Секция «Философия когнитивных наук и искусственного интеллекта»

Принципы развития этичного искусственного интеллекта

Научный руководитель – Разин Александр Владимирович

Смирнова Анна Ивановна
Аспирант
Московский государственный университет имени М.В.Ломоносова, Философский факультет, Кафедра этики, Москва, Россия
E-mail: k.as58046@gmail.com

По мере развития и распространения полуавтономных и автономных систем ИИ в человеческом обществе все чаще звучат опасения различного толка - к примеру, что ИИ может полностью подменить работников различных специальностей и тем самым привести к сокращению рабочих мест и росту безработицы; быть использованным злонамеренными акторами; привести к размыванию ответственности при принятии решений, а также непреднамеренно породить или усугубить предвзятость и тем самым усилить неравенство и подорвать принципы справедливости в обществе. Ответом на эти вызовы служат достаточно многочисленные отчеты и руководящие документы по этичному ИИ, разрабатываемые международными организациями [2, 4, 5], руководящими органами на уровне отдельных стран [8, 9], представителями научно-образовательного сообщества [3, 6, 7], экспертами-практиками в сфере ИИ[1] и т.д. Значительная часть таких документов является примерами так называемых незаконодательных инструментов политики или мягкого права. В отличие от «жесткого» праваруководящие принципы этики не являются юридически обязывающими, но носят рекомендательный характер и призваны быть ориентирами при принятии решений в определенных областях.Действительно, в некоторых случаях регулятивный подход может быть преждевременным, слишком жестким, что приведет кподавлению ценных инноваций. Этический подход более гибок и, хотя и носит рекомендательный характер, все же предъявляет определенные требования, соответствие которым ожидается от разрабатываемых алгоритмов и систем ИИ.

На основании анализа процитированных выше документов можно утверждать, что среди наиболее часто встречающихся ценностных установок и принциповэтичного ИИ можно назвать следующие:

• прозрачность (в использовании данных, в процессах принятия решений алгоритмами ИИ); • справедливость-недопущение предвзятости и дискриминации (что особенно важно в том случае, когда алгоритм ИИ обучался на данных, в которых в скрытом виде присутствует предвзятость или дискриминация), учет интересов уязвимых групп населения, обеспечение равенства; • непричинение вреда - недопущение предвзятости, нарушения конфиденциальности, нарушения основных человеческих прав и свобод, физического ущерба человеку (в более широкой перспективе также любому живому существу и окружающей среде). Однако остаются дискуссионными принципы этичного «поведения» алгоритма ИИ в случаях, когда при любом развитии событий причинение вреда человеку неизбежно (к примеру, так называемая «проблема вагонетки», ставшая чрезвычайно актуальной с развитием беспилотных автомобилей); также дискуссионным с точки зрения этого принципа остается развитие ИИ в военной сфере, особенно развитие и совершенствование беспилотных аппаратов, имеющих своей основной целью убийство живой силы противника; • ответственность - невозможно задать в алгоритме ИИ однозначное решение морально-этической проблемы, которая не имеет однозначного решения в самом человеческом обществе, либо такое решение будет являться нарушением фундаментальных прав человека. Пока что основным путем решения такой проблемы является отстранение ИИ

от решения подобных проблем и возлагание ответственности на человека (разработчика или оператора ИИ); • конфиденциальность (особенно в плане обеспечения безопасности личных данных), • направленность на благо людей - см. дискуссионные ситуации с точки зрения непричинения вреда; • свобода и автономия - сохранение человеком всей полноты прав на самостоятельное принятие решений; • доверие - этот принцип тесно связан с повышением прозрачности алгоритмов ИИ, а также зиждется на просвещении населения относительно сбора, использования и хранения данных, особенностей функционирования алгоритма ИИ (особенно в части принятия решений); • соблюдение человеческого достоинства и иных фундаментальных прав человека; • устойчивость и солидарность - перераспределение благ от использования ИИ таким образом, чтобы сокращать неравенство в обществе и обеспечивать доступ к этим благам для уязвимых групп населения.

Способность систем ИИ не только оперировать по заданным алгоритмам, но самостоятельно делать моральный выбор и нести персональную ответственность за свои действия превратит их в полноценных моральных агентов. Такие системы ИИ еще не созданы, однако проблема морального выбора уже релевантна для имеющихся алгоритмов, как было показано в примерах выше.

Источники и литература

1) Принципы работы с ИИ, разработанные на Асиломарской конференции. URL: https://futureoflife.org/open-letter/ai-principles-russian/
2) ЮНЕСКО. Рекомендация об этических аспектах искусственного интеллекта. ЮНЕСКО, 2021. URL: https://ifap.ru/ofdocs/unesco/airec.pdf
3) An initiative of the Universite de Montreal. URL: https://recherche.umontreal.ca/english/strategic-initiatives/montreal-declaration-for-a-responsible-ai/
4) G20 Ministerial Statement on Trade and Digital Economy. G20 AI Principles / G20. 2019. URL: https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf
5) Independent High-Level Expert Group on Artificial Intelligence Set Up By The European Commission. Ethics Guidelines for Trustworthy AI. Brussels: European Commission, 2019. URL: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
6) Montréal Declaration for a Responsible Development of Artificial Intelligence. 2018 Report.URL: https://monoskop.org/images/b/b2/Report_Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf
7) Montreal Declaration Responsible AI. URL: https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf
8) Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI) / Ministry of Foreign Affairs of the People's Republic of China. 17.11.2022. URL: https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202211/t20221117_10976730.html
9) The Ethical Norms for the New Generation Artificial Intelligence, China // International Research Center for AI Ethics and Governance. 27.09.2021. URL: https://ai-ethics-and-governance.institute/2021/09/27/the-ethical-norms-for-the-new-generation-artificial-intelligence-china/