СОВРЕМЕННЫЕ ВЫЗОВЫ В ЭТИКЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А. Р. Каримов
доктор философских наук, Казанский федеральный университет, Университет Иннополис

М. Г. Хорт
кандидат философских наук, Казанский федеральный университет

АННОТАЦИЯ

В литературе, посвященной ИИ, разделяют три вида ИИ: узкий, общий и суперинтеллект (). Под «узким ИИ» понимается ИИ, направленный на решение какого-то определенного круга задач, таких как распознавание речи и лиц или машинный перевод. «Общий ИИ» обозначает способность достигать множества целей и выполнять множество задач в самых разных контекстах. Такой интеллект будет во-многом подобен человеческому. Ну и наконец, «искусственный суперинтеллект» описывает сценарий, в котором ИИ стремительно самоулучшается и превосходит человеческий интеллект - другими словами, он становится сверхразумным.

Ключевые слова: искусственный интеллект, этика искусственного интеллекта, этика добродетели, машинная этика, моральная ответственность.

До недавнего времени Считалось, что мы далеки от общего ИИ, не говоря уже о суперинтеллекте. Тем не менее, даже «узкий ИИ» способен в определенных областях заменять человеческий интеллект, о чем свидетельствует недавний случай успешной защиты в одном из российских вузов дипломной работы, почти полностью написанной текстовой нейросетью ChatGPT от американской компании OpenAI[1]. Нейросеть от OpenAI примечательна тем, что она является мультимодальной. Это означает, что она не заточена на решение какой-то узкой задачи, например, перевод или распознавание текста или изображения. Потенциально ее применение не имеет ограничений. В связи с этим хотелось бы отметить также недавнее обращение, подписанное лидерами хай-тек индустрии, такими как Илон Маск и Стив Возняк, о необходимости приостановить

[1] https://www.kommersant.ru/doc/5798187

эксперименты с ИИ, пока не будут определены четкие правила его использования[2]. В открытом письме, подписанном свыше 1500 представителей ИТ-индустрии, в частности отмечается, что «усовершенствованный ИИ может представлять собой глубокое изменение в истории жизни на Земле, и его следует планировать и управлять им с соразмерной тщательностью и ресурсами»[3].

В правовых документах РФ искусственный интеллект определяется как «комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека (№ 123-ФЗ от 24 апреля 2020 г.).

В 1990-х и начале 2000-х годов были достигнуты многие знаковые цели развития искусственного интеллекта. Например, в 1997 году Гари Каспаров (гроссмейстер и чемпион мира по шахматам) был побежден компьютером Deep Blue от компании IBM. В том же году программа распознавания речи, разработанная компанией Dragon Systems, была впервые внедрена в Windows, что открыло путь к созданию современных виртуальных помощников (таких как Сири, Алекса, Алиса и т.д.), способных общаться с человеком с помощью синтеза речи. Аналогичным образом, искусственные машины стали способны распознавать человеческие эмоции с помощью анализа лица (Breazeal 2002). Мы живем в эпоху "больших данных", в эпоху, когда у нас есть возможность собирать огромные объемы информации и обрабатывать ее с помощью искусственных машин гораздо эффективнее, чем это мог бы сделать человек (или группа людей). В этом контексте Крижевский и др. (2012), Дин и др. (2012) продемонстрировали огромную силу методов глубокого обучения для нескольких отраслей и различных сфер (от машиностроения и банковского дела до маркетинга и даже развлечений). В долгосрочной перспективе целью исследователей, работающих над ИИ, является создание систем и искусственных машин, которые будут способны превзойти когнитивные способности человека практически во всех задачах и областях. Интересным примером является создании ИИ, выполняющего функции «цифрового прокурора». В конце 2021 г. китайские ученые разработали ИИ, который способен предъявить обвинение с точностью в 97% на основе

[2] https://futureoflife.org/open-letter/pause-giant-ai-experiments/
[3] Там же.

устного описания дела. По мнению ученых, система способна в определенной степени заменить прокуроров в процессе принятия решений. Обвинение цифровой прокурор выносит в зависимости от тысячи характерных черт, которые его система распознает во время анализа описания дела[4]. Соединение ИИ с робототехникой также приводят к впечатляющим результатам. Например, социальные роботы, разработанные в Японии, могут выполнять небольшие задачи, например, приносить еду и воду. Некоторые роботы для ухода за пожилыми людьми удовлетворяют социальные и эмоциональные потребности, развлекая их с помощью игр, напоминая о событиях и встречах, а также обеспечивая социальную активность. Другие роботы для ухода за пожилыми людьми используют мощную гидравлику для обеспечения мобильности и транспортной поддержки пожилых людей [].

В то же время, разработка ИИ чревата множеством проблем, которые стали предметом обширных дискуссий в философской литературе. К таким проблемам обычно относят проблему цифровой безработицы, проблему сбора и сохранения персональных данных, используемых для обучения ИИ, проблему атрибуции ответственности за действия систем ИИ, проблема «черного ящика» (отсутствия транспарентности принятия решений), проблема искусственной тупости (ошибки, совершаемые ИИ), проблема прав роботов и этического отношения к роботам и т.д.

Поскольку в данной статье нас в первую очередь интересуют проблемы этики ИИ, возьмем в качестве примера проблему ответственности за беспилотный автомобиль. Европейские и американские руководства различают шесть типов беспилотных автомобилей, в зависимости от уровня вмешательства человека. Начиная с 3-го уровня человек не обязан постоянно мониторить функции автомобиля. Теперь предположим, что произошла авария, когда автомобиль контролировался автопилотом. Кто виноват и кто отвечает? Водитель? Кажется, что водитель виноват в самой меньшей степени, ведь он просто находился в машине, а сама машина в это время управлялась ИИ. Может быть, виновен владелец автомобиля (предположим, что владелец и водитель не одно и то же лицо)? Но владелец использовал лицензированный и легально приобретенный софт. Компания-разработчик? Да, большинство будет склонно предъявлять обвинение разработчику по аналогии с обычной поломкой. Ведь, если, например, произошла авария, и не сработала подушка безопасности, то как правило, претензии

[4] https://lenta.ru/news/2021/12/27/ai_prosecutor/

предъявляют разработчику. Здесь необходимо различать моральную и юридическую ответственность. Если с вопросом юридической ответственности более или менее понятно, то с вопросом моральной ответственности возникают проблемы. Должен ли разработчик нести моральную ответственность за решение, которое принял ИИ? Все эти вопросы требуют наличия определенной системы принципов, в рамках которой можно было бы оценить поведение разработчика/тех, кто внедряет ИИ/пользователей систем ИИ.

В этике ИИ выделяют две разные, но связанные области (Dignum, 2018):

1) Этика для разработчиков (ethics for designers) – разнообразные этические кодексы, стандарты, которые представляют из себя набор обязательств, которые добровольно принимают на себя разработчики систем ИИ и те, кто их внедряет и эксплуатирует.

2) Машинная этика (ethics by design) – техническая/алгоритмическая интеграция способностей этического рассуждения как часть поведения автономной искусственной системы.

В первом случае нас интересуют нормы, которыми должен руководствоваться человек, а во втором – машина. Что касается различных этических стандартов, то в мире их существуют десятки. Практически у каждой крупной компании-разработчика есть свой кодекс. Например, 7 базовых принципов компании Гугл для разработчиков ИИ приписывают: работу на благо общества, избегать предвзятого отношения по расовому и иному признаку, повышать безопасность использования, быть готовым нести ответственность, сохранять персональные данные, соответствовать наивысшим стандартам качества, ограничивать потенциально опасное и вредные применение.

Проведенный в 2020 г. метаанализ зарубежных руководств по этике ИИ (Fjeld, 2020) выявил наиболее частно встречающиеся в них этические принципы. Наиболее часто упоминаются следующие:

- Справедливость (недискриминация, равенство)
- Честность (транспарентность, признание ошибок)
- Ответственность (за результаты работы ИИ)
- Безопасность (непричинение вреда, сохранение приватности).

В России деятельность разработчиков ИИ регулируется ФЗ № 123-ФЗ от 24 апреля 2020 г., который, в частности, предписывает «1) защищать интересы и права человека, 2) содействовать ответственному использованию технологий ИИ в обществе,

3) стремиться к реализации социального блага, в то же время 4) сохраняя максимальную прозрачность и правдивость в отношении своих возможностей и рисков». Также в 2021 году был согласован и представлен текст «Кодекса этики в сфере ИИ», который носит рекомендательный характер, но среди подписантов все ведущие ИТ-компании России: Яндекс, Сбер, Лаборатория Касперского, Сколтех, Университет Иннополис, VK, МТС – на данный момент всего 100 подписантов. В кодексе (на 10 страницах), в частности, отмечается, что «при развитии технологий ИИ человек, его права и свободы должны рассматриваться как наивысшая ценность» .

Данные четыре установки можно рассматривать как ключевые этические добродетели разработчиков ИИ.

Понятие «техноморальные добродетели» (technomoral virtues) было введено британским философом Ш.Валлор в книге «Технология и добродетели» (2016) для обозначения, качеств, необходимых человеку при взаимодействии с новыми технологиями (биотехнологии, компьютерные технологии, робототехника и т.д.) Шеннон Валлор (Vallor 2016, 2017), опираясь на аристотелевскую, конфуцианскую и буддийскую традиции добродетели, а также на более ранние работы по этике добродетели, созданные западными философами (например, Nussbaum 1999; MacIntyre 1981), предлагает культивировать своего рода морального героя, который выражает эти самые техноморальные добродетели. К ним, по мнению Валлор, относятся: (i) честность, (ii) самоконтроль, (iii) скромность, (iv) справедливость, (v) мужество, (vi) эмпатия, (vii) забота, (viii) цивилизованность, (ix) гибкость, (x) перспективизм, (xi) великодушие и (xii) мудрость. Валлор утверждает, что эти добродетели, вероятно, будут развиваться в будущих техносоциальных контекстах, и поэтому ее таксономию не следует считать исчерпывающей.

REFERENCES:

1. Barrera E (2020) Technology and the virtues: a philosophical guide to a future worth wanting. Glob Med J 12(1):128–131.
2. Bauer WA (2020) Virtuous vs. utilitarian artificial moral agents. AI Soc 35(1):263–271.
3. Breazeal CL (2002) Designing sociable robots. MIT Press, Cambridge.
4. Cath C, Wachter S, Mittelstadt B, Taddeo M, Floridi L (2018) Artificial intelligence and the ‘good society’: the US, EU, and UK approach. Sci Eng Ethics 24(2):505–528

5. Constantinescu M, Crisp R (2022) Can robotic AI systems be virtuous and why does this matter? Int J Soc Robot.
6. Constantinescu M, Voinea C, Uszkai R, Vică C (2021) Understanding responsibility in responsible AI. Dianoetic virtues and the hard problem of context. Ethics Inf Technol 23:803–814.
7. Dignum V (2018) Ethics in artificial intelligence: introduction to the special issue. Springer, Berlin
8. Hallamaa J, Kalliokoski T (2020) How AI systems challenge the conditions of moral agency? In: International conference on human–computer interaction. Springer, Berlin, pp 54–64
9. Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399
10. Peeters A, Haselager P (2021) Designing virtuous sex robots. Int J Soc Robot 13(1):55–66
11. Stahl BC (2021) Concepts of ethics and their application to AI. In:
12. Stahl BC (ed) Artificial Intelligence for a Better Future, Springer. Briefs in Research and Innovation Governance. Springer, Cham, pp 19–33
13. Vallor S (2017) AI and the automation of wisdom. In: Powers T (ed) Philosophy and Computing Essays in Epistemology, Philosophy of Mind, Logic, and Ethics. Springer, Berlin, pp 161–178
14. Wallach W, Vallor S (2020) Moral machine: from value alignment to embodied virtue. In: Liao M (Ed). Ethics of Artificial Intelligence. Oxford University Press, New York, NYC, pp 383–412.