УДК 004.896

Объяснимый ИИ для педагогических целей: психологический аспект

Ли Цзе

Магистрант,
Белорусский национальный технический университет,
220013, Республики Беларусь, Минск, просп. Независимости, 65;
e-mail: 2058963665@qq.com

Тянь Биньюань

Магистрант,
Белорусский национальный технический университет,
220013, Республики Беларусь, Минск, просп. Независимости, 65;
e-mail: 1658230951@qq.com

Ли Юнцзе

Аспирант,
Национальная академия наук Беларуси,
220072, Республики Беларусь, Минск, просп. Независимости, 66;
e-mail: 1454527205@qq.com

Аннотация

Статья посвящена актуальной проблеме объяснимого интеллекта в области образования. Цель статьи заключается в выявлении основных психологических эффектов от взаимодействия студентов с моделями искусственного интеллекта, решения которых трудно объяснимы и недостаточно прозрачны. Задачи исследования состоят в том, чтобы обобщить основные теоретические подходы в процессе применения моделей искусственного интеллекта; систематизировать позиции исследователей в отношении психологических последствий взаимодействия студентов с искусственным интеллектом. Методология исследования основана на системном подходе и состоит из общенаучных методов (сопоставление, обобщение, формально-логический метод), а также ряда специальных методов: анализ историографии по изучаемой теме, метод описательного анализа, методы качественного анализа. Авторы статьи приходят к выводу, что студенту необходимо понимать поведение модели ИИ, используемой в процессе обучения, для собственного психологического комфорта. Развитие объяснимого ИИ в педагогике необходимо для нивелирования негативных психологических последствий взаимодействия студентов с моделями ИИ в процессе обучения.

Для цитирования в научных исследованиях

Ли Цзе, Тянь Биньюань, Ли Юнцзе. Объяснимый ИИ для педагогических целей: психологический аспект // Педагогический журнал. 2024. Т. 14. № 9А. С. 18-24.

Ключевые слова

Искусственный интеллект, объяснимость модели, преподавание, педагогика, обучение на базе искусственного интеллекта.

Введение

Актуальность исследования заключается в том, что в сфере образования ответственность за цифровую среду должны нести две категории пользователей: образовательные учреждения в лице преподавателей и администрации, а также сами студенты. Поэтому проблема объяснимости моделей искусственного интеллекта (далее – ИИ) в педагогике имеет не только собственного дидактический, но и психологический ракурс. Вопрос о том, какие именно психологические последствия будет иметь взаимодействие студентов с машинным интеллектом, особенно в части объяснимости того, как именно «думает» модель. От психологических последствий такого взаимодействия преподаватель может решать, имеют ли эти последствия большое значение для более широкого использования интеллектуальных инструментов в обучении. Это означает разработку систем, которые высшие учебные заведения могут адаптировать к своим потребностям. Для этого преподаватель должен уметь понимать решения или рекомендации системы ИИ в отношении своих студентов с позиций психологии, доверять системе и, в некотором смысле, быть в состоянии «стать единым целым» с ней, чтобы иметь возможность объяснить комбинированное поведение устройства в том виде, в котором оно было установлено [Вислова, 2019, 33].

Для нивелирования негативных последствий взаимодействия человека с системами ИИ требуется четкое представление о процессах принятия решений моделью искусственного интеллекта в учебном процессе. Более того, для применения в системе высшего образования вопрос объяснимости модели ИИ должен быть учтен еще на этапе ее проектирования [Вовк, Супрун, 2022, 79].

В рамках современных обучающих проектов исследователи работают «над разработкой моделей и инструментов для реализации компетентностного подхода для поддержки обучения в персонализированной форме» [Гаврилова, 2023, 35].

Структура компетенций, определенная учебной программой любого вуза, представляет собой для данной области навыки, знания и ноу-хау, которые необходимо приобрести. Педагогические ресурсы (курсы или упражнения) связаны со знаниями и ноу-хау преподавателя в области эталонной системы. По следам взаимодействия обучающегося с образовательными ресурсами рассчитывается профиль компетенций каждого обучающегося. Используя стратегию персонализации, разработанную в модели ИИ, в учебном процессе можно применять образовательные ресурсы, адаптированные к целям конкретного студента и профилю его навыков.

При внедрении моделей ИИ в педагогический процесс, с одной стороны, происходит расчет показателей усвоения, представленных в профиле компетенций, а с другой стороны, необходимо учитывать также и психологические аспекты работы студента с ИИ. Объяснимый интеллект, таким образом, способствует укреплению доверия учащегося к системе, а также

поддержке процесса саморегуляции обучения среди преподавателей и студентов [Tripathy, Seetha, 2022, 114].

В частности, объяснение поведения модели ИИ, основанное на удачных или неудачных упражнениях, может привести учащегося к пересмотру своей самооценки. Например, психологически приемлемое объяснение рекомендуемых ресурсов, того, почему определенные навыки нуждаются в доработке в связи с целями обучения студентов, побуждает их сделать шаг назад от своего обучения, подчеркивая разницу между целями обучения и освоенными концепциями или выявляя пробелы в знаниях студентов.

Объяснимые модели ИИ направлены на укрепление доверия студента к системе обучения, а также на то, чтобы он мог психологически комфортно получать знания в рамках обучения с помощью искусственного интеллекта. В этом контексте способность объяснять необходима, но недостаточна. Поэтому исследования в области психологических эффектов, возникающих в ходе взаимодействия пользователей (преподавателей и студентов) и моделей ИИ являются в настоящее время актуальными [Yohei Okada, Yilin Ning, Marcus Eng, Hock Ong, 2023, 72]. Метазнания, которые позволяют системе анализировать собственное поведение, наблюдая за собой [Чуча, 2023], могут быть использованы при обнаружении неисправности или для улучшения функционирования модели.

Материалы и методы

Материалами исследования послужили работы как отечественных, так и зарубежных авторов. В частности, методические особенности изучения проблемы объяснимости искусственного интеллекта в педагогике рассмотрены в работах таких авторов, как А.В. Вислова [Вислова, 2019], Е.В. Вовк, А.А. Супрун [Вовк, Супрун, 2022], Э.Д. Гаврилова [Гаврилова, 2023], Д.В. Ивлев [Ивлев , 20234], А.Г. Кравцова [Кравцова, 2023] и др. В качестве теоретических материалов исследования были применены подходы, отраженные в работах таких авторов, как И.В. Соболев, Е.А. Потапова [Соболев, Потапова, 2022], М.Р. Тохиржонова [Тохиржонова, 2023], Г.Г. Четвериков, Е.С. Кнышева, И.Д. Вечирская [Четвериков, Кнышева, Вечирская, 2019], С.Ю. Чуча [Чуча, 2023] и др.

В работах эмпирических исследований таких авторов, как М. Амир, К. Ахтар, М. Кумар, А. Найяр [Amir, Akhtar, Kumar, Nayyar, 2023], В. Гровер, М. Догра [Grover, Dogra, 2021], З. Пан, П. Мишра [Pan, Mishra, 2021], рассматриваются практические аспекты применения объяснимости искусственного интеллекта в педагогическом процессе с точки зрения психологии взаимодействия человека и искусственных интеллектуальных систем.

Методология исследования основана на системном подходе и состоит из общенаучных методов (сопоставление, обобщение, формально-логический метод); а также ряда специальных методов: анализ историографии по изучаемой теме, метод описательного анализа, методы качественного анализа.

Результаты и обсуждения

С педагогической точки зрения существует множество ситуаций, в которых крайне важно оснастить систему ИИ объяснительными возможностями.

Данная проблема далеко не нова в искусственном интеллекте, она вновь стала центральной с недавним развитием многих так называемых систем «черного ящика» [Ивлев, 2023, 129]. Здесь

полезно отметить, что использование этого термина имеет тенденцию путать два очень разных контекста: невозможный доступ к модели – в случае проприетарной системы – и предположительно слишком большой сложности модели, что затрудняет понимание лежащей в основе логики [Кравцова, 2023, 37].

Последствия с точки зрения возможных объяснений различаются. Объяснимость может быть важна в первую очередь на этапе разработки приложения, чтобы позволить эксперту отладить, пересмотреть и развить свою модель. На этапе использования приложения запросы о гарантиях подотчетности ИИ-систем также стимулируют разработку таких объяснительных функций. Действительно, даже если это не единственный способ обеспечить подотчетность системы ИИ, способность объяснить результат «мышления» ИИ часто считается необходимым условием ее приемлемости, особенно в силу психологических трудностей восприятия такого ИИ, особенно в случае приложений с «высокими ставками», которые могут существенно повлиять на пользователей [Тохиржонова, 2023, 31].

Таким образом, факторы объяснения ИИ взаимосвязаны: у них есть цель, которая сильно зависит от контекста использования, а форма и уровень детализации, которые они должны принимать, могут сильно варьироваться в зависимости от человека, которому они адресованы. Объяснимые модели ИИ также могут быть интегрированы в более богатую интерактивную среду, в частности предлагая возможность оспаривать их. Современный научный дискурс позволяет составить очень богатую панораму исследований, проведенных в области объяснимого ИИ, и проиллюстрировать разнообразие как приложений, так и используемых методов психологической адаптации пользователей в процессе применения моделей ИИ в учебных целях.

По мнению исследователей, наиболее яркими психологическими последствиями взаимодействия студентов с машинным интеллектом без возможности объяснения поведения модели являются:

– снижение собственной самооценки по причине неясности принятия решений на основе модели ИИ;

– появление симптомов тревожности при взаимодействии с плохо объяснимой моделью ИИ (языковые расстройства, письменная лингвофобия, дислексия и т.п.);

– стресс от постоянного взаимодействия с машинным, превосходящим интеллектом;

–стресс от отсутствия эмпатии у плохо объяснимой модели ИИ;

– невозможность эмоционального контакта в процессе обучения.

Области применения объяснимых моделей ИИ охватывают самые широкие сферы: область образования, планирования маршрутов, умных домов, финансового моделирования, безопасности (обнаружение вторжений), создания материалов, а также – что неудивительно, поскольку последствия могут быть критическими – широкий спектр медицинских приложений (фармакология, визуализация и т.д.). Изучаемые методы варьируются от моделей, о которых известно, что они изначально интерпретируемы (системы, основанные на правилах, деревья решений, графы знаний и т.д.), но, тем не менее, требующие специальной обработки для извлечения приемлемых объяснений, от апостериорных подходов, не зависящих от моделей, таких как контрфактические объяснения, до гибридных подходов, которые стремятся объединить знания и нейронные сети с нейросимволической точки зрения.

Что касается вопроса оценки качества предлагаемых решений, психологические последствия объяснимых моделей иллюстрирует различные стратегии, которые дополняют друг друга: формальные гарантии (минимальные объяснения, верность модели), эксперименты с помощью симуляций или с привлечением пользователей (экспертов).

Заключение

По итогу проведенного исследования можно сформулировать следующие выводы:

1. Приведенные выше данные исследований позволяют выявить, насколько важно студенту понимать поведение модели ИИ, используемой в процессе обучения, для собственного психологического комфорта.
2. Развитие объяснимого ИИ в рамках педагогической сферы, особенно в высшем образовании, необходимо для нивелирования негативных психологических последствий взаимодействия студентов с моделями ИИ в процессе обучения.

Библиография

1. Вислова А.В. Потенциал психологии интеллекта в контексте моделирования искусственного интеллекта // Известия КБНЦ РАН. 2019. № 6 (92). С. 32-47.
2. Вовк Е.В., Супрун А.А. Искусственный интеллект и цифровая педагогика как тренд современной образовательной среды высших учебных заведений // Проблемы современного педагогического образования. 2022. № 77-2. С. 78-99.
3. Гаврилова Э.Д. Психологические аспекты моделирования искусственного интеллекта // Международный журнал гуманитарных и естественных наук. 2023. № 12-3 (87). С. 34-59.
4. Ивлев Д.В. Искусственный интеллект и проблемы этики // Право и практика. 2023. № 4. С. 128-139.
5. Кравцова А.Г. CHATGPT-3: перспективы использования в обучении иностранному языку // МНКО. 2023. № 3 (100). С. 33-36.
6. Соболев И.В., Потапова Е.А. Проблема возможности искусственного интеллекта с точки зрения психологической науки // Коллекция гуманитарных исследований. 2022. № 2 (31). С. 42-58.
7. Тохиржонова М.Р. Роль искусственного интеллекта в педагогике, улучшение опыта обучения с помощью интеллектуальных технологий // Теория и практика современной науки. 2023. № 7 (97). С. 28-49.
8. Четвериков Г.Г., Кнышева Е.С., Вечирская И.Д. Концептуально-психологические аспекты построения многозначных систем искусственного интеллекта. Мозгоподобные преобразователи информации // Онтология проектирования. 2019. № 2 (8). С. 88-97.
9. Чуча С.Ю. Искусственный интеллект в правосудии: юридико-психологические аспекты правоприменения // Правоприменение. 2023. № 2. С. 110-124.
10. Amir M., Akhtar K., Kumar M., Nayyar A. Introduction to Ethical and Socially Responsible Explainable AI // In book: Towards Ethical and Socially Responsible Explainable AI. 2023. No. 3. P. 7-39.
11. Grover V., Dogra M. An Overview of Explainable AI Studies in the Prediction // Conference: Medical Informatics Europe. 2021. No. 4. P. 68-79.
12. Pan Z., Mishra P. Hardware Acceleration of Explainable AI // In book: Explainable AI for Cybersecurity. 2021. No. 7. P. 199-220.
13. Tripathy B.K., Seetha H. Explainable, Interpretable, and Transparent AI Systems // University of Luxembourg. 2022. No. 7. P. 112-145.
14. Yohei Okada, Yilin Ning, Marcus Eng, Hock Ong. Explainable AI in Emergency medicine: An overview // November Clinical and Experimental Emergency Medicine. 2023. No. 10 (4). P. 67-89.

Explainable AI for pedagogical purposes: psychological aspect

Li Jie

Master's Student,
Belarusian National Technical University,
220013, 65, Nezavisimosti ave., Minsk, Republic of Belarus;
e-mail: 2058963665@qq.com

Tian Binyuan

Master's Student,
Belarusian National Technical University,
220013, 65, Nezavisimosti ave., Minsk, Republic of Belarus;
e-mail: 1658230951@qq.com

Li Yongjie

Postgraduate Student,
National Academy of Sciences of Belarus,
220072, 66, Nezavisimosti ave., Minsk, Republic of Belarus;
e-mail: 1454527205@qq.com

Abstract

The article reveals an actual problem of explainable intelligence in the field of education. The purpose of the article is to identify the main psychological effects of students' interaction with artificial intelligence models, the solutions of which are difficult to explain and not transparent enough. The research objectives are to summarize the main theoretical approaches in the process of applying artificial intelligence models, systematization of researchers' positions regarding the psychological consequences of students' interaction with artificial intelligence. The research methodology is based on a systematic approach and consists of general scientific methods (comparison, generalization, formal logical method); as well as a number of special methods: analysis of historiography on the topic under study; the method of descriptive analysis, methods of qualitative analysis. The authors of the article conclude that the student needs to understand the behavior of the AI model used in the learning process for their own psychological comfort. The development of explainable AI in pedagogy is necessary to neutralize the negative psychological consequences of students' interaction with AI models in the learning process.

For citation

Li Jie, Tian Binyuan, Li Yongjie (2024) Ob''yasnimyi II dlya pedagogicheskikh tselei: psikhologicheskii aspekt [Explainable AI for pedagogical purposes: psychological aspect]. *Pedagogicheskii zhurnal* [Pedagogical Journal], 14 (9A), pp. 18-24.

Keywords

Artificial intelligence, model explainability, teaching, pedagogy, learning based on artificial intelligence.

References

1. Amir M., Akhtar K., Kumar M., Nayyar A. (2023) Introduction to Ethical and Socially Responsible Explainable AI. In book: *Towards Ethical and Socially Responsible Explainable AI*, 3, pp. 7-39.
2. Chetverikov G.G., Knysheva E.S., Vechirskaya I.D. (2019) Kontseptual'no-psikhologicheskie aspekty postroeniya mnogoznachnykh sistem iskusstvennogo intellekta. Mozgopodobnye preobrazovateli informatsii [Conceptual and Psychological Aspects of Building Multivalued Artificial Intelligence Systems. Brain-like information converters]. *Ontologiya proektirovaniya* [Ontology of design], 2 (8), pp. 88-97.
3. Chucha S.Yu. (2023) Iskusstvennyi intellekt v pravosudii: yuridiko-psikhologicheskie aspekty pravoprimeneniya [Artificial intelligence in justice: legal and psychological aspects of law enforcement]. *Pravoprimenenie* [Law enforcement], 2, pp. 110-124.

4. Gavrilova E.D. (2023) Psikhologicheskie aspekty modelirovaniya iskusstvennogo intellekta [Psychological Aspects of Artificial Intelligence Modeling]. *Mezhdunarodnyi zhurnal gumanitarnykh i estestvennykh nauk* [International Journal of Humanities and Natural Sciences], 12-3 (87), pp. 34-59.
5. Grover V., Dogra M. (2021) An Overview of Explainable AI Studies in the Prediction. *Conference: Medical Informatics Europe*, 4, pp. 68-79.
6. Ivlev D.V. Iskusstvennyi intellekt i problemy etiki [Artificial Intelligence and Ethical Problems]. *Pravo i praktika* [Law and Practice]. 2023. № 4. S. 128-139.
7. Kravtsova A.G. (2023) CHATGPT-3: perspektivy ispol'zovaniya v obuchenii inostrannomu yazyku [CHATGPT-3: Prospects for Use in Foreign Language Teaching]. *MNKO*, 3 (100), pp. 33-36.
8. Pan Z., Mishra P. (2021) Hardware Acceleration of Explainable AI // In book: Explainable AI for Cybersecurity, 7, pp. 199-220.
9. Sobolev I.V., Potapova E.A. (2022) Problema vozmozhnosti iskusstvennogo intellekta s tochki zreniya psikhologicheskoi nauki [The Problem of the Possibility of Artificial Intelligence from the Point of View of Psychological Science]. *Kollektsiya gumanitarnykh issledovanii* [Collection of Humanitarian Research], 2 (31), pp. 42-58.
10. Tokhirzhonova M.R. (2023) Rol' iskusstvennogo intellekta v pedagogike, uluchshenie opyta obucheniya s pomoshch'yu intellektual'nykh tekhnologii [The Role of Artificial Intelligence in Pedagogy, Improving the Learning Experience with the Help of Intelligent Technologies]. *Teoriya i praktika sovremennoi nauki* [Theory and Practice of Modern Science], 7 (97), pp. 28-49.
11. Tripathy V.K., Seetha H. (2022) Explainable, Interpretable, and Transparent AI Systems. *University of Luxembourg*, 7, pp. 112-145.
12. Vislova A.V. (2019) Potentsial psikhologii intellekta v kontekste modelirovaniya iskusstvennogo intellekta [Potential of the Psychology of Intelligence in the Context of Artificial Intelligence Modeling]. *Izvestiya KBNTs RAN* [Bulletin of the Kabardino-Balkarian Scientific Center of the Russian Academy of Sciences], 6 (92), pp. 32-47.
13. Vovk E.V., Suprun A.A. (2022) Iskusstvennyi intellekt i tsifrovaya pedagogika kak trend sovremennoi obrazovatel'noi sredy vysshikh uchebnykh zavedenii [Artificial Intelligence and Digital Pedagogy as a Trend in the Modern Educational Environment of Higher Education Institutions]. *Problemy sovremennogo pedagogicheskogo obrazovaniya* [Problems of Modern Pedagogical Education], 77-2, pp. 78-99.
14. Yohei Okada, Yilin Ning, Marcus Eng, Hock Ong. (2023) Explainable AI in Emergency medicine: An overview. *November Clinical and Experimental Emergency Medicine*, 10 (4), pp. 67-89.