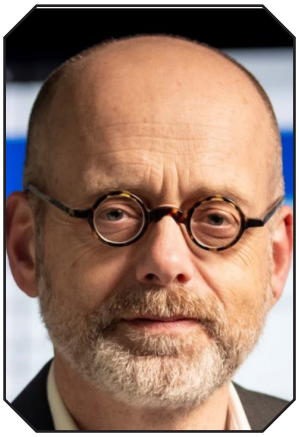
НОВЫЕ ПЕРЕВОДЫ

Б. К. Шталь, Д. Шредер, Р. Родригес

Этика искусственного интеллекта: кейсы и варианты решения этических проблем



ШТАЛЬ Бернд Карстен — профессор критических исследований технологий, Школа компьютерных наук, Ноттингемский университет. Адрес: Великобритания, NG7 2RD, г. Ноттингем, Юниверсити-Парк.

Email: Bernd.Stahl@nottingham.ac.uk

Публикуется с разрешения Издательства Института Гайдара.

Источник: Шталь Б. К., Шредер Д., Родригес Р. (готовится к изданию). *Этика искусственного интеллекта: кейсы и варианты решения этических проблем*. М.: Издательства Института Гайдара. Перев. с англ. Stahl B. C., Schroeder D., Rodrigues R. (2023) *Ethics of Artificial Intelligence. Case Studies for Addressing Ethical Challenges*, Cham, Switzerland: Springer.

Книга «Этика искусственного интеллекта: кейсы и варианты решения этических проблем» опирается на работу, которой занимались авторы в рамках ряда различных исследовательских инициатив. Главным проектом, который свёл авторов вместе и продемонстрировал потребность в кейсах, описывающих этические проблемы ИИ и пути их решения, был финансируемый ЕС проект Shaping the Ethical Dimensions of Smart Information System (SHERPA; 2018–2021). В данной книге представлены примеры этических проблем из реальной жизни в сочетании с комментариями и стратегиями их преодоления. Книга опирается на метод кейс-стади. Тематические исследования — один из лучших способов узнать об этических дилеммах и получить представление о различных сложностях и перспективах заинтересованных сторон.

Журнал «Экономическая социология» публикует первую главу из книги «Этика искусственного интеллекта: введение», где авторы описывают постановку проблемы и объясняют, каким образом структурирована их книга.

Ключевые слова: искусственный интеллект; этика; этические проблемы; анализ кейсов; машинное обучение; метод виньеток.

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 использования ИИ велики и имеют широкую область применения — от операционных улучшений, таких как снижение числа человеческих ошибок (например, при постановке медицинских диагнозов) до роботизации опасных ситуаций (например, обеспечение 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 этическая озабоченность его развитием высказывалась уже с середины XX века [Wiener 1954; Dreyfus 1972; Weizenbaum 1977]. Сейчас эти дебаты активизировались благодаря более широкому интересу к применению и воздействию усовершенствованных алгоритмов, большей доступности вычислительных мощ-

ШРЕДЕР Дорис — профессор философии морали, директор Центра профессиональной этики, Школа здравоохранения, социальной работы и спорта, Университет Центрального Ланкашира. Адрес: Великобритания, PR1 2HE, Ланкашир, г. Престон

Email: dschroeder@uclan.ac.uk

РОДРИГЕС Ровена — руководитель отдела инноваций и исследований, Trilateral Research Ltd. Адрес: Великобритания, SW1X 7QA, г. Лондон, ул. Найтсбридж Грин, д. 1.

Email: rowena.rodrigues@trilateralresearch.com

ностей и растущему объёму данных, которые могут использоваться для анализа [Hall, Pesenti 2017].

Технические достижения благоприятствовали развитию определённых типов ИИ, в особенности машинного обучения [Alpaydin 2020; Faggella 2020], одной из популярных форм которого, в свою очередь, является глубокое обучение (см. экспликацию) [LeCun, Bengio, Hinton 2015]. Успех этих подходов к ИИ привёл к быстрому расширению сферы его применения, что нередко влекло за собой этически неоднозначные последствия (например, несправедливое или незаконное господство, дискриминация и вмешательство в политику).

Глубокое обучение

Глубокое обучение — один из подходов к машинному обучению, в последние годы приведший к значительным успехам в разработке ИИ [Bengio, Lecun, Hinton 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 своё название от глубоких нейронных сетей, то есть сетей со множеством слоёв. 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 ещё 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].

С расширением сферы применения ИИ его этика выходит далеко за пределы академической науки. Так, в феврале 2020 г. в Риме был выпущен документ, призывающий к этичному развитию ИИ (см.: «Call for an AI Ethics»; https://www.romecall.org/), который связывает Ватикан со структурами ООН, занимающимися вопросами продовольствия и сельского хозяйства, с Microsoft, IBM и итальянским Министерством технологических инноваций и цифровизации. Ещё один пример: в июле 2021 г. ЮНЕСКО собрала 24 экспертов со всего мира и запустила международные онлайн-консультации по этике ИИ, чтобы облегчить диалог между всеми странами — членами этой организации. Большой интерес также проявляет пресса, хотя некоторые учёные считают, что вопросы этики ИИ в ней рассматриваются слишком поверхностно [Ouchchy, Coin, Dubljevic 2020].

Одна из больших проблем, с которыми могут 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 знакомы с технологией, хотя никто, включая самих разработчиков ИИ, 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 о том, какие этические вопросы могут возникнуть в ходе его развития и применения. И важно обсуждать эту тему. Подход, который мы здесь выбрали, заключается в изучении кейсов, поскольку реальный опыт этики ИИ предлагает примеры со множеством нюансов для рассмотрения, изучения и анализа [Brusseau 2021]. 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, а потому хорошо подходит для преподавания [Yin 2003].

По этой причине мы выбрали метод анализа кейсов. Основываясь на ряде источников с учётом их обновлений (в первую очередь — на работе: [Andreou, Laulhe Shaelou, Schroeder 2019]), мы отобрали наиболее значимые или релевантные этические вопросы, которые в настоящее время обсуждаются в контексте ИИ, и посвятили по отдельной главе каждому из них.

Главы имеют схожую структуру. Сначала мы приводим небольшие экспликации, или виньетки, моделирующие ситуации из реальной жизни и дающие 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 дискриминации, будет сопровождаться объяснением цели и объёма 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 даём 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 (Unfair and Illegal Discrimination, глава 2)

В первой виньетке-экспликации речь пойдёт об автоматизированном составлении короткого списка кандидатов на вакансию при помощи ИИ, обучавшегося на резюме соискателей за последние 10 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 её приверженностью разнообразию и равенству на рабочем месте.

Во второй виньетке описывается, как заключённый, хорошо проявивший себя в программе реабилитации, лишился возможности условно-досрочного освобождения, поскольку ИИ предсказал, что он

представляет опасность для общества. Выяснилось, что субъективное личное мнение тюремных охранников, возможно, имеющих расовые предрассудки, привело к необоснованно завышенной оценке опасности этого заключённого для общества.

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

Неприкосновенность частной жизни (Privacy, глава 3)

Первая виньетка посвящена китайской системе социального кредита, которая использует самые разные данные для подсчёта рейтинга благонадёжности граждан. Высокий рейтинг позволяет получить льготы, а низкий — привести к отказу в оказании услуг.

Вторая виньетка рассказывает о программе исследования генома человека, запущенной в Саудовской Аравии, которая, по прогнозам, должна привести к прорывам в медицине, но при этом вызывает обеспокоенность возможными нарушениями неприкосновенности частной жизни.

Надзорный капитализм (Surveillance Capitalism, глава 4)

Первая виньетка посвящена сбору фотографий, которые извлекаются из таких сервисов, как Instagram, 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 миллиардов изображений лиц людей со всего мира.

Вторая виньетка рассказывает об утечке данных у поставщика услуг по отслеживанию состояния здоровья, из-за чего в общий доступ попали данные 61 миллиона человек.

В третьей виньетке кратко излагается судебное дело, возбуждённое против Facebook за то, что компания ввела пользователей в заблуждение, своевременно и должным образом не объяснив им при активации учётной записи, что их данные будут собираться в коммерческих целях.

Манипуляция (Manipulation, глава 5)

В первой виньетке рассматривается скандал с компаниями Cambridge Analytica и Facebook, разразившийся из-за того, что были собраны 50 миллионов профилей пользователей, отправлены персонализированные сообщения владельцам учётных записей и вёлся широкий анализ поведения избирателей в преддверии американских президентских выборов 2016 г. и референдума о Брексите в том же году.

Вторая виньетка показывает, как научные исследования используются для того, чтобы навязывать коммерческие продукты потенциальным покупателям тогда, когда те с наибольшей лёгкостью поддаются внушению. Например, косметическая продукция предлагается в том случае, если адресаты рекламы чувствуют себя непривлекательными.

Право на жизнь, свободу и безопасность (Right to Life, Liberty and Security of Person, глава 6)

Первая виньетка посвящена нашумевшей аварии с беспилотным автомобилем Tesla, в которой погиб человек, находившийся в машине.

Во второй виньетке даётся 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.

В третьей виньетке речь идёт о состязательных атаках при постановке медицинских диагнозов, когда система ИИ может быть почти на 70% введена в заблуждение поддельными изображениями.

Достоинство (Dignity, глава 7)

Первая виньетка описывает кейс работника, чьё 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.

Вторая виньетка посвящена секс-роботам и, в частности, вопросу о том, оскорбляют ли они достоинство женщин и девочек.

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

ИИ для добра и цели ООН в области устойчивого развития (AI for Good and the UN's Sustainable Development Goals, глава 8)

Первая виньетка этой главы рассказывает о том, как сезонное предсказание климата в условиях ограниченных ресурсов привело к отказам в кредитовании неимущим фермерам в Зимбабве и Бразилии и к досрочному увольнению работников рыболовной промышленности в Перу.

Вторая виньетка посвящена исследовательской команде из богатой страны, которой потребовались большие объёмы 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 заниматься разрешением нарастающего кризиса с эпидемией Эболы.

Это книга об анализе кейсов, связанных с этикой ИИ, а не философская книга по этике. Тем не менее следует ясно указать на то, что мы понимаем под термином «этика». Мы опираемся на сложившуюся традицию этических дискуссий и ключевых позиций, таких, в частности, как оценка долга этического агента [Kant 1788; 1797], оценка последствий действий [Bentham 1789; Mill 1861], оценка характера агента [Aristotle 2000] и выявление потенциальной предвзятости в собственной позиции, например, с использованием этики заботы [Held 2005]. В нескольких главах мы отдаём предпочтение позиции Канта, но признаём и используем и другие взгляды. Мы отдаём себе отчёт в том, что существует множество этических традиций, помимо упомянутых здесь доминирующих европейских, и приветствуем дебаты о том, как они способны помочь лучше понять разные аспекты этики и технологии. Таким образом, мы используем термин «этика» как плюралистический.

Такой подход открыт для интерпретаций с точки зрения главных этических теорий, а также разных 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, Ienca, Vayena 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 имела большое значение, так как 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 серьёзной критике [Mittelstadt 2019; Rességuier, Rodrigues 2020]. Главный тезис оппонентов состоит в том, что такой подход далёк 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 нам в путешествии 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 важен вопрос о 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, найти способы обойти их недостатки.

Литература

AI HLEG. 2019. *Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intel-Ligence*. Brussels: European Commission. URL: https://ec.europa.eu/newsroom/dae/document.cfm? doc_id=60419

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* (trans. R. Crisp). Cambridge: Cambridge University Press. См. также рус. перев.: Аристотель. 1983. Никомахова этика. В сб.: *Аристотель. Собрание сочинений: в 4 т*. Т. 4. М.: Мысль; 53–293.

Bengio Y., Lecun Y., Hinton G. 2021. Deep Learning for AI. Communications of the ACM. 64 (7): 58–65. URL: https://doi. org/10.1145/3448250

Bentham J. 1789. *An Introduction to the Principles of Morals and Legislation*. Mineola: Dover Publications. См. также рус. перев.: Бентам И. 1998. *Введение в основания нравственности и законодательства*. М.: РОССПЭН.

Brey P. 2000. Disclosive Computer Ethics. *ACM SIGCAS Computers and Society*. 30 (4): 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*. 2: 441–447. 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.* 3: 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*. 10: 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. См. также рус. перев.: Дрейфус Х. 1978. *Чего не могут вычислительные машины*. М.: Прогресс.

Escartín J. et al. 2015. The Impact of Writing Case Studies: Benefits for Students' Success and Well-Being. *Procedia. Social and Behavioral Sciences*. 196: 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/

Floridi L. 1999. Information Ethics: On The Philosophical Foundation of Computer Ethics. *Ethics and Information Technology.* 1: 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/

Gambelin O. 2021. Brave: What It Means To Be an AI Ethicist. *AI Ethics.* 1: 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 References 7. URL:https://assets.publishing.service.gov.uk/media/5a824465e5274a2e87dc2079/Growing_the_artificial_intelligence_industry_in_the_UK.pdf

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. *Nat Mach Intell.* 1: 389–399. URL: https://doi.org/10.1038/s42256-019-0088-2

Kant I. 1788. *Kritik der praktischen Vernunft*. Ditzingen: Reclam. См. также рус. перев.: Кант И. 1994. Критика практического разума. В изд.: Кант И. *Собрание сочинений: в 8 т*. Т. 4. М.: Чоро; 373–565.

Kant I. 1797. *Grundlegung zur Metaphysik der Sitten*. Ditzingen: Reclam. См. также рус. перев.: Кант И. 1994. Основоположения метафизики нравов. В изд.: Кант И. *Собрание сочинений: в 8 т*. Т. 4. М.: Чоро; 153–246.

Knight W. 2017. The Dark Secret at the Heart of AI. *MIT Technology Review*, 11 April. URL: https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/

LeCun Y., Bengio Y., Hinton G. 2015. Deep Learning. *Nature.* 521: 436–444. URL:https://doi.org/10.1038/nature14539

McCarthy J. et al. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine.* 27 (4): 12–14. URL: https://doi.org/10.1609/aimag.v27i4.1904

Mill J. S. 1861. *Utilitarianism*. 2nd revised edn. Indianapolis: Hackett Publishing Co. См. также рус. перев.: Милль Дж. С. 2013. *Утилитаризм*. Ростов-на-Дону: Донской издательский дом.

Mittelstadt B. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence.* 1: 501–507. URL: https://doi.org/10.1038/s42256-019-0114-4

Ouchchy L., Coin A., Dubljevi´c V. 2020. AI in the Headlines: The Portrayal of the Ethical Issues of Artificial Intelligence in the Media. *AI & Society*. 35: 927–936. 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 (2): 2053951720942541. URL: https://doi.org/10.1177/2053951720942541

Schölkopf B. et al. 2021. Toward Causal Representation Learning. Proceedings of the *IEEE.* 109 (5): 612–634. doi: 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. 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.* Press Release, 2 July. Paris: UNESCO. URL: https://en.unesco.org/news/ai-ethics-another-step-closer-adoption-unescos-recommendation-0

Weizenbaum J. 1977. *Computer Power and Human Reason: From Judgement to Calculation*. New York: W. H. Freeman & Co Ltd. См. также рус. перев.: Вейценбаум Дж. 1982. *Возможности вычислительных машин и человеческий разум. От суждений к вычислениям*. М.: Радио и связь.

Wiener N. 1954. *The Human Use of Human Beings*. New York: Doubleday. См. также рус. перев.: Винер Н. 2001. Человеческое использование человеческих существ. В кн.: Винер Н. *Человек управляющий*. СПб.: Питер; 3–196.

Yin R. K. 2003. *Applications of Case Study Research*. 2nd edn. Thousand Oaks: Sage Publications.

NEW TRANSLATIONS

Bernd Carsten Stahl, Doris Schroeder, Rowena Rodrigues

Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges (excerpt)

STAHL, Bernd Carsten — Professor of Critical Research in Technology at the School of Computer Science of the University of Nottingham. Address: University Park, Nottingham. NG7 2RD, UK.

Email: Bernd.Stahl@nottingham.ac.uk

SCHROEDER, Doris — Professor of Moral Philosophy, Director of the Centre for Professional Ethics School of Health, Social Work and Sport of the University of Central Lancashire. Address: Preston, Lancashire, PR1 2 HE, UK.

Email: dschroeder@uclan.ac.uk

RODRIGUES, Rowena — Head of Innovation and Research, Trilateral Research. Address: 1 Knightsbridge Grn, London SW1X 7QA.

Email: rowena.rodrigues@trilateralresearch.com

Abstract

Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges is based on the work that Prof. B. C. Stahl, Prof. D. Schroeder and R. Rodrigues were engaged in within the framework in a number of different projects. The main project that brought the authors together and demonstrated the need for case studies addressing ethical AI problems and ways to solve them was the EU-funded *Shaping the Ethical Dimensions of Smart Information System* (SHERPA; 2018–2021). This book provides real-life examples of ethical issues, along with discussions and coping strategies. The book is based on the case study method. Case studies are one of the best ways to learn about ethical dilemmas and gain insights into the various complexities and perspectives of stakeholders.

The *Journal of Economic Sociology* publishes the first chapter of the book "Ethics of Artificial Intelligence: An Introduction", where the authors present their problem statement and outline the structure of their book.

Keywords: artificial intelligence; ethics of technology; computer ethics; information ethics; case studies; machine learning; responsible research and innovation.

References

AI HLEG. (2019) *Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intel-Ligence*, Brussels: European Commission. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (accessed 25 September 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. Available at: https://doi.org/10.21253/DMU.8181827.v3 (accessed 25 September 2020).

Aristotle. (2000) *Nicomachean Ethics* (trans. R. Crisp), Cambridge: Cambridge University Press.

Bengio Y., Lecun Y., Hinton G. (2021) Deep Learning for AI. *Communications of the ACM,* vol. 64, no 7, pp. 58–65. Available at: https://doi.org/10.1145/3448250 (accessed 18 January 2024).

Bentham J. (1789) *An Introduction to the Principles of Morals and Legislation*, Mineola: Dover Publications.

Brey P. (2000) Disclosive Computer Ethics. *ACM SIGCAS Computers and Society*, vol. 30, no 4, pp. 10–16. Available at: https://doi.org/10.1145/572260.572264 (accessed 18 January 2024).

Brusseau J. (2021) Using Edge Cases to Disentangle Fairness and Solidarity in AI Ethics. *AI Ethics*, no 2, pp. 441–447. Available at: https://doi.org/10.1007/s43681-021-00090-z (accessed 18 January 2024).

Bynum T. W. (2001) Computer Ethics: Its Birth and Its Future. *Ethics and Information Technology*, vol. 3, pp. 109–112. Available at: https://doi.org/10.1023/A:1011893925319 (accessed 18 January 2024).

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, pp. 449–464. Available at: https://doi.org/10.1007/s13748-021-00239-1 (accessed 18 January 2024).

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., Varela-Rey A., Jiménez Y., Vidal T., Rodríguez-Carballeira Á. (2015) The Impact of Writing Case Studies: Benefits for Students' Success and Well-Being. *Procedia. Social and Behavioral Sciences*, vol. 196, pp. 47–51. Available at: https://doi.org/10.1016/j.sbspro.2015.07.009 (accessed 18 January 2024).

Faggella D. (2020) *Everyday Examples of Artificial Intelligence and Machine Learning*, Boston: Emerj. Available at: https://emerj.com/ai-sector-overviews/everyday-examples-of-ai/ (accessed 23 September 2020).

Floridi L. (1999) Information Ethics: On The Philosophical Foundation of Computer Ethics. *Ethics and Information Technology*, vol. 1, pp. 33–52. Available at: https://doi.org/10.1023/A:1010018611096 (accessed 18 January 2024).

Franks B. (2017) The Dilemma of unexplainable artificial intelligence. *Datafloq*, 25 July. Available at: https://dat afloq.com/read/dilemma-unexplainable-artificial-intelligence/ (accessed 18 May 2022).

Gambelin O (2021) Brave: What It Means To Be an AI Ethicist. *AI Ethics*, vol. 1, pp. 87–91. Available at: https://doi.org/10.1007/s43681-020-00020-5 (accessed 18 January 2024).

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 References, no 7. Available at: https://assets.publishing.service.gov.uk/media/5a824465e5274a2e87dc2079/Growing_the_artificial_intelligence_industry_in_the_UK.pdf (accessed 25 September 2020).

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, no 1, pp. 389–399. Available at: https://doi.org/10.1038/s42256-019-0088-2 (accessed 18 January 2024).

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 April. Available at: https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/ (accessed 18 May 2022).

LeCun Y., Bengio Y., Hinton G. (2015) Deep Learning. *Natur*e, no 521, pp. 436–444. Available at: https://doi.org/10.1038/nature14539 (accessed 18 January 2024).

McCarthy J., Minsky M. L., Rochester N., Shannon C. E. (2006) A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. August 31, 1955. *AI Magazine*, vol. 27, no 4, pp. 12–14. Available at: https://doi.org/10.1609/aimag.v27i4.1904 (accessed 18 January 2024).

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

Mittelstadt B. (2019) Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence*, no. , pp. 501–507. Available at: https://doi.org/10.1038/s42256-019-0114-4 (accessed 18 January 2024).

Ouchchy L., Coin A., Dubljevi´c V. (2020) AI in the Headlines: The Portrayal of the Ethical Issues of Artificial Intelligence in the Media. *AI & Society*, vol. 35, pp. 927–936. Available at: https://doi.org/10.1007/s00146-020-00965-5 (accessed 18 January 2024).

Rességuier A., Rodrigues R (2020) AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics. *Big Data & Society*, vol. 7, no 2, art. 2053951720942541. Available at: https://doi.org/10.1177/2053951720942541 (accessed 18 January 2024).

Schölkopf B., Locatello F., Bauer S., Ke N. R., Kalchbrenner N., Goyal A., Bengio Y. (2021) Toward Causal Representation Learning. *Proceedings of the IEEE*, vol. 109, no 5, pp. 612–634. doi: 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. Available at: https://link.springer.com/book/10.1007/978-3-030-69978-9 (accessed 18 January 2024).

UNESCO. (2021) *AI Ethics: Another Step Closer to the Adoption of UNESCO's Recommendation.* Press Release, 2 July, Paris: UNESCO. Available at: https://en.unesco.org/news/ai-ethics-another-step-closer-adoption-unescos-recommendation-0 (accessed 18 May 2022).

Weizenbaum J. (1977) *Computer Power and Human Reason: From Judgement to Calculation*, 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 edn., Thousand Oaks: Sage Publications.

Received: January 9, 2024
Citation: Stahl B. C., Schroeder D., Rodrigues R. (2024) Etika iskusstvennogo intellekta: keysy i variant resheniya eticheskikh problem [Ethics of Artificial Intelligence. Case Studies for Addressing Ethical Challenges (excerpt)]. *Journal of Economic Scoiology = Ekonomicheskaya sotsiologiya*, vol. 25, no 1, pp. 85–95. doi 10.17323/1726-3247-2024-1-85-95 (in Russian).