Этическая мысль
2024. Т. 24. № 1. С. 101–114
УДК 004.8, 179.9

Ethical Thought
2024, Vol. 24, No. 1, pp. 101–114
DOI: 10.21146/2074-4870-2024-24-1-101-114

В.К. Кудряшова

Может ли искусственный интеллект быть «этичным»? Перспектива (современной) этики добродетели

Кудряшова Валерия Константиновна – преподаватель. Национальный исследовательский университет «Высшая школа экономики». Российская Федерация, 105006, г. Москва, ул. Старая Басманная, д. 21/4, корп. 1.

ORCID: 0000-0002-8147-2530
e-mail: vkkudryashova@hse.ru

В статье рассматриваются методологические ограничения применения (современной) этики добродетели в сфере искусственного интеллекта. Первое ограничение заключается в теоретической неопределенности «этики добродетели», так как в аспекте многих вопросов прикладной этики она инструментально истолковывается многими исследователями только как подход, отсылающий к аристотелевскому принципу середины и позволяющий артикулировать ключевые добродетели для ИИ. Такой подход является в корне неверным, так как он выборочно подходит к содержанию этики добродетели, вычленяя из нее лишь те принципы, которые способствуют его апробации. Второе ограничение связано с тем, что в аристотелизме и нео-аристотелизме ИИ не может рассматриваться в качестве морального агента sensu stricto. В связи с этим ИИ представляет собой систему, которая может только имитировать поведение морального агента, являющегося носителем определенных добродетелей. Несмотря на первичный отрицательный диагноз, в нео-аристотелизме можно выделить два направления, которые могут предложить новый подход к этике искусственного интеллекта. Экземпляризм позволяет иначе взглянуть на роль социальных роботов в жизни людей: они могут рассматриваться в качестве моральных образцов и нарративов, которые могут способствовать моральному развитию своих владельцев и менять их моральный облик в лучшую/худшую сторону. Подход возможностей М.К. Нуссбаум также сфокусирован на влиянии ИИ на качество человеческой жизни: в данном контексте речь идет о том, какие изменения могут претерпевать базовые человеческие возможности и как будут обеспечены минимальная социальная справедливость и человеческое достоинство при применении технологий ИИ.

Ключевые слова: этика добродетели, нео-аристотелизм, моральный экземпляризм, подход возможностей, искусственный интеллект, этика искусственного интеллекта

(Современная) этика добродетели в контексте ИИ

Вопрос возможности создания этичного искусственного интеллекта является одним из наиболее обсуждаемых в области прикладной этики, так как он не только отражает актуальную повестку современного общества, но и подсвечивает проблемы, с которыми сталкиваются нормативные этические теории. Каждая из них вносит свой вклад в осмысление феномена ИИ и предлагает ряд этических ограничений, связанных с его практическим применением. В этом контексте речь в большинстве случаев идет о спектре деонтологических и консеквенциалистских подходов: все они претендуют на обладание статусом универсальности, следовательно, их принципы, распространяемые на искусственный интеллект, также носят характер всеобщности. Однако это не всегда представляется возможным в силу того, что предметная область ИИ остается довольно широкой и принципы деонтологии и консеквенциализма, применяемые к интеллектуальным системам, не всегда проходят проверку на прочность в различных контекстуальных условиях [Cappuccio et al., 2020, 3–4; Allen et al., 2010, 252–254; Wallach, Allen, 2009, 84–86, 95–97]. В связи с последним аспектом в исследовательской литературе можно все чаще и чаще встретить упоминание и третьего теоретического подхода, который помогает преодолеть ограничения деонтологии и консеквенциализма, – этику добродетели. Под этикой добродетели в сфере ИИ понимается такой подход, который предлагает рассматривать определенный набор добродетелей, которые могут помочь регулировать работу систем искусственного интеллекта. Несмотря на обилие дискурсов о добродетели, некоторые авторы [Coleman, 2001, 250; Klincewicz, 2016; Berberich, Diepold, 2018; Li, 2021; Hagendorff, 2022, 4–6; Boddington, 2023, 264–272][1] указывают на аристотелевскую этику добродетели в качестве теоретической основы: в их научных работах это подтверждается наличием ссылок на аристотелевское понимание добродетели, контекстуальность фронезиса (практической мудрости, рассудительности) и аристотелевское учение о середине; далее авторы составляют подборку из наиболее релевантных, на их взгляд, добродетелей, которые могут быть эксплицированы в область этики ИИ или же они могут определять функционирование ИИ как морального агента.

Уже на данном этапе разработки такой «этики добродетели», на наш взгляд, имеется ряд теоретических возражений. В первую очередь само выражение «этика добродетели» подразумевает вероятность множественных интерпретаций, если автор не специфицирует его использование. Эта ошибка стала довольно частым явлением, особенно в рамках прикладной этики, так как под этикой добродетели подразумевается концептуальный меланж, не всегда ограничивающийся философией Аристотеля[2], или же очень узкое понимание

[1] Этот список представляет только пример такого рода статей и отдельных глав монографий и его сложно назвать окончательным.

[2] При внимательном чтении можно обнаружить, что к лагерю философов, являющихся последователями «этики добродетели», могут быть причислены абсолютны все мыслители,

этической теории самого Аристотеля из-за точечно заимствуемых идей, которые могут быть вырваны из контекста, не отражать всей глубины его учения о морали или же попросту неправильно интерпретироваться. Хотя эта ошибка и может вызывать некоторое недоумение со стороны специалистов по аристотелевской и нео-аристотелианской философии, она не является столь грубой и, скорее, в очередной раз подчеркивает терминологическую сложность, связанную с тем, что мы включаем в понятие «этика добродетели». М.К. Нуссбаум справедливо обращала внимание на этот момент и настаивала на необходимости использования термина «нео-аристотелизм» для избежания неправильных ассоциаций и интерпретаций [Nussbaum, 1999, 200–201][3]; другие нео-аристотелианцы также отмечали эту особенность в своих трудах [Gardiner, 2005; Crisp, 1996, 2–3; Irwin, 1996, 39], но не настаивали на отказе от использования понятия «этика добродетели» в связи с его исторической привязкой к аристотелевскому этическому дискурсу. На наш взгляд, сложность с термином «этика добродетели» решается довольно просто – с помощью прописывания методологии исследования со стороны авторов. Отдельным положительным моментом в решении этой трудности может стать и повсеместное введение в оборот таких терминов, как «нео-аристотелизм», «старая и новая этика добродетели»[4] [Gardiner, 2005; Шохин, 2014, 19], современная этика добродетели[5].

Другая методологическая сложность, которая возникает из-за неправильной интерпретации аристотелевской этики в сфере ИИ, связана с тем, что авторы предлагают рассматривать систему искусственного интеллекта в качестве морального агента, хотя такая интерпретация не согласуется с воззрениями древнегреческого философа. В аристотелевской этике моральный агент – это субъект, имеющий объективную цель (telos) своего существования – стремление к благу, которое заключается в деятельности души согласно добродетели и способности суждения [Аристотель, 1983, 64, 1098a]. Достижение этой цели со стороны морального агента требует определенных усилий в реализации совершенства собственной природы, которое заключается в сознательном и последовательном упражнении в добродетелях – совершении нравственных поступков. Такое целеполагание может быть применимо к искусственному интеллекту лишь инструментально, так как он не обладает понятием блага, актуализируемого в достижении счастья, стремлением к последовательному самосовершенствованию и разумной частью души: как следствие,

---

которые хоть в каком-либо виде упоминали, предлагали свое учение о добродетелях или же просто перечисляли некоторые добродетели.

[3] Хотя термин «этика добродетели» и закреплен за учением Аристотеля, однако данное выражение часто используется, к примеру, для обозначения конфуцианского или стоического учения о добродетелях.

[4] Разделение на «старую» и «новую» этику добродетели является условным для большинства последователей нео-аристотелизма и используется по большей мере для обозначения хронологических рамок.

[5] Автор этой статьи под термином «(современная) этика добродетели» понимает аристотелизм и нео-аристотелизм в совокупности: в этой статье не рассматриваются концепты, имеющие различное толкование в аристотелизме и нео-аристотелизме, и, напротив, делается акцент на их общих идеях (см.: [Crisp, 1996, 5–6; Crisp, 2022, 1548]).

все принимаемые ИИ решения автоматичны и бессознательны согласно аристотелевской этической теории.

Два других проблемных аспекта связаны с формированием списка добродетелей и аристотелевским принципом середины. Проблема списка добродетелей заключается в том, что он не предписывает первичность одних добродетелей по отношению к другим и является открытым для пересмотра в зависимости от различных социально-культурных и исторических контекстов. Авторы часто не учитывают эту методологическую особенность и не конкретизируют основания для предпочтения одних добродетелей по отношению к другим. К. Аллен и его коллеги также уделяют пристальное внимание этой проблеме и подчеркивают, что обратный сценарий, согласно которому становится возможным создание всеохватывающего списка универсальных добродетелей, послужил бы основой для модели моральной агентности искусственного интеллекта [Allen et al., 2010, 258]. Другое ограничение списка добродетелей связано с возникающим конфликтом добродетелей. В случае социальных роботов этот момент имеет критическую важность: например, если в его дизайне ценностей будут заложены одновременно способность к состраданию и честность, то в случае возникновения этической дилеммы встанет вопрос о примате одной добродетели над другой. Детальное прописывание кода и машинное обучение не могут решить эту проблему в рамках (современной) этики добродетели, так как искусственный интеллект не обладает моральной агентностью и добродетелью фронезиса. Принцип середины также используется инструментально: авторы оставляют без должного внимания тот факт, что Аристотель настаивает на необходимости динамичного понимания добродетели [Аристотель, 1983, 85, 1106a30], и, напротив, чаще отсылают к арифметическому пониманию середины, не учитывая зависимость определения добродетели от ситуативных условий.

Еще один нюанс, на который мы считаем должным обратить внимание, – это отсутствие в нео-аристотелианском дискурсе какой-либо рефлексии по поводу этики ИИ. Единственная статья, опубликованная по этой тематике, вышла из-под пера одного из современных последователей нео-аристотелизма – Р. Криспа – в соавторстве с М. Константинеску [Constantinescu, Crisp, 2022]. В своей работе авторы рассматривают пример социальных роботов и отстаивают позицию, согласно которой социальные роботы не могут выступать в качестве моральных (и добродетельных) агентов, так как они не соответствуют трем критериям:

1. Социальные роботы не могут совершать добродетельные поступки. Понимание поступка является довольно комплексным в (современной) этике добродетели и оно не сводится к возможности тонкой и точечной настройки алгоритмов, которые смогут имитировать поступки человека. Интерпретация того или иного поступка как добродетельного напрямую зависит от характера морального агента, его намерений и мотивов, артикулируемых в процессе совершения поступка добродетелей и умения принимать решения в зависимости от контекста ситуации. В связи с этим невозможно составить предварительный перечень правил для социальных роботов, который предполагал бы возможность ситуативной вариативности. Эти доводы могут быть частично

опровергнуты, и поступки социальных роботов могут внешне напоминать поступки человека: Крисп и Континеску не отрицают ту возможность, что роботы будущего смогут совершать поступки на основании заложенных в них знаний и делать рациональный выбор, учитывая различные контекстуальные условия [Constantinescu, Crisp, 2022, 1550]. Однако нам кажется, что такой сценарий все равно не может быть сопоставим с (нео-)аристотелианским дискурсом, так как он не учитывает аспект морального развития, связанного с изменением характера морального агента.

2. Социальные роботы не являются автономными и не обладают моральным знанием. Безусловно, в случае необходимости принятия какого-либо решения система искусственного интеллекта будет принимать во внимание различные аргументы «за» и «против» перед вынесением окончательного вердикта. Если же мы рассматриваем процесс принятия морального решения о правильности того или иного поступка с точки зрения (нео)-аристотелизма, то акцент переносится с морального решения на морального агента, так как добродетельный поступок должен быть проявлением характера морального (добродетельного) агента. Другими словами, просто совершение поступка, согласующегося с отдельными видами добродетелей, недостаточно, так как (современную) этику добродетели больше интересует, является ли этот поступок результатом актуализации нравственных качеств морального агента. В настоящий момент нейронные сети и машинное обучение способствуют увеличению уровня автономии ИИ и, как следствие, генерации им неочевидных схем и решений. Однако эта автономия основана на работе с огромными наборами данных, т.е. речь идет о некоторой предзаданной ситуации. Таким образом, все принимаемые системой искусственного интеллекта решения сводятся к простым подсчетам всех вариантов развития какой-либо ситуации и выбору наиболее удачного и полезного из них. Моральное знание в этике (нео-)аристотелизма не сводится к таким подсчетам, умение работать с огромными пластами данных не является достаточным условием для получения статуса морального агента.

3. Социальные роботы не обладают способностью суждения. Фронезис – ключевая добродетель для аристотелевской этики, так как она подразумевает умение наилучшим образом поступать в конкретных обстоятельствах [Аристотель, 1983, 176, 1140a25]. Другая отличительная особенность (современной) этики добродетели состоит в том, что она предлагает сместить акцент с самого поступка на морального агента: следовательно, выбор, как правильно поступать в той или иной ситуации, является проявлением характера морального агента и артикуляцией присущих ему добродетелей. Как было отмечено ранее, искусственный интеллект не обладает моральной агентностью в рамках аристотелевской этики, поэтому его обладание способностью суждения инструментально и напрямую коррелирует с предзаданным набором характеристик, которые были заложены его разработчиком и на основе которых он осуществляет ситуативный выбор. Некоторые исследователи пытаются обойти имеющееся теоретическое ограничение с помощью использования нового понятия «функционального или искусственного фронезиса» применительно к ИИ [Sullins, 2021, 141]: в данном контексте речь идет о попытке

разработчиков создать некоторую форму моральной эпистемологии, предполагающую для интеллектуальных систем способность к обучению на примере различных этических кейсов. На наш взгляд, имеются два возражения против такого подхода.

Во-первых, фронезис не является формой пропозиционального знания. Никакой алгоритм не может составить точную инструкцию о том, как совершать правильные поступки согласно (современной) этике добродетели. Это невозможно ввиду теоретической сложности этой нормативной теории. Как было указано ранее, она смещает акцент с поступка на характер морального агента, т.е. поступок не будет считаться добродетельным (даже в случае благих последствий), если он является результатом неправильного рассуждения и актуализации пороков. Более того, последователи неоаристотелианской этики отмечают важность изучения эмоций и их влияние как на формирование морального характера агента, так и на принятие моральных решений.

Во-вторых, введение в научный дискурс понятия «функционального или искусственного фронезиса» связано с критикой антропоцентризма в сфере этики ИИ: этические категории, используемые для описания нравственного поведения и характера человека и задающие моральный идеал, не могут использоваться для социальных роботов. Тезис об ограниченности применения антропоцентрического аргумента широко распространен в дискуссиях, посвященных этике ИИ, и выдвинуть против него весомые контраргументы затруднительно. Но можно задать и каверзный вопрос критикам антропоцентрического аргумента: почему мы должны говорить именно о морали применительно к ИИ, а не о «функциональной морали» [Sharkey, 2020, 289; Wallach, Allen, 2009, 26]? Последовательная критика антропоцентрического аргумента противоречит стремлению рассматривать ИИ как морального агента и накладывать на него спектр наших представлений о благе, учитывая тот факт, что нео-аристотелианская этика накладывает ряд ограничений на возможность распространения своих принципов на ИИ. Использование понятий (современной) этики добродетели становится метафоричным: отвлеченно можно рассуждать и о существовании добродетелей у роботов, однако такие рассуждения не будут иметь ничего общего с философскими концептами. Более того, такая позиция представляется сомнительной с точки зрения академической этики: фрагментация нормативной теории и некорректное использование ее понятийного аппарата для обоснования собственной исследовательской позиции приводят к искажению смыслов как самой теории, так и построению новой концепции на зыбком фундаменте.

Таким образом, сугубо инструментальное понимание и использование подхода «этики добродетели» в сфере ИИ, к которому апеллируют современные авторы, содержательно расходится с этикой аристотелизма и нео-аристотелизма. Мы не предполагаем, что современные исследователи в области практической философии должны предварять свои работы историко-философской реконструкцией и не говорим о неправильности их этических выводов о функционировании систем искусственного интеллекта. Однако такая ситуация усиливает тенденцию к возникновению герменевтических ошибок и постулирует более комплексный вопрос о соотношении моральной теории

и практики: если отклонение от ключевых теоретических постулатов может оправдываться только инструментальным характером этики ИИ, то непонятен смысл существования и степень влияния моральной теории на практику. Как нами было указано ранее, в таком случае использование философских концептов становится метафоричным. Несмотря на указанные нами трудности в интерпретации идей (современной) этики добродетели при обращении к вопросам прикладной этики, можно выделить два направления в рамках этики неоаристотелизма – экземпляризм (exemplarism) и подход возможностей (capability approach), которые предлагают иное осмысление феномена искусственного интеллекта.

Экземпляристский подход в сфере ИИ

Экземпляристский подход в оценке искусственного интеллекта является наиболее распространенным для нео-аристотелианского дискурса. Теоретические основания этого подхода заложены американским философом Л.Т. Загзебски. Его ключевая идея состоит в необходимости следования в своем нравственном поведении моральным образцам [Zagzebski, 2010, 54]. В качестве моральных образцов могут выступать не только выдающиеся личности, которые являются носителями моральных и интеллектуальных совершенств, но и различные нарративы. Такая концептуальная модель не должна истолковываться превратно: экземпляризм не предполагает имитации поведенческих паттернов морального образца, напротив, этап подражания соотносится с идеей аристотелевского мимесиса. Идея мимесиса предполагает не только возможность получения первых представлений о вещах, но также обладает пропедевтической функцией: создание сильной эмоциональной связи между моральным образцом (или нарративом) и индивидом способствует его большей вовлеченности и желанию походить на предмет подражания. Однако такое подражание нисколько не умаляет этапа рефлексии: они тесно связаны между собой в экземпляристской схеме Загзебски, так как подражание предваряет рефлексию. Этот момент, часто превратно истолковываемый, имеет рациональное основание: подражание дает первичные представления о добродетелях и нравственном поведении, без которых невозможна последующая рефлексия. Результатом рефлексии являются формирование критического отношения к образцам и нарративам и умение контекстуального применения полученных знаний, что проявляется в совершении правильных поступков.

Экземпляристская оптика с ее акцентом на моральном агенте переносится и в область машинной этики, что предполагает фундаментально новый подход к пониманию роли и значения социальных роботов в жизни человека. В отличие от обозначенных нами в первом разделе подходов, которые фрагментарно используют идеи (современной) этики добродетели и некорректно приписывают возможность моральной агентности системам искусственного интеллекта, последователи экземпляризма рассматривают их только с точки зрения образцов и нарративов. Другими словами, экземпляристская модель ИИ нацелена на понимание того, как взаимодействие с социальными роботами может

поспособствовать качественному улучшению жизни, созданию условий для морального развития и совершенствованию добродетельного характера индивидов [Cappuccio et al., 2020, 5; Cappuccio et al., 2021, 1]. Сфера применения экземпляристской модели в отношении социальных роботов не ограничена, хотя и можно выделить отдельные ниши, в которых их применимость является наиболее успешной и интересной с точки зрения этики: к ним относятся робототехника для обучения детей и роботы, выполняющие ряд психотерапевтических или тьюторских функций. Наиболее показательным примером позитивного влияния социального робота на моральную рефлексию является опыт детского взаимодействия с роботом-черепахой Шелли [Cappuccio et al., 2020, 6–7]: цель этого робота состоит в снижении у детей уровня агрессии и выработке терпеливого отношения. Непродолжительные эксперименты с детьми приблизительно одной возрастной категории (около 13 лет) позволили сделать вывод о том, что их игровое взаимодействие с роботом привело к формированию устойчивого паттерна бережного отношения к животным; на наш взгляд, в рамках проведенного эксперимента можно также сделать вывод об обучении детей коммуникативным навыкам и выработке у них просоциального поведения, так как игра с роботом организовывалась в группах и было отмечено, что групповая интеркоммуникация способствовала ограничению и снижению уровня деструктивного поведения у отдельных членов группы.

Безусловно, к этому исследованию может возникнуть ряд критических замечаний, так как оно не является лонгитюдным и не дает подробной информации о референтной группе. Другой проблемный аспект взаимодействия между человеком и роботом может касаться контекстуальных условий: так, в случае игры детей с роботом-черепахой Шелли исследователи могли варьировать период времени, в течение которого робот мог прятаться в панцире, если дети были склонны к проявлению агрессии по отношению к нему; если этот период времени был слишком коротким, то дети интерпретировали поведение робота как игру и воспринимали это как поощрение актов агрессии. Это поднимает довольно нетривиальный вопрос о возможности образцов и нарративов усиливать у морального агента проявления деструктивного поведения и закреплять его в виде сформировавшихся пороков. Эта критическая ремарка применима не только к экземпляристской модели в области ИИ, но затрагивает и теоретические основания экземпляризма: в этом случае универсального решения нет, так как отношение к роботам является в какой-то степени отражением нашего морального облика. Наличие добродетели фронезиса и ее применение в различных контекстах взаимодействия с ИИ может способствовать изменению ситуации в положительном ключе, если моральный агент воспринимает это взаимодействие не как сугубо автоматическое и требующее от него механического воспроизведения «правильных» действий, а рефлексирует по поводу полученного опыта [Cappuccio et al., 2019, 14].

В рамках экземпляристского подхода довольно остро стоит проблема асимметричных отношений между моральным агентом и социальным роботом. Эта проблема основана на антропоцентрическом аргументе, который порождает своеобразный парадокс: с одной стороны, социальные роботы в большинстве

случаев сконструированы похожими на людей, чтобы вызывать определенный эмоциональный отклик; с другой стороны, цель социального робота заключается в служении человеку, в связи с чем может возникнуть ситуация когнитивного диссонанса [Cappuccio et al., 2019, 18]. Внешнее сходство робота с человеком и возможность его использования для удовлетворения собственных нужд[6] может стать толчком к возрождению модели отношений «раб – господин». Такие ситуации становятся обратной стороной медали для экземпляристского подхода, так как могут привести к укоренению негативных поведенческих паттернов. Одним из решений этой проблемы может стать точечная настройка алгоритмов социальных роботов, которые смогут стимулировать и поощрять добродетельное поведение морального агента по отношению к ним и, наоборот, случаи проявления деструктивного или девиантного поведения переводить в регистр моральных уроков [Sparrow, 2020, 7]: последний аспект несет в себе не только пропедевтическую, но и корректирующую функцию.

Хотя экземпляристская модель применительно к искусственному интеллекту и имеет ряд трудностей, она является довольно многообещающим направлением для отдельных групп исследователей, которые работают в области социальной робототехники. На данном этапе главное преимущество этого подхода – его ориентация на практику и реалистичность: экземпляризм не задается абстрактными вопросами о том, сможет ли социальный робот обрести моральную агентность в ближайшем будущем и к каким последствиям приведет этот факт, а исходит из понимания ситуации «здесь и сейчас». Благодаря этому экземпляризм не только остается в теоретических рамках нео-аристотелианской этики, но и подходит к проблеме влияния ИИ на жизнь человека с позиции морального индивида.

Подход возможностей в сфере ИИ

Другая нео-аристотелианская теоретическая рамка, позволяющая по-новому взглянуть на роль искусственного интеллекта в жизни людей, – это нуссбаумианский подход возможностей. Для начала мы реконструируем его основные идеи, а затем покажем, какие трансформации они претерпевают в области этики искусственного интеллекта.

В первую очередь важно провести линию разграничения между подходом возможностей и подходом человеческого развития: последний предлагает интерпретировать возможности как некоторые показатели, применяемые при проведении компаративистского анализа. Нуссбаум расширяет это понимание возможностей, прибавляя к нему необходимость обеспечения уважения человеческого достоинства и осуществления минимальной социальной справедливости. Другая отличительная черта подхода заключается в том, что возможности – или субстанциальные свободы – осуществляются каждым человеком по его усмотрению. В связи с этим нуссбаумианский подход не претендует на статус универсальности; напротив, он провозглашает ценностный плюрализм,

[6] В этом контексте речь идет о проявлении деструктивного или девиантного поведения со стороны отдельного индивида.

что позволяет видоизменять его в зависимости от заданных условий. Последнее важное разграничение проводится между возможностями и функционированием: возможности предполагают свободу выбора или отказа от них, а функционирование является производной от результата нашего выбора между осуществлением возможностей [Nussbaum, 2011, 24–25]. Подход артикулирует два вида возможностей – комбинированные и внутренние (combined and internal capabilities). Комбинированные возможности – это набор внешних возможностей, осуществление которых зависит от социально-политических и экономических условий. Внутренние возможности связаны с личностными склонностями, навыками, убеждениями, которые предопределяются или развиваются в зависимости от наличествующих условий. К примеру, возможность (свобода) самовыражения – это внутренняя возможность, а внешний запрет на отдельные формы самовыражения – комбинированная возможность. Согласно Нуссбаум, реализация минимальной социальной справедливости и обеспечение уважения человеческого достоинства в политических сообществах достижимы благодаря десяти основополагающим возможностям [Ibid., 33–34][7].

Для начала стоит отметить, что нуссбаумианский подход возможностей, как и экземпляристский подход, не акцентирует внимание на моральной компоненте искусственного интеллекта; он сконцентрирован вокруг морального индивида (и человеческого сообщества в целом) и качественных характеристик его жизни, которая претерпевает определенный ряд изменений из-за внедрения различных технологий, связанных с функционированием систем искусственного интеллекта. В теоретическом срезе важным становится определение через свободу выбирать или отказываться от различных возможностей: этот аспект используется и по отношению к технологиям ИИ. Следовательно, первым пунктом является доступность технологий искусственного интеллекта для всех членов человеческого сообщества [Buccella, 2023, 1143], а из амбивалентного характера возможностей следует второй пункт – необходимость соблюдения права на использование или отказ от использования ИИ [Buccella, 2023, 1148; London, Heidari, 2023, 8]. Третий пункт кроется в создании таких условий организации социально-политической жизни общества, при которых согласие или отказ на использование технологии ИИ не нарушит уровень минимальной социальной справедливости и не приведет к дискриминации отдельных групп. На наш взгляд, последний тезис представляется наиболее затруднительным для практической реализации, так как повсеместное распространение интеллектуальных систем неизбежно приводит к изменению традиционных форм социальной жизни и внедрению технологий во всех сферах жизни. Решение этого вопроса возможно только методами правового

[7] К этим возможностям относятся: (1) жизнь; (2) физическое здоровье; (3) телесная неприкосновенность; (4) чувства, воображение и мысли; (5) эмоции; (6) практический разум; (7) членство – возможность социального взаимодействия и обеспечение человеческого достоинства; (8) другие виды живых существ; (9) игра; (10) осуществление контроля над собственной средой – участие в политической жизни, обладание собственностью и право вступать в трудовые отношения.

регулирования и через учреждение государственных стандартов в области применения технологий искусственного интеллекта.

Данный подход представляется плодотворным в связи с тем, что он комбинирует человеко-ориентированный и риск-ориентированный подходы. Благодаря этому синтезу технологии ИИ и их развитие напрямую связаны с последствиями, которые они оказывают на человеческую жизнь. Эта связь работает и в обратном направлении: в случае возникновения негативных эффектов при внедрении интеллектуальных систем, они должны купироваться и влечь за собой последующее изменение ценностного дизайна ИИ [Oosterlaken, 2013, 209–210].

По мнению авторов, изменения в связи с распространением искусственного интеллекта коснулись практически всех базовых возможностей, о которых пишет Нуссбаум [Buccella, 2023, 1146–1148; London, Heidari, 2023, 12–16]. В зависимости от выбранной возможности можно наблюдать как положительные, так и негативные изменения. К примеру, один из пунктов возможности осуществления контроля над собственной средой предполагает право вступления в трудовые отношения и регулирует все вопросы, связанные с соблюдением прав работников. Процесс поиска сотрудника компаниями алгоритмизирован: первичный отбор кандидатов осуществляется с помощью технологий на базе искусственного интеллекта, что помогает сокращать время поиска и исключать из выборки неподходящих соискателей вакансии. Идентичная процедура распространяется и на компании, так как соискатели могут с помощью настраиваемых вручную алгоритмов расширять или сужать поисковой запрос. С первого взгляда такое положение дел выглядит выигрышным для обеих сторон. Однако компании при поиске сотрудника могут задавать критерии поиска, дискриминирующие сотрудника по признакам половой, расовой и религиозной принадлежности: этот фактор создает определенные риски для правильного функционирования возможности.

Подход возможностей в контексте этики ИИ на текущий момент находится на этапе становления: исследования в этой области весьма немногочисленны, хотя и обещают стать востребованными в связи с необходимостью соотнесения человеко-ориентированного и риск-ориентированного подходов.

Заключение

Принципы (современной) этики добродетели, которая является альтернативным нормативным подходом, противопоставленным деонтологии и консеквенциализму, комплексны и имеют свои интерпретативные особенности в сфере прикладной этики, в частности, применительно к искусственному интеллекту. В строгом смысле полное соответствие этики ИИ аристотелевским и нео-аристотелианским идеям невозможно, так как в этом случае не выполняется целый ряд критериев: невозможность рассматривать ИИ в качестве морального агента, невозможность дискурса о добродетелях, отсутствие фронезиса, неправильное понимание аристотелевского учения о середине и другие. Сугубо инструментальное понимание (современной) этики добродетели, на наш

взгляд, создает ряд теоретических и герменевтических трудностей именно для философского исследования в области этики искусственного интеллекта; как не раз подчеркивалось, указанные трудности снимаются в случае прописывания методологии исследования и указания на смысловые особенности использования тех или иных философских концептов. Несмотря на имеющиеся историко-философские методологические ограничения, два нео-аристотелианских подхода – экземпляризм и подход возможностей, предлагают качественно иной взгляд на роль искусственного интеллекта в жизни человека и задают новую перспективу для будущих исследований.

Can Artificial Intelligence Be "Ethical"? The Outlook of (Contemporary) Virtue Ethics

Valeriia K. Kudriashova

HSE University. 21/4, 1 Staraya Basmannaya Str., Moscow, 105006, Russian Federation.

ORCID: 0000-0002-8147-2530
e-mail: vkkudryashova@hse.ru

The article examines the methodological limitations of the application of (contemporary) virtue ethics in the field of artificial intelligence. The first limitation lies in the theoretical vagueness of "virtue ethics", since it is only instrumentally interpreted in case of applied ethics' issues by many researchers: for instance, they often refer only to the principle of golden mean that helps to understand the nature of virtues and to articulate them for AI. This approach is fundamentally wrong, since it selectively considers the main ideas of (contemporary) virtue ethics and uses only those that approve it. The second limitation is that in Aristotelianism and neo-Aristotelianism AI cannot be considered as a moral agent *sensu stricto*. In this regard, AI is a system that can only imitate the behaviour of a moral agent who is the bearer of a certain set of virtues. Despite the initial negative diagnosis, there are still two approaches in neo-Aristotelianism that can bring light to the possibility of AI ethics. Exemplarism allows us to take a different look at the role of social robots in human life: they can be seen as moral exemplars or narratives that can contribute to the moral development of their owners and change their moral portrait for the better or worse. The capability approach created by M.K. Nussbaum also focuses on the impact of AI on the quality of human life: in this context, we are highlighting what changes basic human capabilities can undergo and how the minimal level of social justice and the respect of human dignity are ensured while using AI technologies.

Keywords: virtue ethics, neo-Aristotelianism, moral exemplarism, capability approach, AI, ethics of AI

Литература / References

Аристотель. Никомахова этика / Пер. Н.В. Брагинской // *Аристотель*. Соч.: в 4 т. Т. 4. М.: Мысль, 1983. С. 295–374.

Aristotle. "Nikomahova etika" [The Nicomachean Ethics], trans. by N.V. Braginskaya, in: Aristotle, *Sochineniya* v 4 t. [Works in 4 Vols.], Vol. 4. Moscow: Mysl' Publ., 1983, pp. 295–374. (In Russian)

Шохин В.К. Этика добродетели – «старая» и «новая» // *Шохин В.К.* Агатология: Современность и классика. М.: Канон+, 2014. С. 19–49.

Shohin, V.K. "Etika dobrodeteli – «staraya» i «novaya»" [Virtue Ethics – "Old" and "New"], in: V.K. Shohin, *Agatologiya: Spvremennost' i klassika* [Agathology: Modernity and Classics]. Moscow: Kanon+ Publ., 2014, pp. 19–49. (In Russian)

Allen, C., Varner, G., Zinser, G. "Prolegomena to Any Future Artificial Moral Agent", *Journal of Experimental & Theoretical Artificial Intelligence*, 2000, Vol. 12, No. 3, pp. 251–261.

Berberich, N., Diepold, K. *The Virtuous Machine – Old Ethics for New Technology?* [https://arxiv.org/pdf/1806.10322.pdf, accessed on 17.03.2024].

Boddington, P. *AI Ethics*. London: Springer, 2023.

Buccella, A. "AI for All Is a Matter of Social Justice", *AI and Ethics*, 2023, No. 3, pp. 1143–1152.

Cappuccio, M., Peeters, A., McDonald, W. "Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition", *Philosophy & Technology*, 2019, pp. 1–23.

Cappuccio, M., Sandoval, E.B., Mubin, O., Obaid, M., Velonaki, M. "Can Robots Make Us Better Humans? Virtuous Robotics and the Good Life with Artificial Agents", *International Journal of Social Robotics*, 2020, pp. 1–16.

Cappuccio, M., Sandoval, E.B., Mubin, O., Obaid, M., Velonaki, M. "Robotics Aids for Character Building: More than Just Another Enabling Condition", *International Journal of Social Robotics*, 2021, No. 13, pp. 1–5.

Coleman, K.G. "Android Arete: Toward a Virtue Ethic for Computational Agents", *Ethics and Information Technology*, 2001, No. 3, pp. 247–265.

Constantinescu, M., Crisp, R. "Can Robotic AI Systems Be Virtuous and Why Does This Matter?", *International Journal of Social Robotics*, 2022, No. 14, pp. 1547–1557.

Crisp, R. "Modern Moral Philosophy and the Virtues", *How Should One Live? Essays on the Virtues*, ed. by R. Crisp. Oxford: Oxford UP, 1996, pp. 1–18.

Gardiner, S.M. *Virtue Ethics, Old and New*. Ithaka and London: Cornell UP, 2005.

Hagendorff, Th. "A Virtue-Based Framework to Support Putting AI Ethics into Practice", *Philosophy & Technology*, 2022, pp. 1–24.

Irwin, T.H. "The Virtues: Theory and Common Sense in Greek Philosophy", *How Should One Live? Essays on the Virtues*, ed. by R. Crisp. Oxford: Oxford UP, 1996, pp. 37–56.

Klincewicz, M. "Artificial Intelligence as a Means to Moral Enhancement", *Studies in Logic, Grammar and Rhetoric*, 2016, No. 48, pp. 171–187.

Li, O. "Problems with "Friendly AI", *Ethics and Information Technology*, 2021, pp. 1–8.

London, A.J., Heidari, H. *Beneficent Intelligence: A Capability Approach to Modeling Benefit, Assistance, and Associated Moral Failures through AI Systems* [https://arxiv.org/abs/2308.00868, accessed on 29.02.2024].

Nussbaum, M.C. *Creating Capabilities: The Human Development Approach*. Cambridge, Massachusetts; London: The Belknap Press of Harvard UP, 2011.

Nussbaum, M.C. "Virtue Ethics: A Misleading Category?", *The Journal of Ethics*, 1999, No. 3, pp. 163–201.

Oosterlaken, I. *Taking a Capability Approach to Technology and Its Design: A Philosophical Exploration*. Doctoral Dissertation. Enschede, 2013.

Sharkey, A. “Can We Program or Train Robots to Be Good?”, *Ethics and Information Technology*, 2020, No. 22, pp. 283–295.

Sparrow, R. “Why Machines Cannot Be Moral”, *AI & Society*, 2020, pp. 1–9.

Sullins, J.P. “Artificial Phronesis: What It Is and What It Is Not”, *Science, Technology, and Virtues: Contemporary Perspectives*, eds. E. Ratti, Th.A. Stapleford. Oxford: Oxford UP, 2021, pp. 136–146.

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

Zagzebski, L.T. “Exemplarist Virtue Theory”, *Metaphilosophy*, 2010, Vol. 41, No. 1–2, pp. 41–57.