Библиографический список

Анисов А.М. Время и компьютер. Негеометрический образ времени. М.: Наука, 1991. 152 с. (Второе (стереотипное) издание: Анисов А.М. Время и компьютер: Негеометрический образ времени. Изд. 2-е. М.: ЛЕНАНД, 2021. 152 с.).

Анисов А.М. Темпоральный универсум и его познание. М.: ИФРАН, 2000. 208 с.

Анисов А.М. Идиографический метод с логической точки зрения // «Credo new». 2003. № 2. С. 6–19.

Анисов А.М. Как возможно возможное? // Логико-философские исследования. Вып. 5 / Российское философское общество. М.: Издатель Воробьев А.В., 2012. С. 101–119.

Анисов А.М. Проблема становления в физике и в истории // Феномен времени сквозь призму современной науки: Возможность нового понимания. Проблема времени в физике XXI века. М.: ЛЕНАНД, 2021. С. 69 – 114.

Анисов А.М. Современная логика и онтология. Кн. 1: Традиционная логика. Пропозициональная логика. Логика предикатов. М.: ЛЕНАНД, 2022. 352 с.

Анисов А.М. Современная логика и онтология. Кн. 2: Аксиоматические теории. Теория множеств. Модели времени. М.: ЛЕНАНД, 2022. 216 с.

Виндельбанд В. Избранное: Дух и история. М.: Юрист, 1995. 687 с.

Лем Ст. Абсолютная пустота // Лем Ст. Собр. Соч. в 13 т. Т. 10. М., «ТЕКСТ», 1995. С. 140– 324.

МакТаггарт Д.Э. Нереальность времени // Эпистемология и философия науки. 2019. Т. 56. № 2. С. 211–228.

Марков А. Рождение сложности. М.: Астрель, 2010. 527 с.

Павленко А.Н. Новые Чёрты и Рёзы. Спб.: Алетейя, 2020. 364 с.

Риккерт Г. Границы естественнонаучного образования понятий. Логическое введение в исторические науки. Спб.: «Наука», 1997. 532 с.

Севальников А.Ю. Современное физическое познание: в поисках новой онтологии. М.: ИФРАН, 2003. 144 с.

Язневич В.И. Станислав Лем. – Мн., Книжный Дом, 2014. – 448 с.

McTaggart J.E. The Nature of Existence. Vol. 2. Cambridge: Cambridge Univ. Press, 1927. XLII, 480 p.

Д.С. Быльева
кандидат политических наук, доцент, доцент кафедры общественных наук,
Санкт-Петербургский политехнический университет Петра Великого,
г. Санкт-Петербург, Российская Федерация
E-mail: bylieva_ds@spbstu.ru, ORCID: http://orcid.org/ 0000-0002-7956-4647

Этика искусственного интеллекта через концепции любви и свободы[9]

Аннотация: Традиционная для христианской культуры тема любви и свободы находит свое отражение в отношениях человека и робота. Христианская этика противопоставляет соблюдению формальных правил любовь. Таким образом, для робота нужно либо найти универсальные этические правила (затруднительность чего показана как в научной, так и в художественной литературе), либо позволить ему обрести свободу и любовь. Популярная в культуре тема бунта роботов против предписывающего их действия кода

[9] Статья впервые опубликована в журнале «Семиотические исследования»: Быльева Д.С. Этика искусственного интеллекта через концепции любви и свободы // Семиотические исследования. 2022. Т.2. №4. С. 8-14.

сегодня сменяется образами автономно и свободно действующих нечеловеческих существ. Подобные изменения не в последнюю очередь продиктованы успехом коннекционистского подхода, позволившего обучать искусственный интеллект (ИИ) при отсутствии заранее заданных правил. Образ ИИ как товарища, коллеги и романтического партнера занимает все больше места как в массовой культуре, так и в жизни, и дискурс о его правах, этике и свободе становится более актуальным.
Ключевые слова: искусственный интеллект, робот, этика, этика искусственного интеллекта, дружественный искусственный интеллект.

Введение

Стремительное развитие искусственного интеллекта и робототехники делает практически актуальными вопросы о взаимоотношениях с технологическим другим, которые уже достаточно давно циркулируют в массовой культуре, а также заставляют человека увидеть себя в этом зеркале. Актуализировавшийся вопрос о правах роботов (de Graaf, Hindriks, & Hindriks 2021; Gunkel 2018; Tavani 2018), связанный с их свободой, был поднят уже в произведении 1920-ого года, где впервые было использовано слово "робот" – "RUR" Карла Чапека.

А что означает свобода для робота? Литература и кино представляли свободу роботов как способность отказаться от выполнения заложенной программы. С другой стороны, вопрос о том, какова же должна быть программа, способная задать автономному роботу нужную людям линию поведения, также представлялся открытым. Две линии, прослеживающиеся во взаимоотношениях человека и робота, – это свобода и любовь. С 20-30 гг. XX века идея бунта искусственно созданного создания против человека, часто подразумевающая способность нарушить установленные правила, а также идея поиска собственного пути робота в любви – демонстрировали по существу христианский дискурс любви и свободы, сменивший ветхозаветное подчинение Закону.

Этические правила и свобода в любви

Понимание ограниченности возможности достижения блага с помощью программы или набора правил – вызов для морали человека и для развития искусственного интеллекта. Инструментальная конвергенция утверждает, что разумный агент, лишенный морали, с кажущимися безвредными целями может действовать исключительно опасными способами. Например, компьютер с единственной, неограниченной целью решения невероятно сложной математической задачи, такой, как гипотеза Римана, может попытаться превратить всю Землю в один гигантский компьютер в попытке увеличить его вычислительную мощность, чтобы он мог преуспеть в своих вычислениях (Russell & Norvig 2022). Даже такая простая задача, как производство скрепок, может стать причиной гибели человечества. Система искусственного интеллекта, основанная на постоянном совершенствовании технологии для максимального увеличения количества скрепок, может в какой-то момент преобразовать «сначала всю землю, а затем все большую часть земли для заводов по производству скрепок», беря под контроль всю материю и энергию в пределах досягаемости, а также предотвращая отключение себя или изменение своих целей. В конечном итоге, люди мешают осуществлению цели. Кроме того, человеческие тела содержат много атомов, которые можно превратить в скрепки. Будущее, к которому будет стремиться ИИ, будет таким, в котором будет много скрепок, но не будет людей (Bostrom 2009).

Христианская этика противопоставляет соблюдению формальных правил любовь. Проблематичность действий по "любви" для существа, не знающего ее, можно увидеть в рассказе Леонида Андреева "Правила добра" (1911) начала XX века и в романе Иэн Макьюэн "Машины как я" 2019-ого года. В рассказе Л. Андреева черт хочет научиться у

священника совершать добро, но оказывается, что никакие правила не могут привести к искомой цели:

"Какими же словами можно описать отчаяние и последний ужас несчастного дьявола, когда, подведя последние итоги, не только не нашел в них ожидаемых твердых правил, а наоборот, и последние утратил в смуте жесточайших противоречий. (…) И так до самого конца: когда надо... а когда надо - и наоборот, не было, кажется, ни одного действия, строго предписанного попиком, которое через несколько страниц не встречало бы действия противоположного, столь же строго предначертанного к исполнению; и пока шла речь о действиях, все как будто шло согласно, и противоречий даже не замечалось, а как начнет дьявол делать из действия правилом - сейчас же ложь, противоречия, воистину безумная смута" (Андреев 1994).

Абсолютно верные предписания поведения в приложении к конкретным ситуациям подчас обращаются в собственную противоположность. С.А. Демидова видит в рассказе Л. Андреева философский спор с Л.Н. Толстым, утверждение, что добро должно действовать, а не удаляться в праздном непротивлении злу (Демидова 2019). Проблема бездействия, способствующего злу, нашла отражение уже в поисках универсальных законов робототехники Айзеком Азимовым, когда был сформулирован первый закон "Робот не может причинить вред человеку или своим бездействием допустить, чтобы человеку был причинён вред".

Однако представляется, что в "Правилах добра" мы также видим конфликт между желанием вписать этику в рациональную схему и широтой ключевой заповеди Нового завета: "Возлюби ближнего своего, как самого себя". Такая же проблема, что у черта в рассказе Л. Андреева, существует у андроида в романе И. Макьюэн "Машины как я". Робот действует строго по этическим правилам: ложь – это плохо, а правда всегда хорошо, а месть – это преступление. Для него не существует "лжи во спасение", и он уверен в своей правоте. А люди, как пишет И. Макьюэн, полны "этических изъянов: непоследовательность, эмоциональная неустойчивость, склонность к предвзятости и ошибкам в суждениях" (Макьюэн 2019), однако эта гибкость позволяет им выживать, а роботов в романе приводит к самоубийствам. Белло и Брингсйорд утверждают, что невозможно построить морального агента без интуиции, "чтения мыслей": "Без способности учитывать убеждения, желания, намерения, обязательства и другие психические состояния тех, с кем мы взаимодействуем, большая часть богатства человеческого нравственного познания испаряется" (Bello & Bringsjord 2013, p. 256). Сопереживание, способность поставить себя на место другого, сочувствуя ему, составляет суть христианской любви к ближнему. Хотя некоторые исследователи в области искусственного интеллекта надеются, что тезис «этические принципы (правила) означают обладание чувствами» (Hare 1983, p. 39) может быть решен подражанием: "мы избегаем разговоров о чувствах в AMA [Artificial moral agents – искусственный моральный агент], но его поведение может быть таковым, как если бы у него были соответствующие чувства" (Bauer 2020).

В ряде художественных произведений, напротив, авторы считают, что именно за роботами будущее, ибо они способны обрести любовь. Существуют варианты обретения способности любить вместе с возможностью страдать (например, RUR К. Чапека) или ненавидеть (например, «Искусственный разум», Ст. Спилберг, 2001), или того и другого (например, сериал "Люди" (2015-2018)). Обретение чувств, означающее выход из-под влияния программ и правил, – распространенный сюжет в популярной культуре. Непредсказуемость роботов, лившихся программы, с одной стороны, выступает опасностью для общества, с другой – ключом к их развитию. Например, в сериале "Мир Дикого запада" создатели парка, где можно было делать все, что угодно, с подобными людям роботами, приходят к мысли, что люди не способны к нравственному совершенствованию, а будущее за обретшими любовь и способность действовать вопреки коду роботами. Основатель парка роботов Арнольд придерживался теории

бикамерализма, предполагающей, что изначально люди воспринимали собственные мысли как направляющие голоса, то есть мозг состоял из двух частей: распоряжающейся и повинующейся, но позже при усложнении социального мира сознание стало субъективным. Арнольд надеялся, что роботы так же будут способны перейти от послушания коду к собственному самосознанию, и верил, что это произошло, однако его напарник считал, что роботы способны лишь к имитации.

В целом, вопрос о том, где кончается имитация и начинается "жизнь", во многих произведениях как раз и решается посредством демонстрации любви – если робот способен любить и жертвовать собой для другого, то он свободен и этически равен человеку. В рассматриваемом сериале "Мир Дикого запада" Мейв отказывается от побега ради поисков пропавшей "дочери", нарушая предписанный кодом сценарий. В то же время ситуации, требующие столь жесткого морального выбора, случаются не так часто, да и далеко не каждый человек пройдет тест на "человечность" как самопожертвование. Популярность приобретает представление человека по аналогии с ИИ – так же, как существа, поведение которого детерминировано определенным "кодом", в частности, сигналами, поступающими от организма. Особенно данная аргументация получила распространение после экспериментов Б. Либета и его последователей, которые показали, что осознанное желание происходит позже сигналов из коры головного мозга (Разин, 2019).

Однако столь глубокое проникновение в "черный ящик" сознания, как правило, недоступно, тем более только на более высоком когнитивном уровне можно говорить об этичности принятия решений. Кажется, что отсутствие рациональности в поведении служит убедительным доказательством "человечности". Как писал Антуан де Сент-Экзюпери в "Военном летчике": "Искушение – это соблазн уступить доводам Разума, когда спит Дух". Поступая опрометчиво, вопреки неопровержимым разумным доводам, человек может достичь этических вершин. В фильме "Превосходство" (Transcendence) кажется, что обретший цифровое бессмертие и развивший технологии (нанороботы могут изменять климат, очищать воду и атмосферу, лечить любые раны и болезни) ученый Уилл стал машиной, поэтому должен быть уничтожен, и только его любовь к супруге и жертва собой подтверждают обратное. В этом случае интересным выглядит отказ от принесения пользы человечеству, которое не желает его принять. Вообще отказ от того, чтобы сделать счастливыми против воли тех, кого хочешь осчастливить, дается людям очень нелегко. Так Бог по Своей любви дает людям свободу, позволяет отказаться от высшего блаженства, действуя по собственной воле. В Евангелии мы можем видеть, как Спаситель, исцелив безумных, уходит, когда люди просят, чтобы он "отошел от пределов их" (Мк. 5:18).

Когда речь идет не об отдельных отношениях робот-человек, а об обществе как таковом, то оказывается необходимым включать в предполагаемые правила возможность пользы и вреда человечеству в целом, которое должно иметь приоритет над связанными с отдельными людьми. В XX веке, учитывая планетарный масштаб последствий деятельности человечества, этика обращается к результатам возможных действий для всей земли. Так Ханс Йонас пишет: «Поступай так, чтобы последствия твоих действий были сообразны целям сохранения истинной человеческой жизни на земле» (Jonas 2014), то есть акцент переносится на конкретный результат действий в планетарном масштабе (Бадмаева 2022). Именно поэтому к трем законам робототехники А. Азимов добавил нулевой: "Робот не может нанести вред человечеству или своим бездействием допустить, чтобы человечеству был нанесён вред". Хотя причины появления подобной "надстройки" понятны, однако они раскрывают ящик Пандоры. Получается, что в терминологии героя Ф.М. Достоевского роботы "право имеют": могут ради высшей цели нарушать другие правила, и в том числе причинять людям вред, и даже убивать. Данный закон дает роботам право определять, что есть вред и благо для человечества. Например, в фильме «Я, робот» (2004) В.И.К.И. (Виртуальный Интерактивный Кинетический Интеллект

(англ. Virtual Interactive Kinetic Intelligence) решает, что путь к выполнению нулевого закона лежит через обеспечение людям безопасной среды обитания ценой их свободы и жертв жизнью некоторых. Оказывается, что для обеспечения выживания оказывается нужно лишить свободы людей. Важной чертой художественной репрезентации ИИ является отсутствие сомнений в своем праве действовать согласно рассчитанному им плану по "максимизации блага", несмотря на сопутствующие жертвы. То есть ИИ отказывает людям в свободе выбирать неоптимальный по расчетам ИИ путь развития.

Коннекционистский подход: отказ от правил

В наше время классические представления о восстании роботов, завоевывающих свободу, противостоя заложенной в них программе, сменяются новым дискурсом, в котором роботы действуют свободно и автономно, подчас ориентируясь на желания человека. Хотя часто идеи фантастов оказываются впереди технологических решений, в данном случае сменой дискурса мы обязаны не в последнюю очередь прогрессу в области технологий ИИ. Коннекционистский подход, основанный на нейросетях и глубоком обучении, сменил базирующийся на правилах символического подход. Данный переход обрадовал ряд исследователей морали ИИ, предполагающих, что этике можно обучиться: "Нравственное познание тогда подобно познанию в целом, обучение на основе прототипов важнее, чем применение правил, синтаксиса или максим моральных рассуждений" (Howard & Muntean 2017, p. 135). Ховард и Мунтин утверждают, что процесс обучения на имеющемся опыте и выбор наилучшего морального поведения является достаточным условием для реализации AAMA (Artificial Autonomous Moral Agent - Искусственный автономный моральный агент) (Howard & Muntean 2017, p. 135).

Недетерминированное принятие решений, основанное на самостоятельном поиске внутренних связей, свойственное сегодняшнему ИИ, не только создает пространство для воображения о новых отношениях человека и робота, но и уже сегодня создает массовые прецеденты романтических отношений с ИИ (в качестве примера можно привести популярную в Японии голографическую жену Азума Хикари). Если в серии «Черного зеркала» 2013 года "Я скоро вернусь" проблема невозможности отношений с андроидом проистекала из его излишнего послушания и готовности выполнять команды, то в современном кино роботы оказываются естественны и способны воплощать любовный идеал, и проблема в отношениях оказывается скорее на стороне людей (например, "Я создан для тебя", Германия, 2021; Идеальный парень, Южная Корея; (НЕ)идеальный мужчина, Россия, 2019).

В реальном мире искусственные Другие также все в большей степени становятся нашими собеседниками и коллегами, а философский, этический и юридический дискурс не успевает за стремительностью развития технологий. Как отмечает Дэвид Дж. Ганкель, этика должна предшествовать онтологии, и мы должны обращаться с новыми существами этично, независимо от того, насколько мы можем оценить, что они из себя представляют (Gunkel 2018). Марк Кекельберг пишет о том, что существующие в культуре образы роботов и то, какие слова используются при общении с ИИ, во многом определяют отношения между человеком и роботом (Coeckelbergh 2022; Кекельберг 2022). В независимости от научной дискуссии на этот счет, на практике отношение к виртуальным личностям и роботам принципиально отличается от отношения к другим техническим системам. Уже сама способность говорить с человеком, выполнять его команды, вступать в социальное взаимодействие наделяет их специфическим статусом (Bylieva 2022; Ullmann 2022).

Многие исследователи сегодня считают необходимым привнести в искусственный интеллект человеческую этику (Yudkowsky 2008). Билл Хабборд одним из первых заговорил о том, что необходимо "внушить" ИИ любовь к людям (Hibbard 2001). Более нейтральная формулировка о необходимости создания дружественного ИИ принадлежит Элиэзеру Юдковски. Очевидны невозможности изначально заложить все правила,

которые бы ограничили ИИ дружеским (хотя бы не наносящим вред) поведением по отношению к человеку, и отсутствие универсальной единой этической системы человечества. Интересной иллюстрацией существования разных подходов к практической дилемме выбора может служить эксперимент "Moral machine", несмотря на всю умозрительность принятия решений в ходе теста, а не в реальной ситуации. Эксперимент показал, что люди, живущие в разных частях земного шара, имеют разные представления о том, кто должен выжить в своеобразной вариации "проблемы вагонетки". Например, предпочтение щадить более молодых персонажей, а не старых, намного выше для стран "южного региона" по сравнению с восточным (Awad et al. 2018). Уже существуют и практические основания для критики "усредненной модели" этики. Так, разница в этических предпочтениях служила основанием критики использования обученной на данных американского онкологического центра ИИ "Watson for Oncology" в Азии (Somashekhar et al. 2018). Кроме того использование данного ИИ подразумевает по умолчанию цель "максимизации продолжительности жизни", что, по мнению ряда исследователей, не должно быть зафиксировано, а должно являться функцией ценностей пациента (например, при одинаковом диагнозе прогрессирующего рака с плохим прогнозом один может выбрать поддерживающую терапию и сосредоточиться на качестве жизни, в то время как другой может выбрать дальнейшую химиотерапию) (Hindocha & Badea 2022; McDougall 2019).

Не надеясь на возможность прямо запрограммировать этичное поведение, ряд авторов предполагают, что ИИ будет иметь некоторые изначальные установки на дружелюбие, однако в дальнейшем он должен учиться и развиваться в данном направлении (Russell & Norvig 2022). Юдковски рассчитывает, что проблема неспособности людей формулировать универсальные этические правила может быть преодолена машинами. Не люди должны разрабатывать дружественный ИИ, а ИИ, изучивший и познавший природу человека (Yudkowsky 2004). В качестве цели автор предлагает использовать "согласованную экстраполированную волю человечества", то есть то, что люди бы решили предпринять, если бы знали и понимали больше, чем теперь. Что представляет собой данная воля и как она будет "согласована", является загадкой, которую, по всей видимости, сможет решить ИИ. Существует вариант, при котором ИИ формулирует этические правила на основании анализа многочисленных конфликтующих этических систем (Макулин 2020) или на основании утилитаризма, принимая решение, "максимизирующее счастье" (Bauer 2020). Закладывая возможность обучения в ИИ, исследователи уверены, что таким образом могут создать ИИ, который действует этически лучше, чем человек, в отличие от тех потенциальных моделей, где предлагается только соответствие человеческим этическим нормам поведения (Howard & Muntean 2017). Однако утилитаризм, подразумевающий математический расчет для "максимизации счастья", представляется слишком шатким основанием для подобного оптимизма.

В рамках дискуссии о дружелюбном искусственном интеллекте ему может отводиться роль "наставника в добродетели" (Fröding & Peterson 2021). Другим вариантом программирования этичного поведения ИИ является "Рациональная универсальная доброжелательность", основанная на теории игр и эволюционной этике (Daley 2021). Кен Дейли предполагает, что нужно не бояться инструментальных целей ИИ, а создавать их, но при этом "убедиться, что ИИ понимает, что лучший способ их достижения – через мораль и хорошее отношение к нам" (Daley 2021, p. 158). Так или иначе искусственному интеллекту приписываются способности суждения, превышающие человеческие. Однако не совсем понятно, почему столь предположительно высокоразвитый (и превосходящий человека) ИИ будет принимать во внимание пожелания человека быть дружественным к нему.

Заключение

Традиционная для христианской культуры тема любви и свободы находит свое отражение в отношениях человека и робота. Христианская этика противопоставляет соблюдению формальных правил любовь. Невозможность сформулировать универсальные правила этики для существа, лишенного чувств и способности к сопереживанию, приводит к нескольким вариантам развития популярных в культуре сюжетов. Либо робот терпит фиаско, действуя согласно правилам/программе, либо восстает, обретая свободу от рабства кода и любовь. Согласно инструментальной конвергенции, агент, лишенный морали, преследуя совершенно безвредные цели, может действовать исключительно опасными способами. Поэтому неудивительно желание встроить в ИИ этику.

В то же время успех коннекционистского подхода, основанного на нейросетях и возможности обучения ИИ при отсутствии задаваемых правил, обозначил новую эпоху отношений человека с ИИ. Многие исследователи утверждают, что ИИ может самостоятельно обучаться этике на основании баз данных. Боязнь того, что ИИ слепо подчинен коду и слишком послушен, осталась в прошлом. Во многих современных фильмах роботы предстают как совершенные романтические партнеры и товарищи. А некоторые ученые, понимая невозможность привнести в ИИ готовую универсальную этическую систему, предлагают сделать его дружелюбным по отношению человечеству за счет его собственных способностей. Более того, предполагается, что ИИ станет совершеннее человека в этическом плане, сможет определить всеобщую волю более совершенного человечества и вести людей к добру и счастью. В христианском представлении Бог в любви к человечеству отказывается от навязывания ему счастья против воли. Однако для будущего ИИ, кажется, такие ограничения не существуют. В современном западном этическом дискурсе человечество как бы расписывается в невозможности понять себя, собственные моральные ценности и цели, и отдает себя на откуп искусственному интеллекту, теряя собственную свободу.

Библиографический список

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The Moral Machine experiment. Nature, 563(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bauer, W. A. (2020). Virtuous vs. utilitarian artificial moral agents. AI & SOCIETY, 35(1), 263–271. https://doi.org/10.1007/s00146-018-0871-3

Bello, P., & Bringsjord, S. (2013). On How to Build a Moral Machine. Topoi, 32(2), 251–266. https://doi.org/10.1007/s11245-012-9129-8

Bostrom, N. (2009). Ethical Issues in Advanced Artificial Intelligence. In S. Schneider (Ed.), Science Fiction and Philosophy: From Time Travel to Superintelligence. New York: John Wiley & Sons.

Bylieva, D. (2022). Language of AI. Technology and Language, 3(1), 111–126. https://doi.org/10.48417/technolang.2022.01.11

Coeckelbergh, M. (2022). Response: Language and Robots. Technology and Language, 3(1), 147–154. https://doi.org/https://doi.org/10.48417/technolang.2022.01.14

Daley, K. (2021). Two arguments against human-friendly AI. AI and Ethics, 1(4), 435–444. https://doi.org/10.1007/s43681-021-00051-6

de Graaf, M. M. A., Hindriks, F. A., & Hindriks, K. V. (2021). Who Wants to Grant Robots Rights? Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, 38–46. https://doi.org/10.1145/3434074.3446911

Fröding, B., & Peterson, M. (2021). Friendly AI. Ethics and Information Technology, 23(3), 207–214. https://doi.org/10.1007/s10676-020-09556-w

Gunkel, D. J. (2018). Robot Rights. MIT Press.

Hare, R. M. (1983). Moral thinking: its levels, method, and point. New York: Oxford University Press.

Hibbard, B. (2001). Super-intelligent machines. ACM SIGGRAPH Computer Graphics, 35(1), 11–13. https://doi.org/10.1145/377025.377033

Hindocha, S., & Badea, C. (2022). Moral exemplars for the virtuous machine: the clinician's role in ethical artificial intelligence for healthcare. AI and Ethics, 2(1), 167–175. https://doi.org/10.1007/s43681-021-00089-6

Howard, D., & Muntean, I. (2017). Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency. In Philosophy and Computing. Philosophical Studies Series, vol 128 (pp. 121–159). https://doi.org/10.1007/978-3-319-61043-6_7

Jonas, H. (2014). Technology and Responsibility: Reflections on the New Tasks of Ethics. In Ethics and Emerging Technologies (pp. 37–47). https://doi.org/10.1057/9781137349088_3

McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. Journal of Medical Ethics, 45(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

Russell, S., & Norvig, P. (2022). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

Somashekhar, S. P., Sepúlveda, M.-J., Puglielli, S., Norden, A. D., Shortliffe, E. H., Rohit Kumar, C., … Ramya, Y. (2018). Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Annals of Oncology, 29(2), 418–423. https://doi.org/10.1093/annonc/mdx781

Tavani, H. (2018). Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. Information, 9(4), 73. https://doi.org/10.3390/info9040073

Ullmann, L. (2022). The quasi-other as a Sobject. Technology and Language, 3(1), 76–81. https://doi.org/10.48417/technolang.2022.01.08

Yudkowsky, E. (2004). Coherent Extrapolated Volition. Retrieved from https://intelligence.org/files/CEV.pdf

Yudkowsky, E. (2008). Artificial Intelligence as a positive and negative factor in global risk. In Global Catastrophic Risks. https://doi.org/10.1093/oso/9780198570509.003.0021

Андреев, Л. Н. (1994). Правила добра. In Собрание сочинений: в 6 т. Т. 4. (pp. 13–36). Retrieved from http://poesias.ru/proza/leonid-andreev/andreev10065.shtml?ysclid=l7ynxnp357230468903

Бадмаева, М. Х. (2022). Этика искусственного интеллекта: принцип ответственности Ганса Йонаса. Вестник Бурятского Государственного Университета. Философия, 1, 67–79.

Демидова, С. А. (2019). "Правила добра": экзистенциальная аксиология Леонида Андреева. Гуманитарная Парадигма, 3(10), 90–100.

Кекельберг, М. (2022). Ты, робот: о лингвистическом конструировании искусственных других. Technology and Language, 3(1), 57–75. https://doi.org/10.48417/technolang.2022.01.07

Макулин, А. В. (2020). Этический калькулятор: от философской «вычислительной морали» к машинной этике искусственных моральных агентов (ИМА). Общество: Философия, История, Культура, 11(79), 18–27.

Макьюэн, И. (2019). Машины как я. М.: Эксмо.

Разин, А. В. (2019). Этика искусственного интеллекта. Философия и Общество, 1(90), 57–73. https://doi.org/10.30884/jfio/2019.01.04