Вестник Томского государственного университета. 2020. № 456. С. 99–103. DOI: 10.17223/15617793/456/11

УДК 167.7+179

А.В. Шиллер

МЕСТО ЭТИЧЕСКОЙ СИСТЕМЫ В АРХИТЕКТУРЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Изложены аргументы в пользу актуальности и важности создания этической системы или свода правил этики и морали для искусственного агента. Рассмотрены некоторые методологические сложности, возникающие при разработке этической системы для искусственного интеллекта (ИИ), приведены примеры действующих этических кодексов, а также рассмотрена роль эмоциональной системы в архитектуре «этичного ИИ». Выявлены уровни соприкосновения этики и ИИ.
Ключевые слова: искусственный интеллект; архитектура ИИ; этическая система; эмоциональная система; методология ИИ.

Введение

Различные исследования в области искусственного интеллекта (ИИ) в настоящее время являются популярной темой не только в научных кругах, но и в массовой культуре и СМИ. Одной из причин повышенного интереса общественности к ИИ являются опасения, что скоро ИИ станет «разумным», т.е. перестанет быть инструментом в руках человека и станет самостоятельным агентом. Появление автономности ИИ неизбежно поднимает вопрос о возможном влиянии этих систем на человека и, соответственно, необходимости этической системы у такого искусственного агента. Что означает для системы ИИ принять решение? Каковы моральные и правовые основания действий и решений у таких систем? Может ли система ИИ брать на себя ответственность за совершенные действия (а также обладает ли она достаточным уровнем саморегуляции)? Как можно контролировать такие системы, если они в процессе обучения придут в состояние, очень далекое от исходно заложенного? Можно ли (и нужно ли) допускать такое прогрессивное развитие для коммерческих систем (до уровня саморегуляции и «протосознания»), как их использование и разработка должны регулироваться? Эти и многие другие вопросы являются действительно важными и требуют максимального внимания ученых. То, как научное сообщество и существующие социальные институты будут способны взаимодействовать с системами ИИ, во многом будет определять уровень доверия к ним и в конечном счете влияние ИИ на общество и даже само существование ИИ.

Методологические сложности разработки этической системы

Разработка более прогрессивных версий систем ИИ чаще всего ведется с использованием машинного обучения. На сегодняшний день сложно найти другую такую технологию, которая кажется настолько же перспективной. Количество публикаций с описанием преимуществ и возможных способов использования машинного обучения, предсказывающих невероятные успехи и достижение нового уровеня развития науки, только растет. Однако есть и ряд скептиков, которые задаются вопросом, может ли произойти такое, что при достижении сингулярности искусственный интеллект не много ни мало, а станет главенствовать над человеком. Среди них Ник Бостром, известный по своей работе «Искусственный интеллект. Этапы. Угрозы. Стратегии» [1]. Среди оптимистов развития систем ИИ находится Рэй Курцвейл. Он предсказывает, что к 2029 г. искусственный интеллект превзойдет человеческий, а в 2045 г. наступит сингулярность. Несмотря на то, что такие предсказания кажутся фантастическими, они основаны на осмыслении закона Мура и закона прогрессирующей отдачи, придуманных самим Курцвейлом. Согласно этим законам, компьютерные возможности удваиваются каждые два года, т.е. растут по экспоненте [2]. В отличие от пугающих сценариев будущего, популярных в медиа и СМИ и поддерживаемых некоторыми учеными, в настоящее время ИИ уже меняет повседневную жизнь людей во многих сферах – здоровье, безопасность, продуктивность [6]. Именно поэтому для того, чтобы обеспечить принятие и отсутствие страха в обществе перед системами ИИ, они должны быть представлены таким образом, чтобы вызывать доверие и быть понятными, учитывать общечеловеческие и гражданские права.

Учитывая все эти данные, необходимость в разработке этической программы ИИ уже в настоящее время является актуальной задачей. Одной из сторон этой этической программы должна быть разработка системы включения ИИ в существующие социальные институты, включая правовые системы. Также необходимо разрабатывать автономную этическую систему для включения ее в модели искусственных агентов. Понимание этой необходимости приводит к появлению инициатив как со стороны теоретиков ИИ, так и со стороны разработчиков программного обеспечения, включая инициативу по Этике для Автономных Систем от IEEE [10] или этический кодекс для беспилотных автомобилей, разработанный правительством Германии [11]. Кроме того, уже действующие сейчас программы и искусственные агенты требуют внимания исследователей для анализа того, как эти программы делают моральный выбор в каждодневных ситуациях взаимодействия. Например, это связано с алгоритмами скоринга (принятия банком решений о том, кто получает доступ к ссудам и кредитам) или касается того, как машины с автопилотом принимают решение и подсчитывают ценность человеческой жизни в случае неожиданных ситуаций на дороге (по сути, как они решают моральные дилеммы по типу дилеммы вагонетки).

Формализация морали

Основной проблемой при обучении искусственных агентов нормам морали является сложность формализации морали в доступную для обработки компьютерными программами форму, а также тот факт, что не все люди являются носителями и разделяют одинаковые нормы морали и ценности, что не позволяет создать единую модель морали. При решении моральных дилемм люди склонны опираться на эмоции и даже инстинкты, а не на точные вычисления [12, 13]. А искусственные агенты и программы в существующем виде нуждаются в эксплицитных и объективных метриках, измеримых и оптимизированных.

Вероятным решением может быть попытка исследователей и философов сформулировать этические параметры в виде количественных показателей. Другими словами, необходимо сопроводить ИИ эксплицитными ответами и решениями на любые этические дилеммы, с которыми может столкнуться агент. Для этого потребуется совместно выработать согласованный список решений каждой из этических проблем, что является сложной, но не невыполнимой задачей. В настоящее время некоторые разработчики машин на автопилоте уже включают в архитектуру ИИ набор этических ценностей для обеспечения защиты человеческой жизни. При этом машинам «запрещено взаимно компенсировать жертвы», т.е. ИИ не сможет сделать выбор – кого убить – в случае неизбежности аварии со смертельным исходом, на основе индивидуальных особенностей жертвы (физического состояния, возраста, гендера и т.д.) [8].

Еще одним примером заложенного этического принципа является принцип из этического кодекса для разработчиков беспилотного транспорта, разработанного министерством транспорта и цифровой инфраструктуры Германии [11]. Согласно этому принципу, в случае аварийной ситуации автомобиль должен руководствоваться следующим утверждением: жизнь и здоровье человека важнее, чем жизнь животного и сохранность частной собственности. Ранее автоконцерн Mercedes-Benz заявил, что его беспилотные автомобили будут запрограммированы в аварийных ситуациях защищать жизни людей, которые находятся внутри машины, а не снаружи, даже если придется выбирать между спасением одного пассажира и десяти пешеходов [9]. Эти два примера используемых принципов показывают, что разработка общих этических рекомендаций является крайне сложной задачей и в настоящее время единого мнения на счет списка принципов нет.

Обучение системы

Из-за особенностей архитектуры и способов обработки информации ИИ может превосходить человека в компьютерных играх с четкими правилами и границами их применимости благодаря обучению и оптимизации процесса игры для получения большего числа очков с помощью повторных прохождений и обучения на прошлом опыте игр. Именно на этом была основана убедительная победа AlphaGo над человеком при игре в го [14]. ИИ OpenAI обучался при игре в реальном времени с противниками в течение двух недель, после чего обыграл лучшего в мире игрока в Dota2 [15].

Проблемы оптимизации работы системы и обработки информации в реальной жизни гораздо сложнее, чем оптимизация процессов в компьютерной игре. Сложно представить, какой пул тренировочных данных позволит обучить ИИ становиться «справедливее» или избегать расовых и гендерных предрассудков. Кроме того, как уже было сказано выше, невозможно разработать алгоритм обучения тому, что не формализовано в доступной и понятой форме, т.е. пока исследователи не создадут некоторую модель справедливости. Это непростая задача, поскольку предоставление реальных необработанных данных по поведению людей для обучения ИИ может привести к усилению структурной дискриминации путем воспроизведения стереотипов и аттитюдов, заложенных в исходных данных, как уже произошло с чат-ботом Tay (Тэй), разработанным компанией «Майкрософт», который пришлось удалить через несколько дней после его запуска [5]. По самому негативному сценарию обучение на «сырых» данных будет приводить к отказу в обслуживании некоторых категорий людей (система «скоринга» в банковской сфере), затруднять трудоустройство (система тестирований и оценки кандидатов при приёме на работу) или приводить к нарушениям в процессе политического выбора, т.е., по сути, приводить к дискриминации.

Так или иначе, разработчикам необходимо собрать достаточное количество данных для тренировки алгоритмов ИИ. Даже после того, как будут определены специальные метрики для этических ценностей, ИИ может сталкиваться с ситуациями нехватки «чистых» от человеческих влияний данных для тренировки модели. Этические нормы не всегда можно стандартизировать, поэтому сбор «чистых» данных представляется автору отдельной сложной задачей. Разные ситуации требуют разных подходов, а в некоторых ситуациях может быть не одно решение, или даже такие решения, которые нарушают этические нормы, например разрешение на эвтаназию как проявление ценности самостоятельности и свободы воли в противовес ценности человеческой жизни. Один из путей решения проблемы с данными – собрать решения моральных дилемм у миллионов разных людей по типу сбора средств в краудфандинговых проектах. Такой метод уже используется MIT для обучения Moral Machine принятию наилучших этических решений [16]. Но этого тоже может оказаться недостаточно, поскольку исследования показывают, что этические ценности являются культурно-специфическим конструктом и сильно различаются в разных культурах [7]. Ценности очень часто присутствуют имплицитно в процессах рассуждения. Это означает, что используемые методологические подходы и сами данные должны содержать в себе ценности, разделяемые всеми носителями. Если сделать данные ценности эксплицитными, то это повысит уровень доверия к интеллектуальным независимым системам. Поэтому

система рассуждений у искусственных агентов должна быть построена таким образом, чтобы агент был способен принимать во внимание ценности, этические соображения, оценивать и расставлять приоритеты в списке ценностей, разделяемых другими агентами с учетом контекста культуры и религиозных убеждений, а также был способен к mindreading (пониманию мышления другого агента) [17]. И при этом такая система рассуждений должна оставаться прозрачной и понятной для создателей искусственного агента.

Пересмотр ответственности

Кроме того, поскольку способности независимого принятия решений у искусственных агентов возрастают, возможно, самым важным вопросом, требующим внимания ученых, является вопрос пересмотра ответственности [4]. Несмотря на уровень автономности, социальную осознанность и способность обучаться, системы ИИ являются объектами, созданными людьми для достижения каких-либо целей.

Согласно авторскому взгляду, первым шагом на пути создания «ответственного» ИИ будет выбор параметров, по которым решения ИИ относительно этических вопросов будут более прозрачными. В таком случае, если ИИ совершит ошибку, которая будет иметь заметные последствия, в качестве оправдания нельзя будет использовать отговорку про «особенности алгоритма». Однако также понятно, что требования полной алгоритмической прозрачности технически несостоятельны. Нейросети подходящей сложности для поддержки этической системы ИИ должны быть настолько сложны, что осуществление проверки их алгоритмов человеком будет затруднено. Кроме того, существует проблема черного ящика – разработчики не могут проследить путь, по которому нейросеть действует [18]. Поэтому требования прозрачности должны применяться к процессу количественного определения этических ценностей и последствий совершаемого выбора до того, как включить их в архитектуру модели. Используемые теории, методы и алгоритмы должны содержать в себе отобранные ценности и учитывать их на всех этапах разработки (анализ, дизайн, конструирование, развертывание и оценка).

Важно отметить, что разработка «ответственного ИИ» отражает соответствие этой искусственной системы фундаментальным человеческим принципам и ценностям и подчеркивает ответственность разработчиков за свой продукт. По факту, ответственный ИИ – это нечто большее, чем просто «галочка» в списке о наличии этической системы, или разработка некоторых дополнительных функций или отключение «кнопки» в ИИ системе. Ответственность является базовой характеристикой для автономии и должна быть одним из ключевых оснований при исследовании ИИ.

При этом вопрос о том, кто ответствен за действия искусственной системы, остается открытым. В настоящее время для беспилотных автомобилей в случае аварии ответственность несет либо компания-производитель автомобиля, либо компания-разработчик программного обеспечения для этого автомобиля, если это разные юридические лица. Но в будущем, когда ИИ будет обладать автономией, станет ли он считаться субъектом права – этот вопрос является дискуссионным и требует дополнительных исследований.

Эмоциональная система

Этическая система не сможет существовать без разработанной аффективной или же целостной аффективно-когнитивной системы в том случае, если за образец этической системы будет взята человеческая модель, поскольку эмоции во многом влияют на принятие решений человеком. Поэтому, по мнению автора статьи, еще одной важной задачей является разработка действующей модели эмоций, применимой для ИИ, раскрытие и последующий дизайн связей между эмоциональной и этической системами в архитектуре искусственных агентов. Вероятно, необходимо также предусмотреть возможность отключения эмоциональной системы для некоторых видов ИИ по желанию пользователя, взаимодействующего с искусственным агентом. Исследования показывают, что люди могут испытывать негативные эмоции при взаимодействии с ИИ с несовершенной эмоциональной системой (по аналогии с эффектом зловещей долины) [19], кроме того, люди склонны считать эмоциональность фактором-деструктором для процессов мышления, с чем до недавнего времени соглашались и ведущие нейроученые [3].

Заключение

Можно отметить, что все представленные выше идеи свидетельствуют о связи этики и исследований ИИ на нескольких уровнях:

– уровень технической / алгоритмической интеграции способностей к этическим рассуждениям как части поведения интеллектуальной автономной системы;

– уровень регуляторных и инженерных методов, которые поддерживают для ИИ проведение анализа и оценки этических вопросов, помогают интегрировать или перемещать традиционные социальные конструкты;

– уровень стандартов и процессов сертификации, который обеспечивает взаимодействие разработчиков и пользователей в то время, как они исследуют, разрабатывают, конструируют, используют и управляют системами ИИ.

Важно отметить, что с учетом всё возрастающей включенности ИИ в повседневную жизнь людей, необходимо уже сейчас разрабатывать этические системы для искусственных агентов. Для решения этой задачи необходимы такие подходы, которые стали бы основой для выбора архитектуры искусственных агентов, регулировали возможности развития систем ИИ, чтобы обеспечить качественное управление данными и определить степень вовлеченности человека в процесс управления этими системами. В разработке методологии и архитектуры систем ИИ не последнюю роль игра-

ет философия как наука – источник множества этических программ и принципов: этики И. Канта, Д. Юма, принципа благоговения перед жизнью А. Швейцера, основанного на этике ненасилия, и многих других.

ЛИТЕРАТУРА

1. Бостром Н. Искусственный интеллект. Этапы. Угрозы. Стратегии. М. : Манн, Иванов и Фербер, 2016. 760 с.
2. Курцвейл Р. Transcend: девять шагов на пути к вечной жизни. М. : Манн, Иванов и Фербер, 2015. 520 с.
3. Damasio A., Meyer K. Behind the looking-glass // Nature. 2008. Vol. 454. P. 167–168.
4. Dignum V. Responsible autonomy // Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'2017). 2017. P. 4698–4704.
5. Price R. Microsoft is deleting its AI chatbot's incredibly racist tweets. 2016. URL: http://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3
6. Stone P., Brooks R., Brynjolfsson E., Calo R., Etzioni O., Hager G., Hirschberg J., Kalyanakrishnan S., Kamar E., Kraus S., Leyton-Brown K., Parkes D., Press W., Saxenian A., Shah J., Tambe M., Teller A. Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel. URL: https://ai100.sites.stanford.edu/sites/g/files/sbiybj9861/f/ai100report10032016 fnl_singles.pdf
7. Turiel E. The culture of morality: Social development, context, and conflict. Cambridge : Cambridge University Press, 2002. 336 p.
8. Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars? 2015. URL: https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be-programmed-to-kill/
9. Save the driver, says Benz. 2016. URL: https://www.motoring.com.au/kill-the-rest-says-benz-104114/
10. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. URL: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html
11. Ethics commission. Automated and connected driving. URL: https://www.bmvi.de/SharedDocs/EN/Documents/G/ethic-commission-report.pdf?__blob=publicationFile
12. Сериков А.Е. Дискуссия об инстинктах человека в психологии и этологии // Вестник Самарской гуманитарной академии. Серия «Философия. Филология». 2015. № 1 (17). С. 65–87.
13. Восковская Л.В., Куликов Д.К. Когнитивные функции дилеммы в свете проблем искусственного интеллекта // Инженерный вестник Дона. 2014. № 1, ч. 2. URL: https://cyberleninka.ru/article/n/kognitivnye-funktsii-dilemmy-v-svete-problem-iskusstvennogo-intellekta/viewer
14. Кловайт Н., Ерофеева М. Работа в эпоху разумных машин: зарождение невидимой автоматизации // Философско-литературный журнал «Логос». 2019. Т. 29, № 1. С. 53–84.
15. OpenAI five finals. URL: https://openai.com/blog/openai-five-finals/
16. MoralMachine. URL: https://moralmachine.mit.edu
17. Questa V. Mindreading and Empathy // Rivista Internazionale di Filosofia e Psicologia. 2015. Vol. 6, № 2. P. 261–271.
18. Buckner C. Empiricism without magic: transformational abstraction in deep convolutional neural networks // Synthese. 2018. 195 (12). P. 5339–5372. DOI: 10.1007/s11229-018-01949-1
19. Bechmann A., Lomborg S. Why people hate the paperclip. New Media and Society. 2013. DOI: 10.13140/RG.2.1.2508.1047.

Статья представлена научной редакцией «Философия» 1 июля 2020 г.

The Place of the Ethical System in the Architecture of Artificial Intelligence

Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal, 2020, 456, 99–103.
DOI: 10.17223/15617793/456/11

Alexandra V. Shiller, Lomonosov Moscow State University (Moscow, Russian Federation). E-mail: shiller.a@gmail.com

Keywords: artificial intelligence; AI architecture; ethical system; emotional system; AI methodology.

The aim of the article is to analyze the existing views on the need to develop an ethical system for artificial intelligence and to identify methodological difficulties arising from the implementation of the idea of "ethical AI". The article presents arguments in favor of the relevance and importance of creating an ethical system or a set of rules of ethics and morality for an artificial agent. The stated ideas are based on the research of theorists and philosophers of AI, as well as on the results of the author's analysis of documents: current ethical codes of a number of companies and states (MIT Moral Machine, IBM Watson, Mercedes Benz autonomous cars' code, Code of Ethics for autonomous vehicles of the Government of Germany, analysis of the work of Tay and Zo chat bots from Microsoft). The article describes some methodological difficulties arising in the development of an ethical system for artificial intelligence. The complexity or impossibility of formalizing morality in a form accessible for processing by computer programs stand out among the difficulties. For example, it is challenging to describe morality in a quantitative format; put restrictions on the "purity" and amount of data for training the system—training on real data of people's behavior will lead to unpredictable results, since people do not tend to make decisions based on logic or rational arguments. There is a problem of data opacity or "black box"—the inability to trace the decision-making process of the neural network due to its extreme complexity. There is a need for an emotional system in the architecture of "ethical AI"; the presence of such a system makes AI more "humanlike" and brings in elements of "negligence" and errors, which is necessary when requesting to create an anthropomorphic AI. In addition, the issue of legal liability for the actions of AI and the legal status of AI in the future is raised. The levels at which AI and ethics are connected and interact with each other are revealed: the level of technical/algorithmic integration of ethics into the intellectual system; the level of regulatory and engineering methods for analyzing and evaluating ethical issues; the level of action codes, standards and certification processes, which allows developers and users to interact with AI in real time. As a conclusion, it is noted that due to the increasing involvement of AI in the everyday life of people, the development of an ethical system for AI is no longer just a matter of theoretical studies, but becomes an urgent problem with the ability to use research results or check ideas in practice. In the development of the methodology and architecture of AI systems, a great role belongs to philosophy as a source of many ethical programs and principles; therefore, philosophers need to participate in the interdisciplinary research of existing AI systems and in the development of new ones.

REFERENCES

1. Bostrom, N. (2016) *Iskusstvennyy intellekt. Etapy. Ugrozy. Strategii* [Superintelligence: Paths, Dangers, Strategies]. Translated from English. Moscow: Mann, Ivanov i Ferber.
2. Kurzweil, R. (2015) *Transcend: devyat' shagov na puti k vechnoy zhizni* [Transcend: Nine Steps to Living Well Forever]. Translated from English. Moscow: Mann, Ivanov i Ferber.

3. Damasio, A. & Meyer, K. (2008) Behind the looking-glass. *Nature*. 454. pp. 167–168.
4. Dignum, V. (2017) Responsible autonomy. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'2017)*. pp. 4698–4704.
5. Price, R. (2016) *Microsoft Is Deleting Its AI Chatbot's Incredibly Racist Tweets.* [Online] Available from: http://www.businessinsider.com/microsoft-deletes-racist-genocidal-tweets-from-ai-chatbot-tay-2016-3.
6. Stone, P. et al. (2016) *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel*. [Online] Available from: https://ai100.sites.stanford.edu/sites/g/files/sbiybj9861/f/ai100report10032016 fnl_singles.pdf.
7. Turiel, E. (2002) *The Culture of Morality: Social Development, Context, and Conflict*. Cambridge: Cambridge University Press.
8. MIT Technology Review. (2015) *Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?* [Online] Available from: https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be-programmed-to-kill/.
9. Taylor, M. (2016) *Save the driver, says Benz*. [Online] Available from: https://www.motoring.com.au/kill-the-rest-says-benz-104114/.
10. IEEE. (n.d.) *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. [Online] Available from: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.
11. Ethics Commission. (2017) *Automated and Connected Driving.* [Online] Available from: https://www.bmvi.de/SharedDocs/EN/Documents/G/ethic-commission-report.pdf?__blob=publicationFile.
12. Serikov, A.E. (2015) Discussion on Human Instincts in Psychology and Ethology. *Vestnik Samarskoy gumanitarnoy akademii. Seriya "Filosofiya. Filologiya"*. 1 (17). pp. 65–87.
13. Voskovskaya, L.V. & Kulikov, D.K. (2014) Cognitive Functions of Dilemma in Light of Artificial Intelligence Problems. *Inzhenernyy vestnik Dona – Engineering Journal of Don.* 1 (2). [Online] Available from: https://cyberleninka.ru/article/n/kognitivnye-funktsii-dilemmy-v-svete-problem-iskusstvennogo-intellekta/viewer. (In Russian).
14. Klowait, N. & Erofeeva, M. (2019) Work in the Age of Intelligent Machines: The Rise of Invisible Automation. *Logos*. 29 (1). pp. 53–84. (In Russian).
15. OpenAI. (2019) *OpenAI Five Finals*. [Online] Available from: https://openai.com/blog/openai-five-finals/.
16. *Moral Machine.* [Online] Available from: https://moralmachine.mit.edu.
17. Questa, V. (2015) Mindreading and Empathy. *Rivista Internazionale di Filosofia e Psicologia.* 6 (2). pp. 261–271.
18. Buckner, C. (2018) Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese*. 195 (12). pp. 5339–5372. DOI: 10.1007/s11229-018-01949-1
19. Bechmann, A. & Lomborg, S. (2013) Why people hate the paperclip. *New Media and Society*. DOI: 10.13140/RG.2.1.2508.1047

Received: 01 July 2020