Секция «Этика»

Моральный код искусственного интеллекта: возможно ли запрограммировать этику?

Научный руководитель – Скворцов Алексей Алексеевич

Муминова Парвина Диловаровна
Студент (магистр)
Московский государственный университет имени М.В.Ломоносова, Философский факультет, Москва, Россия
E-mail: muminovaprvn@gmail.com

Искусственный интеллект (ИИ) всё глубже проникает в различные сферы нашей жизни, существенно меняя наши способы жизнедеятельности и наши взгляды на какие-либо вещи. Однако вместе с развитием технологий ИИ возникают закономерные вопросы: способен ли ИИ действовать этично? Может ли он быть моральным агентом? Возможно ли запрограммировать моральный код, который позволит алгоритмам принимать справедливые и гуманные решения? Эти и другие вопросы лежат в основе современных исследований в области этики и являются ключевой темой для дискуссий среди философов, разработчиков и законодателей.

Мораль — это сложная система норм и ценностей, формирующаяся под влиянием различных факторов, таких, например, как культура, религия и общество [1]. В отличие от людей, машины не обладают сознанием, эмпатией и интуицией, что ставит под сомнение их способность действительно понимать моральные дилеммы и осознанно принимать моральные решения [3]. Тем не менее, существует несколько подходов к созданию этичного ИИ.

Один из них — подход, основанный на правилах (rule-based AI), когда алгоритмы действуют по заранее сформулированным и запрограммированным принципам и правилам, исключая нежелательные сценарии. Зависимость от правил гарантирует, что каждую операцию и решение, которые принимает ИИ, можно отследить до определенного набора руководящих принципов и правил, разработанных специалистами. Среди преимуществ данного подхода — прозрачность и предсказуемость. Данный подход широко применяется, например, в здравоохранении, когда ИИ необходимо определить диагноз на основе симптомов или набора анализов. Но данный подход имеет и недостатки, среди которых, например: возможность возникновения конфликта или пересечения правил и необходимость следить за актуальностью правил. Другой вариант — подход, основанный на использовании машинного обучения (ML-based AI), когда алгоритмы анализируют данные и принимают решения на их основе. Данный подход более универсален, поскольку в то время rule-based AI действует по заранее запрограммированным правилам, данный подход способен обрабатывать большие массивы данных и различать закономерности, преодолевая сложность и неоднозначность различных ситуаций. Однако этот метод несет в себе риск переноса предвзятости из обучающего материала [7].

Различные международные организации, включая IEEE (Institute of Electrical and Electronics Engineers) и Европейский Союз, занимаются разработкой стандартов и других документов для этичного ИИ. Например, отчет Европейской комиссии об этических принципах для заслуживающего доверия ИИ служит четким ориентиром для ответственного развития ИИ, способствуя международной поддержке решений, которые приносят пользу человечеству [2]. Однако разработка этичного ИИ и соответствующих документов сталкивается с рядом проблем, одна из главных — культурная относительность морали. То, что считается допустимым в одной стране, может быть неприемлемым в другой. Например,

критерии справедливого распределения ресурсов или допустимость сбора персональных данных могут существенно различаться [6]. Еще одна проблема — моральные дилеммы, с которыми сталкивается ИИ. Классический пример — дилемма вагонетки, когда автономный автомобиль должен выбрать между наездом на одного пешехода или группу людей. Каким принципом он должен руководствоваться? В таких ситуациях даже люди не всегда могут прийти к единому решению [4]. Кроме того, встает вопрос ответственности. Если ИИ допустил ошибку или совершил преступление, кто будет ответственным за это — разработчик, владелец технологии или сам ИИ? В настоящее время правовая система большинства стран не предусматривает самостоятельную ответственность ИИ, что затрудняет регулирование таких ситуаций [5].

Одно остается очевидным: если вопросы этики не будут должным образом учтены, это может привести к серьезным последствиям. Всем нам известны случаи, когда алгоритмы принимали дискриминационные решения. Например, автоматизированные системы оценки резюме некоторых компаний отдавали предпочтение мужчинам, поскольку обучались на исторически предвзятых данных [8]. Другой риск — использование ИИ для манипуляции и дезинформации. Например, технологии deepfake позволяют создавать реалистичные поддельные видео, которые могут использоваться для дискредитации людей или распространения ложной информации.

Некоторые исследователи полагают, что в будущем ИИ сможет самостоятельно разрабатывать моральные принципы на основе анализа массивов данных. Однако пока такие перспективы далеки от реальности. Большинство экспертов сходятся во мнении, что ключевым элементом этичного ИИ должен оставаться человеческий контроль [6]. Оптимальная стратегия — это гибридный подход, при котором ИИ работает в рамках четко определенных этических норм и регулируется людьми. Это требует развития законодательных инициатив, международных стандартов и прозрачных алгоритмов [5]. Только так можно будет создать искусственный интеллект, который не просто выполняет задачи, но и действует в интересах общества, соблюдая моральные принципы.

Вопрос о том, возможно ли запрограммировать мораль, на данный момент остается открытым. Продвинутые технологические решения уже существуют, но они сталкиваются с рядом ограничений, на преодоление которых понадобится время и ресурсы. Несмотря на то, что существует широкое согласие в отношении этических принципов, которыми следует руководствоваться при разработке ИИ, их эффективное программирование сопряжено с определенными трудностями. Чтобы ИИ служил человечеству, а не был источником угроз, его развитие должно сопровождаться продуманными этическими и правовыми механизмами [8].

Источники и литература

1) Апресян, Р. Г. Мораль / Р. Г. Апресян // Новая философская энциклопедия : в 4 т. / пред. науч.-ред. совета В. С. Степин. - 2-е изд., испр. и доп. - Москва : Мысль, 2010. - Т. 2. - С. 607-614.

2) Floridi, L., 2019. Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 1, pp. 261-262. https://doi.org/10.1038/S42256-019-0055-Y

3) Miller, K., 2017. Can We Program Ethics into AI? [Reflections]. IEEE Technol. Soc. Mag., 36, pp. 29-30. https://doi.org/10.1109/MTS.2017.2697085

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

5) Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L., 2021. Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. Minds and Machines, 31, pp.

239-256. https://doi.org/10.1007/s11023-021-09563-w

6) Piano, S., 2020. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Palgrave Communications, 7, pp. 1-7. https://doi.org/10.1057/S41599-020-0501-9

7) Siau, K., & Wang, W., 2020. Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. J. Database Manag., 31, pp. 74-87. https://doi.org/10.4018/jdm.2020040105

8) Shukla, S., 2024. Principles Governing Ethical Development and Deployment of AI. International Journal of Engineering, Business and Management. https://doi.org/10.22161/ijebm.8.2.5