2025. "Normative Conflicts and Shallow AI Alignment." *Philosophical Studies*, 1–44.

Mittelstadt, B. 2019. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1:501–507.

Rahwan, I., et al. 2019. "Machine Behaviour." *Nature* 568 (7753): 477–486.

Rawls, J. 1995. *Teoriya spravedlivosti [A Theory of Justice]* [in Russian]. Trans. from the English by V. V. Tselishchev. Novosibirsk: Novosibirskiy universitet [Novosibirsk State University Press].

Russell, S. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin.

Savulescu, J., C. Gyngell, and G. Kahane. 2021. "Collective Reflective Equilibrium in Practice (CREP) and Controversial Novel Technologies." *Bioethics* 35:652–663.

Strathern, M. 1997. "Improving Ratings: Audit in the British University System." *European Review* 5 (3): 305–321.

Sturgeon, N. 2006. "Ethical Naturalism." In *Oxford Handbook of Ethical Theory*, ed. by D. Copp, 91–121. Oxford: Oxford University Press.

Umbrello, S., and I. van de Poel. 2021. "Mapping Value Sensitive Design onto AI for Social Good Principles." *AI Ethics* 1:283–296.

Vinge, V. 1993. "The Coming Technological Singularity: How to Survive in the Post-human Era." In *Vision 21 : Interdisciplinary Science and Engineering in the Era of Cyberspace*, 11–22. Cleveland: NASA Lewis Research Center.

Wallach, W., and C. Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Wei, J., et al. 2022. "Emergent Abilities of Large Language Models." arXiv preprint. Accessed Sept. 11, 2025. https://arxiv.org/abs/2206.07682.

Xie, J.-Q., et al. 2023. "The Neural Correlates of Value Hierarchies: A Prospective Typology Based on Personal Value Profiles of Emerging Adults." *Frontiers in Psychology* 14.

Yudkowsky, E. 2011. "Complex Value Systems in Friendly AI." In *Artificial General Intelligence*, ed. by J. Schmidhuber, K. R. Thórisson, and M. Looks, 388–393. Berlin: Springer.