СЕРИЯ
ИССЛЕДОВАНИЯ
КУЛЬТУРЫ

Bernd Carsten Stahl
Doris Schroeder
Rowena Rodrigues

Ethics of Artificial Intelligence

Case Studies and Options for Addressing Ethical Challenges

Бернд Карстен Шталь
Дорис Шредер
Ровена Родригес

Этика искусственного интеллекта

Кейсы и варианты решения этических проблем

Перевод с английского
Инны Кушнаревой
под научной редакцией
Александра Павлова

Второе издание

Издательский дом
Высшей школы экономики
Москва, 2025

УДК 004.8+172
ББК 32.813+87.7
Ш87

https://elibrary.ru/mfedco

Проект серийных монографий
по социально-экономическим
и гуманитарным наукам

Руководитель проекта Александр Павлов

Шталь, Бернд Карстен и др.

Ш87 Этика искусственного интеллекта: Кейсы и варианты решения этических проблем / Б. К. Шталь, Д. Шредер, Р. Родригес; пер. с англ. И. Кушнаревой; под науч. ред. А. Павлова; Нац. исслед. ун-т «Высшая школа экономики». — 2-е изд. — М.: Изд. дом Высшей школы экономики, 2025. — 200 с. — (Исследования культуры). — 600 экз. — ISBN 978-5-7598-4316-0 (в пер.). — ISBN 978-5-7598-4291-0 (e-book).

Принято считать, что потенциальные выгоды от применения искусственного интеллекта (ИИ) велики: от операционных улучшений, таких как снижение числа человеческих ошибок, до использования роботов в опасных ситуациях. В то же время все понимают, что применение ИИ сопряжено со множеством этических проблем — от предвзятости в работе алгоритмов и цифрового разрыва до проблем здоровья и безопасности. В книге рассматриваются реальные кейсы этических проблем, которые ставит перед нами искусственный интеллект, и варианты их решения. Разбор кейсов — один из лучших способов изучения этических дилемм и понимания связанных с ними сложностей и точек зрения заинтересованных в работе ИИ сторон.

С учетом всеобъемлющего характера этики искусственного интеллекта в академических, политических, философских и медийных дебатах книга будет полезна широкой аудитории, включая исследователей, представляющих самые различные дисциплины, а также политиков, сотрудников неправительственных организаций, преподавателей и образованную общественность.

УДК 004.8+172
ББК 32.813+87.7

Перевод выполнен по изданию: *Stahl B.C., Schroeder D., Rodrigues R.* Ethics of Artificial Intelligence. Case Studies and Options for Addressing Ethical Challenges (Springer, 2023)

Опубликовано Издательским домом Высшей школы экономики
http://id.hse.ru

doi:10.17323/978-5-7598-4316-0

ISBN 978-5-7598-4316-0 (в пер.)
ISBN 978-5-7598-4291-0 (e-book)
ISBN 978-3-031-17040-9 (англ.)

© The author(s) 2013 / CC BY 4.0
© Перевод на русский язык. Национальный исследовательский университет «Высшая школа экономики», 2024; 2025

ОГЛАВЛЕНИЕ

Благодарности 10

Глава 1. Этика искусственного интеллекта: введение 12
Несправедливая и незаконная дискриминация (глава 2) 16
Неприкосновенность частной жизни (глава 3) 17
Надзорный капитализм (глава 4) 17
Манипуляция (глава 5) 18
Право на жизнь, свободу и личную безопасность (глава 6) 18
Человеческое достоинство (глава 7) 19
«ИИ во благо» и Цели ООН в области устойчивого развития (глава 8) 19
Библиография 21

Глава 2. Несправедливая и незаконная дискриминация 24
2.1. Введение 24
2.2. Кейсы дискриминации, которой способствовал ИИ 25
2.2.1. Кейс 1: Предубежденность в инструментах найма на работу 25
2.2.2. Кейс 2: Дискриминирующее использование ИИ в правоохранительной системе и для профилактических полицейских мер 28
2.2.3. Кейс 3: Дискриминация на основе цвета кожи 32
2.3. Этические вопросы, связанные с дискриминацией, которой способствуют ИИ . . . 33

2.4. Меры по борьбе с несправедливой/
незаконной дискриминацией 38
2.4.1. Оценка воздействия ИИ 40
2.4.2. Этика, заложенная в дизайн 42
2.5. Выводы 44
Библиография 45

Глава 3. Неприкосновенность
частной жизни. 50
3.1. Введение 50
3.2. Кейсы нарушения неприкосновенности
частной жизни по вине ИИ 53
3.2.1. Кейс 1: Использование
персональных данных
авторитарными режимами 53
3.2.2. Кейс 2: Неприкосновенность
генетических данных 55
3.2.3. Кейс 3: Биометрический надзор 59
3.3. Защита данных
и неприкосновенность частной жизни 61
3.4. Меры по борьбе с угрозами
неприкосновенности частной жизни,
порождаемыми ИИ 64
3.5. Выводы 67
Библиография 70

Глава 4. Надзорный капитализм 74
4.1. Введение 74
4.2. Кейсы надзорного капитализма,
использующего ИИ 76
4.2.1. Кейс 1: Присвоение данных 76
4.2.2. Кейс 2: Монетизация данных
о здоровье 77
4.2.3. Кейс 3: Нечестные коммерческие
практики 79
4.3. Этические вопросы надзорного капитализма . . 80
4.4. Меры противодействия
надзорному капитализму 83

4.4.1. Антимонопольное законодательство . . . 84
4.4.2. Доступ и шеринг данных 86
4.4.3. Укрепление права собственности потребителей/индивидов на их данные 87
4.5. Выводы 89
Делить компании, опираясь на антимонопольное законодательство, сложно 89
Являются ли «работники крупных технологических компаний одним из главных сдерживающих факторов роста ее власти»? . . . 91
Библиография 92

Глава 5. Манипуляция 98
5.1. Введение 98
5.2. Кейсы манипулирования при помощи ИИ . . . 99
5.2.1. Кейс 1: Манипуляции на выборах 99
5.2.2. Кейс 2: Продвижение продаж в «моменты уязвимости для внушения» 100
5.3. Этика манипуляции 101
5.4. Меры для борьбы с манипуляциями 106
5.5. Выводы 109
Библиография 110

Глава 6. Право на жизнь, свободу и личную безопасность. 113
6.2. Кейсы, в которых ИИ негативно влиял на право на жизнь, свободу и личную безопасность 116
6.2.1. Кейс 1: Смертельная авария с участием беспилотного автомобиля 116
6.2.2. Кейс 2: Уязвимости в системе управления умными домами 119
6.2.3. Кейс 3: Состязательные атаки в медицинской диагностике 121
6.3. Этические вопросы 122
6.3.1. Безопасность человека 122

6.3.2. Неприкосновенность частной жизни . . . 123
6.3.3. Ответственность и подотчетность 124
6.4. Меры по защите жизни, свободы и безопасности людей 127
6.4.1. Внедрение и укрепление режимов ответственности 127
6.4.2. Управление качеством систем ИИ 130
6.4.3. Состязательная устойчивость 132
6.5. Выводы 132
Библиография 134

Глава 7. Человеческое достоинство 140
7.1. Введение 140
7.2. Кейсы с потенциальным конфликтом между ИИ и человеческим достоинством 143
7.2.1. Кейс 1: Несправедливое увольнение . . . 143
7.2.2. Кейс 2: Секс-роботы 147
7.2.3. Кейс 3: Роботы-сиделки 151
7.3. Этические вопросы, касающиеся ИИ и достоинства 154
7.4. Выводы 161
Библиография 162

Глава 8. «Искусственный интеллект во благо» и Цели устойчивого развития. 167
8.1. Введение 167
8.2. Приносит ли «ИИ во благо» благо? Кейсы 170
8.2.1. Кейс 1: Сезонное прогнозирование климата в условиях ограниченных ресурсов 171
8.2.2. Кейс 2: «Вертолетные исследования» . . . 172
8.3. Этические вопросы, возникающие в связи с «ИИ во благо» и ЦУР 175
8.3.1. Пустыня данных или неравномерное распределение доступности данных 176
8.3.2. Двойные стандарты 177

8.3.3. Невнимание к социальным составляющим проблем, которые призваны решить ЦУР 178
8.3.4. Слон в комнате: цифровой разрыв и нехватка талантов для ИИ 179
8.3.5. Нерешенные важные проблемы, в которых ИИ и ЦУР вступают в конфликт . . . 181
8.4. Выводы 182
Библиография 183

Глава 9. Этика искусственного интеллекта: заключение 188
Библиография 196

БЛАГОДАРНОСТИ

Эта книга опирается на работу, которую авторы вели в рамках целого ряда различных проектов. Главный проект, который свел нас вместе и продемонстрировал потребность в кейсах, описывающих этические проблемы ИИ и пути их решения, — SHERPA (2018–2021), финансируемый ЕС. Все три автора были его участниками, причем Бернд Шталь возглавлял его. Мы также хотели бы отметить вклад участников консорциума SHERPA, который вдохновлял и оказывал влияние на работу над данной книгой.

Все три автора активно участвовали и продолжают работать в других проектах, прямо или косвенно оказавших влияние на эту книгу. Эти проекты и люди, которые в них заняты, заслуживают благодарности. Среди этих проектов финансируемые ЕС «Проект “Человеческий мозг”», «ТехЭтос» и SIENNA, а также проекты «Ответственная индустрия», CONSIDER, TRUST и ETICA.

Мы также благодарим за поддержку коллег из наших институтов и организаций, в частности из Центра компьютерных наук и социальной ответственности при Университете Де Монфорт, Центра профессиональной этики при Университете Центрального Ланкашира и Trilateral Research Ltd.

Мы хотели бы поблагодарить Пауля Вайса за замечательную редактуру, Джулию Кук за проницательные замечания к первому варианту работы и Джайанти Кришнамурти, Джулиану Питанги и Тони Милевой из Springer Nature за эффективное управление процессом публикации. Благодарим Костаса Иатридиса за замечательную академическую поддержку. Мы в долгу перед Амандой

Шарки за разрешение использовать ее превосходную зарисовку о роботах-сиделках и пожилых людях.

В самом конце, хотя этим список, безусловно, не исчерпывается, мы хотели бы поблагодарить трех анонимных рецензентов, чьи замечания помогли нам внести изменения в исходные идеи книги.

Это исследование получило финансирование от программы EC Horizon 2020 Frame-work Programme for Research and Innovation по грантам No. 786641 (SHERPA) и No. 945539 (Human Brain Project SGA3).

Глава 1
Этика искусственного интеллекта: введение

Этические вызовы, которые бросает искусственный интеллект (ИИ), — одна из главных тем XXI века. Принято считать, что потенциальные выгоды от применения ИИ велики: от операционных улучшений, таких как снижение числа человеческих ошибок (например, при постановке медицинских диагнозов), до использования роботов в опасных ситуациях (например, для обеспечения безопасности ядерной электростанции после аварии). В то же время ИИ ставит множество этических проблем — от предвзятости в работе алгоритмов и цифрового разрыва до проблем здоровья и безопасности.

Сфера ИИ превратилась в широкомасштабный проект, в который вовлечены самые разные заинтересованные лица. Однако в этике ИИ нет ничего нового. Концепции ИИ почти 70 лет [McCarthy et al., 2006], и озабоченность его развитием с точки зрения этики высказывалась уже в середине XX века [Wiener, 1954; Винер, 2001; Dreyfus, 1972; Дрейфус, 1978; Weizenbaum, 1977; Вейценбаум, 1982]. Сейчас эти дебаты активизировались благодаря более широкому интересу к применению и воздействию усовершенствованных алгоритмов, большей доступности вычислительных мощностей и растущему объему данных, которые могут использоваться для анализа [Hall, Pesenti, 2017].

Эти технические достижения благоприятствовали развитию определенных типов ИИ, в особенности машинного обучения [Alpaydin, 2020; Faggella, 2020], одной из популярных форм которого, в свою очередь, является глубокое обучение (см. врезку) [LeCun et al., 2015]. Успех этих подходов к ИИ привел к быстрому расширению сферы его применения, что нередко влекло за собой этически неоднозначные последствия, например, несправедливое или незаконное господство, дискриминацию и вмешательство в политику.

С расширением сферы применения ИИ его этика выходит далеко за пределы академической науки. Так,

---

Глубокое обучение

Глубокое обучение — один из подходов к машинному обучению, в последние годы приведший к значительным успехам в разработке ИИ [Bengio et al., 2021]. Развитие глубокого обучения — результат использования искусственных нейросетей, пытающихся реплицировать или симулировать функции мозга. Естественный интеллект зарождается в параллельных сетях нейронов, которые обучаются, регулируя силу своих связей. Глубокое обучение пытается воспроизвести деятельность, напоминающую деятельность мозга, используя статистические параметры, чтобы определить, хорошо ли функционирует сеть. Глубокое обучение получило свое название от глубоких нейронных сетей, то есть сетей со множеством слоев. Оно успешно применялось к целому ряду проблем, от распознавания образов до обработки естественной речи. Несмотря на свои успехи, глубокое обучение упирается в ряд ограничений [Cremer, 2021]. Идут дебаты о том, как далеко еще может продвинуться машинное обучение, основанное на таких подходах, как глубокое обучение, и не потребуются ли в будущем фундаментально иные принципы, например, внедрение моделей каузальности [Schölkopf et al., 2021].

«Римская петиция об этике ИИ»[1], выпущенная в феврале 2020 года, связывает Ватикан с Продовольственной и сельскохозяйственной организацией ООН (ФАО), Microsoft, IBM и итальянским Министерством технологических инноваций и цифровизации. Еще один пример: в июле 2021 года ЮНЕСКО собрало 24 экспертов со всего мира и запустило международные онлайн-консультации по этике ИИ, чтобы облегчить диалог между всеми странами — членами этой организации. Большой интерес также проявляет пресса, хотя некоторые ученые считают, что вопросы этики ИИ в ней рассматриваются слишком «поверхностно» [Ouchchy et al., 2020].

Одна из больших проблем, с которыми может столкнуться этика ИИ и те, кто ею занимаются, — непрозрачность того, что происходит внутри ИИ. При том что хорошее понимание самой этой деятельности очень важно для рассмотрения этических вопросов.

> В обязанности специалиста по этике ИИ не входит программирование самих систем, и едва ли от него можно ждать, что он с ним справится. Вместо этого он должен понимать, среди прочего, чем отличается обучение с учителем от обучения без учителя, что такое разметка данных, как получают согласие пользователя — то есть иметь представление о том, как проектируется, разрабатывается и используется система. Другими словами, специалист по этике ИИ должен понимать процесс настолько, насколько это нужно для того, чтобы отследить моменты, когда необходимо вмешаться и ответить на ключевые этические вопросы [Gambelin, 2021].

Таким образом, ожидается, что специалисты по этике ИИ будут знакомы с технологией, хотя, включая самих разра-

[1] What is the matter with AI Ethics? // RenAIssance Foudation [Электронный ресурс]. URL: https://www.romecall.org/ (обращение 27.02.2024).

ботчиков ИИ, «никто по-настоящему не знает, каким образом самые передовые алгоритмы делают то, что они делают» [Knight, 2017].

Несмотря на этот непрозрачный характер работы ИИ в его современной форме, важно размышлять и обсуждать то, какие этические вопросы могут возникнуть в ходе его развития и применения. Подход, который мы здесь выбрали, заключается в изучении кейсов, поскольку «реальный опыт этики ИИ предлагает [...] примеры со множеством нюансов» [Brusseau, 2021] для обсуждения, изучения и анализа. Данный подход даст нам возможность проиллюстрировать основные этические вызовы, связанные с ИИ, часто с отсылкой к правам человека [Franks, 2017].

Анализ кейсов — хорошо зарекомендовавший себя метод углубления понимания теоретических концепций на примере ситуаций из реального мира [Escartín et al., 2015]. Он также позволяет привлечь студентов и расширить опыт обучения [Ibid.], а потому хорошо подходит для преподавания [Yin, 2003].

Поэтому мы выбрали для этой книги метод анализа кейсов. Мы отобрали наиболее значимые или релевантные этические вопросы, которые в настоящий момент обсуждаются в контексте ИИ (основываясь на Андреу [Andreou et al., 2019] и других источниках с учетом обновлений), и посвятили по отдельной главе каждому из них.

Главы имеют следующую структуру. Сначала мы приводим небольшие примеры из реальной жизни, чтобы дать общее представление о конкретном этическом вопросе. Затем представляем нарративную оценку этой виньетки и широкий контекст. В конце мы предлагаем пути решения этических вопросов, которые она затрагивает. Часто это делается в форме обзора инструментов, позволяющих бороться с той или иной этической опасностью. Например, кейс предвзятости в алгоритме, ведущей к дис-

криминации, будет сопровождаться объяснением цели и объема оценки воздействия ИИ. Там, где подходящие инструменты отсутствуют, так как люди должны принять решение, основываясь на этических размышлениях (например, в случае с секс-роботами), мы даем резюме различных стратегий аргументации. В центре нашего внимания случаи из *реальной жизни*, большинство из которых нашли отражение в прессе или в научных журналах. Ниже приводится краткий обзор этих кейсов.

Несправедливая и незаконная дискриминация (глава 2)

В первом примере речь пойдет об автоматизированном составлении короткого списка кандидатов на вакансию при помощи ИИ, обучавшегося на резюме соискателей за последние десять лет. Несмотря на попытки решить проблему гендерного перекоса на самом первом этапе, компания в итоге отказалась от этого метода, так как он был несовместим с ее приверженностью разнообразию и равенству на рабочем месте.

Во втором примере описывается, как заключенному, хорошо проявившему себя в программе реабилитации, было отказано в условно-досрочном освобождении из-за того, что ИИ предсказал, будто он представляет опасность для общества. Выяснилось, что субъективное личное мнение тюремных надзирателей, возможно, имеющих расовые предрассудки, привело к необоснованно завышенной оценке опасности этого заключенного для общества.

Третий пример рассказывает об истории студента-инженера азиатского происхождения, чья фотография на паспорте не была принята государственными системами Новой Зеландии на том основании, что на ней у него якобы закрыты глаза. Это была ошибка в распознавании фотографии на паспорте, связанная с этническим происхожде-

нием, которую подобные системы совершали и в других местах, например, в Великобритании в случае с темнокожими женщинами.

Неприкосновенность частной жизни (глава 3)

Первый пример связан с китайской системой социального кредита, которая использует самые разные данные для подсчета рейтинга благонадежности граждан. Высокий рейтинг позволяет получить льготы, а низкий — приводит к отказу в оказании услуг.

Второй пример рассказывает о запущенной в Саудовской Аравии «Программе исследования генома человека», которая, по прогнозам, должна привести к прорывам в медицине, но при этом вызывает обеспокоенность возможными нарушениями неприкосновенности частной жизни.

Надзорный капитализм (глава 4)

Первый пример посвящен сбору фотографий, которые извлекаются из таких сервисов, как Instagram*[2], LinkedIn и YouTube, без ведома пользователей и в нарушение соглашения с ними. По сообщениям, одна компания при помощи ИИ, специализирующегося на распознавании лиц, собрала 10 млрд изображений лиц людей со всего мира.

Второй пример содержит факты об утечке данных у поставщика услуг по отслеживанию состояния здоровья, из-за чего в общий доступ попали данные 61 млн человек.

---

[2] Здесь и далее звездочкой (*) маркированы корпорация Meta и принадлежащие ей социальные сети Instagram и Facebook, деятельность которых признана экстремистской и запрещена на территории РФ; информация используется в исследовательских целях и не направлена на одобрение экстремистской деятельности. — *Примеч. ред.*

В третьем примере кратко излагается судебное дело, возбужденное против Facebook* за то, что компания ввела пользователей в заблуждение, своевременно и должным образом не объяснив им при активации учетной записи, что их данные будут использоваться в коммерческих целях.

Манипуляция (глава 5)

В первом примере рассматривается скандал с Cambridge Analytica и Facebook*, который позволил этой компании собрать 50 млн профилей пользователей, отправлять персонализированные сообщения владельцам учетных записей и вести широкий анализ поведения избирателей в преддверье американских президентских выборов 2016 года и референдума о брексите в том же году.

Второй пример показывает, как научные исследования используются для того, чтобы навязывать коммерческие продукты потенциальным покупателям в моменты, когда те с наибольшей легкостью поддаются внушению. Например, косметическая продукция предлагается в те моменты, когда адресаты рекламы чувствуют себя непривлекательными.

Право на жизнь, свободу и личную безопасность (глава 6)

Первый пример посвящен нашумевшей аварии с беспилотным автомобилем Tesla, в которой погиб человек, находившийся в машине.

Во втором примере приводится обзор уязвимостей в безопасности систем умного дома, которые могут привести к атакам по принципу «человек посередине», то есть такому виду кибератак, в котором нарушение безопасности системы позволяет хакеру перехватывать конфиденциальную информацию.

Третий пример посвящен состязательным атакам при постановке медицинских диагнозов, когда система ИИ может быть почти на 70% введена в заблуждение поддельными изображениями.

Человеческое достоинство (глава 7)

Первый пример описывает кейс работника, чье человеческое достоинство было унижено, когда его незаконно уволили и грубо выдворили из офиса компании. Решение об увольнении было принято, основываясь на непрозрачной рекомендации, данной автоматической системой.

Второй пример посвящен секс-роботам и, в частности, вопросу о том, оскорбляют ли они достоинство женщин и девочек.

В том же ключе в третьем примере рассматривается вопрос о том, являются ли роботы-сиделки оскорблением достоинства пожилых людей.

«ИИ во благо» и цели ООН в области устойчивого развития (глава 8)

Первый пример рассказывает о том, как сезонное предсказание климата в условиях ограниченных ресурсов привело к отказам в кредитовании неимущих фермеров в Зимбабве и Бразилии и к досрочному увольнению работников рыболовецкой промышленности в Перу.

Второй пример посвящен исследовательской команде из богатой страны, которой потребовались большие объемы данных мобильных телефонов пользователей из Сьерра-Леоне, Гвинеи и Либерии, чтобы отслеживать передвижения населения во время эпидемии Эболы. Комментаторы утверждают, что вместо того, чтобы тратить время на переговоры по этому вопросу, государственные структуры, страдающие от нехватки кадров, должны были зани-

маться разрешением нарастающего кризиса с эпидемией Эболы.

Это книга об анализе кейсов, связанных с этикой ИИ, а не философская работа по этике. Тем не менее мы должны ясно указать, что мы понимаем под термином «этика». Здесь мы опираемся на сложившуюся традицию этических дискуссий и ключевых позиций, таких, в частности, как оценка долга этического агента [Kant, 1788; 1797; Кант, 1994], оценка последствий действий [Bentham, 1789; Бентам, 1998; Mill, 1861; Милль, 2013], оценка характера агента [Aristotle, 2000; Аристотель, 1983] и выявление потенциальной предвзятости в собственной позиции, например, с использованием этики заботы [Held, 2005]. В нескольких главах мы отдаем предпочтение позиции Канта, но признаем и используем также другие подходы. Мы признаем, что существует множество других этических традиций, помимо упомянутых здесь доминирующих европейских, и приветствуем дебаты о том, как они могут помочь нам лучше понять и другие аспекты этики и технологии. Таким образом, мы используем термин «этика» в плюралистическом смысле.

Этот подход носит плюралистический характер, так как он открыт для интерпретаций с точки зрения главных этических теорий, а также других теоретических позиций, включая недавние попытки разработать этические теории, в большей степени нацеленные на новые технологии, такие как «раскрывающая этика» (*disclosive ethics*) [Brey, 2000], компьютерная этика [Bynum, 2001], информационная этика [Floridi, 1999] и этика процветания человека [Stahl, 2021].

Наша плюралистическая интерпретация этики ИИ согласуется с большей частью литературы по этой теме. Преобладающий подход к этике ИИ — разработка руководящих принципов [Jobin et al., 2019], основывающаяся по большей части на этических принципах среднего уровня,

как правило, выводимых из принципов биомедицинской этики [Childress, Beauchamp, 1979]. Работа Группы экспертов высокого уровня по ИИ при ЕС [AI HLEG, 2019] имела большое значение, так как она оказала сильное влияние на дискуссии в Европе, где мы находимся физически и откуда получаем финансирование для нашей работы. Однако подход к этике ИИ, основанный на руководящих этических принципах, был подвергнут серьезной критике [Mittelstadt, 2019; Rességuier, Rodrigues, 2020]. Ее главный тезис состоит в том, что этот подход далек от практики применения ИИ и не объясняет, как этика может внедряться в практику. Наш подход, основанный на анализе кейсов, нацелен на то, чтобы преодолеть эту критику, усилить этическую рефлексию и продемонстрировать возможные практические меры.

Мы приглашаем критически настроенного читателя присоединиться к нашему путешествию по кейсам этики ИИ. Мы также просим его не ограничиваться в своих размышлениях представленными в книге кейсами и задавать фундаментальные вопросы, например, о том, типичны ли обсуждаемые здесь проблемы или они касаются исключительно ИИ, и можно ли ждать их разрешения.

ИИ — пример современной и динамически развивающейся технологии. Поэтому важный вопрос состоит в том, можем ли мы продолжать размышлять об этике ИИ и научиться чему-то, что можно применить к будущим поколениям технологий, чтобы обеспечить человечеству выгоды от технологического прогресса и развития и найти способы обойти их недостатки.

Библиография

Аристотель (1983). Никомахова этика // Аристотель. Собр. соч.: в 4 т. Т. 4. М.: Мысль.

Бентам И. (1998). Введение в основания нравственности и законодательства. М.: РОССПЭН.

Вейценбаум Дж. (1982). Возможности вычислительных машин и человеческий разум. От суждений к вычислениям. М.: Радио и связь.
Винер Н. (2001). Человеческое использование человеческих существ // Винер Н. Человек управляющий. СПб.: Питер.
Дрейфус Х. (1978). Чего не могут вычислительные машины. М.: Прогресс.
Кант И. (1994). Собр. соч.: в 8 т. Т. 4. М.: Чоро.
*Милль Дж.*С. (2013). Утилитаризм. Ростов-н/Д: Донской издательский дом.

AI HLEG (2019). Ethics guidelines for trustworthy AI. High-level expert group on artificial intel-ligence. European Commission, Brussels. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (обращение 25.09.2020).
Alpaydin E. (2020). Introduction to machine learning. Cambridge: The MIT Press.
Andreou A., Laulhe Shaelou S., Schroeder D. (2019). D1.5 Current human rights frameworks. De Montfort University. Online resource. URL: https://doi.org/10.21253/DMU.8181827.v3.
Aristotle (2000). Nicomachean ethics / transl. by R. Crisp. Cambridge: Cambridge University Press.
Bengio Y., Lecun Y., Hinton G. (2021). Deep learning for AI // Communications of the ACM. Vol. 64. P. 58–65. URL: https://doi. org/10.1145/3448250.
Bentham J. (1789). An introduction to the principles of morals and legislation. Mineola: Dover Publications.
Brey P. (2000). Disclosive computer ethics // ACM SIGCAS Computers & Society. Vol. 30 (4). P. 10–16. URL: https://doi.org/10.1145/572260.572264.
Brusseau J. (2021). Using edge cases to disentangle fairness and solidarity in AI ethics. AI Ethics. URL: https://doi.org/10.1007/s43681-021-00090-z.
Bynum T.W. (2001). Computer ethics: its birth and its future. Ethics and Information Technology. Vol. 3. P. 109–112. URL: https:// doi.org/10.1023/A:1011893925319.
Childress J.F., Beauchamp T.L. (1979). Principles of biomedical ethics. New York: Oxford University Press.
Cremer C.Z. (2021). Deep limitations? Examining expert disagreement over deep learning // Progress in Artificial Intelligence. Vol. 10. P. 449–464. URL: https://doi.org/10.1007/s13748-021-00239-1.
Dreyfus H.L. (1972). What computers can't do: a critique of artificial reason. New York: Harper & Row.
Escartín J., Saldaña O., Martín-Peña J. et al. (2015). The impact of writing case studies: benefits for students' success and well-being. Procedia — Social and Behavioral Sciences. Vol. 196. P. 47–51. URL: https://doi.org/10.1016/j.sbspro.2015.07.009.
Faggella D. (2020). Everyday examples of artificial intelligence and machine learning. Boston: Emerj. URL: https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/ (обращение 23.09.2020).
Floridi L. (1999). Information ethics: on the philosophical foundation of computer ethics // Ethics and Information Technology. Vol. 1. P. 33–52. URL: https://doi.org/10.1023/A:1010018611096.

Franks B. (2017). The dilemma of unexplainable artificial intelligence // Datafloq. 25 July. URL: https://datafloq.com/read/dilemma-unexplainable-artificial-intelligence/ (обращение 18.05.2022).

Gambelin O. (2021). Brave: what it means to be an AI ethicist // AI Ethics. Vol. 1. P. 87–91. URL: https://doi.org/10.1007/s43681-020-00020-5.

Hall W., Pesenti J. (2017). Growing the artificial intelligence industry in the UK. Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy, London.

Held V. (2005). The ethics of care: personal, political, and global. New York: Oxford University Press.

Jobin A., Ienca M., Vayena E. (2019). The global landscape of AI ethics guidelines // Nature Machine Intelligence. Vol. 1. P. 389–399. URL: https://doi.org/10.1038/s42256-019-0088-2.

Kant I. (1788). Kritik der praktischen Vernunft. Ditzingen: Reclam.

Kant I. (1797). Grundlegung zur Metaphysik der Sitten. Ditzingen: Reclam.

Knight W. (2017). The dark secret at the heart of AI // MIT Technology Review. 11 Apr. URL: https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/ (обращение 18.05.2022).

LeCun Y., Bengio Y., Hinton G. (2015). Deep learning // Nature. Vol. 521. P. 436–444. URL: https://doi.org/10.1038/nature14539.

McCarthy J., Minsky M.L., Rochester N., Shannon C.E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence // AI Mag. Vol. 27. P. 12–14. URL: https://doi.org/10.1609/aimag.v27i4.1904.

Mill J.S. (1861). Utilitarianism. 2nd revised ed. Indianapolis: Hackett Publishing Co.

Mittelstadt B. (2019). Principles alone cannot guarantee ethical AI // Nature Machine Intelligence. Vol. 1. P. 501–507. URL: https://doi.org/10.1038/s42256-019-0114-4.

Ouchchy L., Coin A., Dubljevic V. (2020). AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media // AI & SOC. URL: https://doi.org/10.1007/s00146-020-00965-5.

Rességuier A., Rodrigues R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics // Big Data & Society. 7:2053951720942541. URL: https://doi.org/10.1177/2053951720942541.

Schölkopf B., Locatello F., Bauer S. et al. (2021). Toward causal representation learning // Proceedings of the IEEE. Vol. 109 (5). P. 612–634. URL: https://doi.org/10.1109/JPROC.2021.3058954.

Stahl B.C. (2021). Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies. Cham: Springer Nature Switzerland AG. URL: https://doi.org/10.1007/978-3-030-69978-9.

UNESCO (2021). AI ethics: another step closer to the adoption of UNESCO's recommendation // UNESCO, Paris. Press release. 2 July. URL: https://en.unesco.org/news/ai-ethics-another-step-closer-adoption-unescos-recommendation-0 (обращение 18.05.2022).

Weizenbaum J. (1977). Computer power and human reason: from judgement to calculation. New ed. New York: W.H. Freeman & Co Ltd.

Wiener N. (1954). The human use of human beings. New York: Doubleday.

Yin R.K. (2003). Applications of case study research. 2nd ed. Thousand Oaks: Sage Publications.