УДК 172.4 DOI: 10.18698/2306-8477-2018-9-547

Вызовы современности: искусственный интеллект. Этический аспект

© М.В. Моисеенко

Российский университет дружбы народов, Москва, 117198, Россия

Рассмотрен искусственный интеллект как неотъемлемая составляющая жизни современного общества, а также проанализированы дальнейшие перспективы его развития и пути недопущения неблагоприятного воздействия искусственного интеллекта на человечество. Благодаря проникновению искусственного интеллекта в большинство сфер жизни человека, цена ошибки при возникновении сбоев в системе возрастает с каждым днем. В статье описаны способы предотвращения ошибок в работе искусственного интеллекта с учетом современных технологий. Показаны возможные сценарии развития технологии, в результате анализа которых автор приходит к выводу, что без соблюдения этических принципов в разработке искусственного интеллекта гармоничное взаимодействие между людьми и машинами не будет представляться возможным. По итогам Асиломарской конференции, прошедшей в январе 2017 г., был принят ряд универсальных этических принципов, выполнение которых способно уменьшить экзистенциальные риски при сохранении перспективы совершить крупнейший скачок в развитии человечества. Несмотря на отдаленность перспективы создания сверхразумного искусственного интеллекта, уже сегодня разработчики во всем мире должны придерживаться принятых этических норм и нести ответственность за создаваемые технологии, а подготовка талантливых программистов — будущих социально ответственных этических лидеров — становиться приоритетным направлением в образовании.

Ключевые слова: *искусственный интеллект, этические принципы, этика, Асиломарская конференция, экзистенциальные риски, Google, сверхразум*

Термин «искусственный интеллект» ввел в научный оборот американский информатик Джон Маккарти в 1955 г. для обозначения науки создания интеллектуальных компьютерных программ. На настоящий момент под данным термином понимается специализированный искусственный интеллект, который также называют слабым искусственным интеллектом, т. е. разработанным для решения конкретной задачи. Долгосрочной глобальной целью разработчиков по всему миру является создание общего искусственного интеллекта, называемого сильным, который предназначен для решения неограниченного круга задач. Предполагается, что общий искусственный интеллект сможет превзойти человека практически в любой когнитивной работе.

Уже сегодня искусственный интеллект становится неотъемлемой частью жизни людей: мобильные телефоны оснащены голосовыми

помощниками, появляется все большее количество «умных» домов, активно проводятся тестирования беспилотных автомобилей. Если раньше научная фантастика ограничивалась описанием человекоподобных роботов для изображения искусственного интеллекта, то сейчас технология включает в себя довольно широкий спектр разработок: распознавание образов, машинный перевод, вождение автомобилей, прогнозы погоды, рекомендательные системы, чтение рукописей, творчество, финансовые модели, игры, создание лекарств, а также синтез речи. Постепенно внедряясь в жизнь человечества, искусственный интеллект все больше участвует в социальных, экономических и даже политических взаимоотношениях.

Данное исследование опирается на научные работы в сфере искусственного интеллекта и этики, а также на материалы, опубликованные ведущими международными организациями, занимающимися вопросами искусственного интеллекта. Анализируя пути развития искусственного интеллекта, автор оценивает значимость соблюдения этических принципов при разработке новых технологий.

Искусственный интеллект как новая реальность. Искусственный интеллект стал приоритетным направлением для многих компаний и разработчиков. Эта область привлекла серьезное внимание со стороны как ученых и инженеров, так и инвесторов. В настоящее время крупнейшими игроками на рынке являются Google, Apple, Facebook, Microsoft и Uber. В России основной компанией-разработчиком является Яндекс.

В конце 2015 г. в США была создана некоммерческая исследовательская организация OPEN AI. Ее основатели, одним из которых является известный предприниматель Илон Маск, утверждают, что целью деятельности данной компании является разработка дружественного к человеку искусственного интеллекта, а основным критерием разработки — открытость материала и безопасность. Изначально компания создавалась в противовес крупным корпорациям, которые могли получить слишком большое влияние на человечество в случае создания сильного искусственного интеллекта, при этом преследуя коммерческие цели.

Илон Маск предлагает рассматривать применение искусственного интеллекта как продолжение себя. Как использование социальных сетей, приложений в смартфонах и в Интернете делает человека более эффективным в решении повседневных задач, так и повсеместное внедрение искусственного интеллекта должно стать частью человека, а не отдельной противоборствующей единицей, во много раз сильнее каждого из людей. По мнению предпринимателя, искусственный интеллект должен быть доступен максимальному количеству разработчиков и пользователей, и это станет лучшей защитой от его неправомерного использования [1].

Как у любой компьютерной программы, у искусственного интеллекта могут возникать сбои в работе, в связи с чем необходимо учитывать риски возможных последствий. Кроме того, важным критерием является принятие ответственности за произошедший сбой. Если система искусственного интеллекта не справилась с поставленной задачей, то необходимо разграничивать ответственность между программистами, создавшими систему и конечными пользователями [2].

Например, перебои в работе персонального компьютера могут остаться незамеченными, в то время как ошибка в работе искусственного интеллекта, управляющего транспортным средством (автомобилем или самолетом), большой электросетью, кардиостимулятором, может стоить человеческих жизней.

В связи с этим возникают вопросы о необходимости изучения машинного поведения, распознавания психологических патологий на ранней стадии, а также проверки легитимности материала, на котором происходит обучение искусственного интеллекта.

Исследование машинного поведения может быть действенной альтернативой изучению внутренней структуры программы и ее кода в условиях непрозрачности, вызванной защитой интеллектуальных прав. Значительная часть алгоритмов, используемых на постоянной основе, является «черными ящиками» по причине необходимости хранения индустриальных секретов.

Современный искусственный интеллект часто демонстрирует паттерны поведения, предсказать которые не могут даже их собственные программисты. Конечное поведение проявляется только при взаимодействии с миром и другими действующими лицами. Существуют фундаментальные теоретические пределы возможности проверить, что определенная часть кода всегда будет удовлетворять нужным параметрам, если только не запустить код и не увидеть, как он себя ведет [3].

Еще одним важным элементом профилактики сбоев в работе искусственного интеллекта является проверка и изучение качества данных, на которых тренируются алгоритмы. Айлин Калискан из Принстонского университета отметила: «Многие люди считают, что у машин нет предубеждений. Но машины тренируются на человеческих данных. А у людей предубеждения есть» [4].

Самые популярные алгоритмы, используемые сегодня в работе, — это алгоритмы, копирующие архитектуру человеческого мозга, выстраивая сложные модели информации. По причине схожести с человеческим мозгом риск возникновения у данных программ психологических проблем возрастает. Такие проблемы не представляется возможным выявить путем изучения программного кода, потому что они напрямую связаны с внутренним представлением информации.

Есть вероятность, что в ближайшем будущем может появиться новая востребованная специальность в медицине — психотерапия для алгоритмов [4].

Создание искусственного интеллекта и глубинное изучение деятельности мозга являются параллельным процессами. Исследуя мозг для усовершенствования и манипулирования его функциями, многие компании продвинулись гораздо дальше изначально поставленной цели понимания мозга. Таким образом, возникают новые перспективы в расшифровке умственных процессов и манипулировании механизмами мозга, лежащими в основе его намерений, эмоций и решений [5]. Если каждый человек сможет общаться с другими людьми путем передачи мыслей, а мощные вычислительные системы будут связаны напрямую с мозгом, то это станет новым витком в развитии человеческой цивилизации.

Этические принципы разработки искусственного интеллекта и перспективы развития. Среди программных разработчиков и ученых идут споры относительно сроков появления сильного искусственного интеллекта, но уже сейчас становится очевидно, что искусственный интеллект будет развиваться в безопасном для людей направлении только при соблюдении этических принципов его разработки.

В январе 2017 г. на территории государственного парка-пляжа Асиломар в городе Пасифик Гроув в штате Калифорния прошло мероприятие, которое получило название Асиломарская конференция. Решение о проведении конференции было принято в результате признания разработки и внедрения искусственного интеллекта одним из важнейших направлений развития человеческой цивилизации и необходимости оценки потенциальной выгоды, а также возможных экзистенциональных рисков.

Целью данного мероприятия стала разработка перечня этических норм и правил, необходимых для дальнейшей работы ведущих ученых и специалистов со всего мира [6].

В рамках Асиломарской конференции были опубликованы следующие наиболее важные этические принципы и ценности:

- безопасность (системы искусственного интеллекта должны быть безопасны и защищены на протяжении всего срока эксплуатации);
- открытость сбоев в системе;
- ответственность (ответственность за последствия использования и действий ложится на разработчиков и создателей систем искусственного интеллекта);
- синхронизация ценностей (цели и поведение искусственного интеллекта с высокой степенью автономности должны быть согласованы с человеческими ценностями на всем протяжении работы);

- человеческие ценности (работа систем искусственного интеллекта должна быть согласована с идеалами человеческого достоинства, прав, свобод и культурного разнообразия);
- совместная выгода (технологии разработки искусственного интеллекта должны приносить пользу максимально возможному числу людей);
- совместное процветание (экономические блага, созданные с помощью искусственного интеллекта, должны получить максимальное распространение ради принесения пользы всему человечеству);
- контроль человеком (процедура и степень необходимости передачи системе искусственного интеллекта функции принятия решений должна быть определена человеком);
- устойчивость систем (те, кто обладает влиянием, управляя продвинутыми системами искусственного интеллекта, должны уважать общественные процессы и улучшать здоровье социума) [7].

В начале июня 2018 г. технологический гигант Google также опубликовал свои принципы искусственного интеллекта. Многие из них перекликаются с этическими принципами Асиломарской конференции. Публикация Google стала ответом на коллективное письмо сотрудников компании к работодателю, которое подписали более 3100 человек [8]. Основной идеей письма стал отказ от создания военных технологий, использующих искусственный интеллект.

Сундар Пичаи привел список технологий, которые неприемлемы для искусственного интеллекта Google:

- технологии, которые причиняют или могут причинить общий вред;
- оружие или другие технологии, основная цель или осуществление которых состоит в причинении вреда людям;
- технологии, которые собирают или используют информацию для наблюдения, нарушают международные признанные нормы;
- технологии, цель которых противоречит общепризнанным принципам международного права и прав человека [9].

Принципы искусственного интеллекта Google:

- социальная польза;
- борьба с дискриминацией;
- безопасность;
- подотчетность;
- принципы приватности;
- высокие стандарты научного совершенства;
- предоставление технологии искусственного интеллекта только тем, кто соблюдает эти принципы [9].

Придерживаясь этических принципов при создании искусственного интеллекта, человечество сможет подойти к созданию сверхразумного искусственного интеллекта без опасения за свое будущее. Разработанный исключительно в целях, соответствующих этическим идеалам и для общей пользы, а не в интересах одной организации, сильный искусственный интеллект способен помочь человечеству искоренить войны, болезни и бедность и это станет крупнейшим событием во всей истории человечества.

Изменения коснутся сотен миллионов рабочих мест. Люди все больше и больше будут перекладывать часть своих служебных заданий и многие рутинные задачи на машину, что позволит им сосредоточиться на творческой работе.

В сознании людей само понятие интеллекта коррелируется с умом, креативностью, уверенностью в себе, с аналитическим мышлением, развитыми коммуникативными навыками, эффективностью и надежностью. Можно говорить также о национальной специфике восприятия интеллекта. Например, китайцы ассоциируют данное понятие с аналитическим мышлением, четкой памятью, аккуратностью и скромностью, в то время как африканцы больше склоняются к аналогиям с развитыми коммуникативными навыками. Таким образом, люди на уровне устоявшихся образов наделяют интеллект положительными чертами и переносят это восприятие на сверхразумный искусственный интеллект [10]. Однако большинство исследователей считают, что сверхразумный искусственный интеллект вряд ли будет проявлять эмоции, например, любовь или ненависть. По мнению экспертов, наиболее опасными системы искусственного интеллекта могут стать в результате двух следующих сценариев:

1) если искусственный интеллект был изначально разработан с разрушительной целью. Риск существует даже при использовании слабого искусственного интеллекта, но с возрастанием степени его разумности и автономности угроза многократно увеличивается;

2) если искусственный интеллект изначально разработан для позитивной задачи, однако в процессе ее исполнения избирает деструктивный метод достижения цели: данный сценарий может произойти в случае допущения ошибки в синхронизации целей искусственного интеллекта и человечества [11].

Общий искусственный интеллект может быть настолько сильным, что ведущим разработчикам в этой сфере еще предстоит найти пути анализа и предсказания его поведения. У искусственного интеллекта будет меньше известных человеку мотиваций, чем у любого инопланетного существа, так как последний будет являться биологическим существом, прошедшим определенный путь эволюции [12].

Возвращаясь к принципам, разработанным на Асиломарской конференции, в долгосрочной перспективе наиболее важными могут стать следующие аспекты:

- опасность недооценки верхнего порога возможностей сильного искусственного интеллекта;
- потенциальные риски, связанные с системами искусственного интеллекта, должны купироваться действиями, соразмерными возможному масштабу воздействия;
- системы искусственного интеллекта, разработанные для улучшения эффективности собственных алгоритмов и самовоспроизведения, должны стать объектом жесткого регулирования и контроля [7].

В текущей ситуации ответственность за будущее мира во многом ложится на молодое поколение программистов, связавших свою жизнь с разработкой искусственного интеллекта. Осознанность выбора профессии и внутренние этические нормы, составляющие основу личности, могут иметь отражение в деятельности программистов, а значит, и в судьбе человечества.

В 2012 г. группа разработчиков искусственного интеллекта и специалистов по системной безопасности организовала двухнедельную летнюю программу SPARC (Summer Program in Applied Rationality and Cognition — Летняя программа по прикладной рациональности и познанию). Для участия в данной программе каждый год со всего мира отбираются 30–35 наиболее математически одаренных старшеклассников [13]. Основной целью данной программы является обучение будущих технологических экспертов социальным навыкам (эмпатии), а также закрепление важных этических норм, основ эффективного альтруизма. Гуманистическое, гармоничное развитие человеческой личности — стратегическая цель духовного производства, мощной отраслью которого является система образования [14]. В рамках данной программы складывается сообщество юных разработчиков, которые готовы взять на себя ответственность за развитие человечества и стать этическими лидерами будущего.

Международная некоммерческая организация Future of Life, целью деятельности которой является продвижение и поддержка исследований и инициатив, направленных на сохранение жизни на Земле, выделила четыре основных направления своей деятельности: искусственный интеллект, биотехнологии, ядерное оружие, а также климатические изменения. При этом изучение разработок в области искусственного интеллекта является приоритетным направлением, что говорит о всевозрастающем влиянии искусственного интеллекта на жизнь человечества.

Вне зависимости от наличия антропоморфных признаков искусственный интеллект уже сегодня представляет собой новый класс

действующих лиц, а возможно, и новых вид, населяющий Землю. Для реализации возможности совместного процветания людей и машин в здоровом и взаимовыгодном сотрудничестве человечество должно с усердием взяться за изучение машинного поведения, а также четко придерживаться основополагающих этических принципов при разработке искусственного интеллекта.

ЛИТЕРАТУРА

[1] Levy. S. How Elon Musk and Y Combinator Plan to Stop Computers From Taking Over. *Medium*. URL: https://medium.com/backchannel/how-elon-musk-and-y-combinator-plan-to-stop-computers-from-taking-over-17e0e27dd02a (дата обращения 06.07.2018).

[2] Bostrom N., Yudkowsky E. *The Ethics Of Artificial Intelligence.* Cambridge, Cambridge University Press, 2011.

[3] Rahwan I., Cebrian M. Machine Behavior Needs to Be an Academic Discipline. *Nautilus*. URL: http://nautil.us/issue/58/self/machine-behavior-needs-to-be-an-academic-discipline (дата обращения 04.06.2018).

[4] Hills T. Does my algorithm have a mental-health problem? *Aeon*. URL: https://aeon.co/ideas/made-in-our-own-image-why-algorithms-have-mental-health-problems (дата обращения 28.05.2018).

[5] Cussin J. Developing Ethical Priorities for Neurotechnologies and AI. *Future of life institute*. URL: https://futureoflife.org/2017/11/09/developing-ethical-priorities-neurotechnologies-ai/ (дата обращения 02.06.2018).

[6] Леонов В.В. Двадцать три принципа Асиломара. *Современное машиностроение.* URL: https://www.sovmash.com/node/348 (дата обращения 22.05.2018).

[7] Asilomar AI Principles. *Future of life institute*. URL: https://futureoflife.org/ai-principles/ (дата обращения 22.05.2018).

[8] ‘The Business of War’: Google Employees Protest Work for the Pentagon. *The New York Times*. URL: https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html (дата обращения 01.06.2018).

[9] Artificial Intelligence at Google. Our Principles. *Google Al.* URL: https://ai.google/principles (дата обращения 10.06.2018).

[10] Muehlhauser L., Helm L. *Intelligence Explosion and Machine Ethics*. Berlin, Machine Intelligence Research Institute, 2012.

[11] Benefits and Risks of Artificial Intelligence. *Future of life institute*. URL: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/ (дата обращения 25.05.2018).

[12] Bostrom N. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. Oxford, Oxford University Press, 2012.

[13] What is SPARC? *SPARC*. URL: https://sparc-camp.org/ (дата обращения 01.06.2018).

[14] Цвык И.В. Компьютерные технологии в современном образовательном процессе: этический аспект. *Вестник Российского университета дружбы народов. Сер. Философия*, 2017, т. 21, № 3, с. 379–388.

Статья поступила в редакцию 20.07.2018

Ссылку на эту статью просим оформлять следующим образом:

Моисеенко М.В. Вызовы современности: искусственный интеллект. Этический аспект. *Гуманитарный вестник*, 2018, вып. 9.

http://dx.doi.org/10.18698/2306-8477-2018-9-547

Моисеенко Марина Валентиновна окончила философский факультет Московского государственного университета им. М.В. Ломоносова в 1986 г. Канд. филос. наук, доцент кафедры этики факультета гуманитарных и социальных наук Российского университета дружбы народов. Область научных интересов: духовно-нравственное наследие российского зарубежья, этические аспекты искусства, этические аспекты религии, соотношение политики и морали, проблемы профессиональной этики, этические аспекты развития искусственного интеллекта.

e-mail: moiseenko_mv@rudn.university

The challenges of modernity: Artificial intelligence. Ethical aspect

© M.V. Moiseenko

Peoples' Friendship University, Moscow, 117198, Russia

The article considers artificial intelligence as an integral part of modern society life, and analyzes further prospects for its development and ways to prevent the adverse impact of artificial intelligence on humanity. The cost of the error, if it occurs in the system, increases every day due to the penetration of artificial intelligence in most areas of human life. The article describes ways to prevent errors in the work of artificial intelligence, using modern technologies. The article also considers possible scenarios of technology development, their analysis results in the conclusion that without observance of ethical principles in the development of artificial intelligence harmonious interaction between people and machines is not possible. Following the results of the Asilomar conference held in January 2017, a number of universal ethical principles were adopted; their implementation can reduce existential risks while preserving the prospect of making the biggest leap forward in the development of mankind.Although the prospect of creating superintelligence seems to be far away, right now the developers all over the world must follow the accepted ethical rules and be responsible for the technologies being created, and training talented programmers — future socially responsible ethical leaders must become a priority in education.

Keywords: *artificial intelligence, ethical principles, ethics, Asilomar conference, existential risks, superintelligence*

REFERENCES

[1] Levy. S. How Elon Musk and Y Combinator Plan to Stop Computers From Taking Over. *Medium*. Available at: https://medium.com/backchannel/how-elon-musk-and-y-combinator-plan-to-stop-computers-from-taking-over-17e0e27dd02a (accessed July 6, 2018).

[2] Bostrom N., Yudkowsky E. *The Ethics of Artificial Intelligence.* Cambridge, Cambridge University Press Publ., 2011.

[3] Rahwan I., Cebrian M. Machine Behavior Needs to Be an Academic Discipline. *Nautilus*. Available at: http://nautil.us/issue/58/self/machine-behavior-needs-to-be-an-academic-discipline (accessed June 4, 2018).

[4] Hills T. Does my algorithm have a mental-health problem? *Aeon*. Available at: https://aeon.co/ideas/made-in-our-own-image-why-algorithms-have-mental-health-problems (accessed May 28, 2018).

[5] Cussin J. Developing Ethical Priorities for Neurotechnologies and AI. *Future of life institute*. Available at: https://futureoflife.org/2017/11/09/developing-ethical-priorities-neurotechnologies-ai/ (accessed June 2, 2018).

[6] Leonov V.V. Dvadtsat tri printsipa Asilomara [Twenty-three Asilomar Principles]. *Sovremennoe mashinostroenie — Sovmash.com*. Available at: https://www.sovmash.com/node/348 (accessed May 22, 2018).

[7] Asilomar AI Principles. *Future of life institute*. Available at: https://futureoflife.org/ai-principles/ (accessed May 22, 2018).

[8] 'The Business of War': Google Employees Protest Work for the Pentagon**.** *New York Times.* Available at: https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html (accessed June 1, 2018).
[9] Artificial Intelligence at Google. Our Principles. *Google AL.* Available at: https://ai.google/principles (accessed June 10, 2018).
[10] Muehlhauser L., Helm L. *Intelligence Explosion and Machine Ethics*. Berlin, Machine Intelligence Research Institute Publ., 2012.
[11] Benefits and Risks of Artificial Intelligence. *Future of life institute.* Available at: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/ (accessed May 25, 2018).
[12] Bostrom N. *The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents*. Oxford, Oxford University Press Publ., 2012.
[13] What is SPARC? *SPARC*. Available at: https://sparc-camp.org/ (accessed June 1, 2018).
[14] Tsvyk I.V. *Vestnik RUDN, seriya Filisofiya — Peoples' Friendship University Journal of Philosophy,* 2017, vol. 21, no. 3, pp. 379–388.

Moiseenko M.V., Cand. Sc. (Philosophy), Associate Professor, Department of Ethics, Faculty of Humanities and Social Sciences, Peoples' Friendship University of Russia. Research interests: spiritual and moral heritage of the Russian philosophers abroad, the ethical aspects of the art, ethical aspects of religion, the balance between politics and morality, problems of professional ethics, ethical aspects of the development of artificial intelligence. e-mail: moiseenko_mv@rudn.university