УДК 316.775

ЧЕЛОВЕК ИЛИ AI: К ВОПРОСУ ОБ АВТОРСТВЕ

А. С. Никитина

Статья посвящена проблеме определения текстов, сгенерированных искусственным интеллектом (AI). Рассматриваются разные способы обнаружения сгенерированных текстов, а именно машинные детекторы и мышление человека. Сделан вывод о том, что детекторы сильно ограничены и не могут быть использованы для проверки академической честности. В связи с этим особое внимание уделяется работам исследователей, которые предлагают сотрудничество людей и машин. Также в статье приведены примеры последних исследований и методов, с помощью которых учёные выясняют, способны ли люди определять авторство текстов. В заключение предлагается использовать как качественные, так и количественные подходы социологического исследования для поиска более надёжного определения авторства в условиях быстро развивающихся технологий.

Ключевые слова: искусственный интеллект; AI-детектор; тест Тьюринга; академическая честность; исследовательские методы; анализ текста.

Ещё в 1950 году британский математик Алан Тьюринг задался вопросом: «Могут ли машины мыслить?» [1]. Он предположил, что «мышление» компьютера может проявляться подобно нашему, а именно вести диалог с человеком на естественном языке. Тогда Тьюринг разработал тест, который считался пройденным, если в процессе общения с машиной хотя бы двое из трёх собеседников принимали искусственный интеллект (AI) за живого человека [2]. Год появления этого теста показывает, что учёных давно интересует, способны ли люди определять, кто написал текст, который они читают, – искусственный интеллект или человек. Сегодня тест Тьюринга не может проверить способности машин в полной мере, потому что AI-технологии ушли далеко вперёд. Однако учёные продолжают применять принципы этого давно известного теста на некоторых этапах своих исследований либо модифицируют его, приспосабливая к новым реалиям [3; 4].

В этой статье мы посмотрим, какие подходы и методы применяются для исследования данного вопроса сегодня. Хотя в наши задачи не входит подробный рассказ о результатах, к которым пришёл тот или иной исследователь, мы отметим самые интересные и значимые из них, чтобы проиллюстрировать специфику рассматриваемых методов и заложить хороший теоретический фундамент для наших дальнейших исследований в этой области.

Способы определения сгенерированных текстов можно разделить на две большие группы: машины и люди. С одной стороны, ведётся разработка и применение различных детекторов искусственного интеллекта. С другой стороны, учёные исследуют способность людей определять авторство сгенерированных текстов.

Детекторы могут использовать для обнаружения искусственно созданных текстов набор определённых параметров, например, статистический анализ текста, лингвистический анализ, проверку фактов или алгоритмы классификации, которые позволяют обученным моделям приписывать данным категорию или класс на основе признаков текста [5]. Однако специалисты сомневаются в том, что детекторы способны безошибочно справляться с поставленными задачами. Такое скептическое отношение подкрепляется рядом исследований, некоторые из которых будут рассмотрены ниже.

Учёные из Европейской сети академической честности (ENAI) протестировали 14 инструментов обнаружения сгенерированного

Никитина Анна Сергеевна (nikitinanna.s@yandex.ru),
магистрант I курса социологического факультета Самарского университета,
443086, Россия, г. Самара, Московское шоссе, 34.

текста (такие как Check For AI, Compilatio, DetectGPT, GPT Zero, OpenAI Text Classifier, Turnitin и др.). Почти все эти системы определяли тексты с точностью ниже 80 % и только 5 из них показали результат выше 70 %. Кроме того, было установлено, что эффективность подобных методов обнаружения ухудшается, когда по отношению к проверяемым текстам предварительно применяются методы обхода (например, перефразирование или машинный перевод). Около 50 % таких видоизменённых текстов были ошибочно приписаны человеку [6].

Нэш Андерсон и др. [7] подчёркивают, что возможность обхода детекторов вызывает серьёзное беспокойство. Они исследовали точность GPT-2 Output Detector. Для эксперимента сгенерировали две научные статьи в ChatGPT, которые затем были проверены в программе. Программа сделала вывод, что первая статья написана человеком с вероятностью 0,02 %, а вторая статья – с вероятностью 61,96 %. Однако затем исследователи изменили тексты с помощью перефразирования, и результаты значительно изменились. GPT-2 Output Detector посчитал, что статьи написаны человеком с вероятностью 61,96 % и 99,98 % соответственно.

В целом, проанализировав последние исследования в области машинного обнаружения сгенерированных текстов, мы выделили следующие основные проблемы:

• детекторы сгенерированного контента выдают ложноположительные и ложноотрицательные результаты, то есть принимают сгенерированные тексты за написанные человеком и наоборот;

• сгенерированные тексты, к которым были применены методы обхода, например, перефразирование, с большей долей вероятности определяются как написанные человеком;

• кроме того, детекторы развиваются гораздо медленнее, чем большие языковые модели, а это приводит к тому, что тексты, сгенерированные в более ранних моделях (например, GPT-2), труднее обнаружить, чем те, которые созданы в более поздних версиях (например, GPT-4).

Ряд исследователей изучает проблему определения авторства текстов в сфере образования. Эта область является одной из наиболее уязвимых, ведь у каждого студента появляется возможность выдавать искусственно созданные тексты за свои, а это расшатывает этические нормы и влияет на качество получаемых знаний. Но уличить такую нечестность не всегда возможно – исследователи утверждают, что современные детекторы сгенерированного контента сильно ограничены и не пригодны для использования в качестве доказательства академической непорядочности. Поэтому преподавателям предлагают сосредоточиться не на стратегиях обнаружения, а на самом процессе развития навыков студентов [6].

Однако необходимо помнить, что проблема определения авторства сгенерированных текстов простирается гораздо дальше. Это не только вопрос академической честности. Даже тот студент, который не использует искусственный интеллект для написания выпускной квалификационной работы, может столкнуться со сгенерированным текстом на этапе поиска информации для нее, посчитать, что он написан человеком, и принять на веру все его недостоверные факты. Так, исследователи из Корнельского университета выяснили, что люди находили достоверными фейковые новостные статьи, сгенерированные GPT-2, примерно в 66 % случаев [8]. Это значит, что человеку необходимо уметь определять, кем был написан тот или иной текст, чтобы обезопасить себя при потреблении контента.

Так как машинные методы обнаружения не способны помочь с решением этой проблемы, вопрос остаётся открытым, а мы продолжаем искать новые ответы.

Некоторые исследователи предлагают рассмотреть сотрудничество людей и машин. Дафни Ипполито и др. [9] сделали любопытное наблюдение, что люди и детекторы принимают решения, основываясь на разных качествах текста: люди легче замечают семантические ошибки, а детекторы лучше улавливают статистические искажения, связанные с выбором наиболее вероятных слов. Мэтью Гро и др. [10] предложили людям распознать сгенерированные изображения и обнаружили, что человек может улучшить свою способность определять сгенерированный контент, тренируясь и получая обратную связь. После получения обратной связи по 10 парам изображений в среднем за 1 минуту 14 секунд способность участника распознавать сгенерированный контент улучшилась на 10 %. Эван Кротерс, Натали Япкович и

Херна Виктор [11] считают, что для снижения вероятности ложноположительных результатов и других этических рисков необходимо привлекать для выявления сгенерированных текстов людей-аналитиков. Ахмед М. Эльхатат, Халед Эльсаид и Саид Альмир делают вывод, что не следует использовать инструменты обнаружения в качестве единственного определителя академической честности. Но вместо этого нужно применять более целостный подход, который включает в себя учёт контекста и ручную проверку [12]. Машинам необходим контроль со стороны человека, который обладает нужными для такого контроля компетенциями.

Чтобы выяснить, способны ли люди определять авторство текстов, исследователи используют различные методы. Например, Штеффен Хербольд и др. [13] подготовили набор эссе на английском языке, в который входили студенческие работы и сгенерированные тексты, и разработали ***анкету.*** В качестве респондентов выступили эксперты-преподаватели. Они должны были заполнить бланк, состоящий из трёх частей: 1) нужно было самостоятельно оценить свой уровень владения английским языком; 2) оценить эссе по критериям (тема и законченность, логика и композиция, выразительность и полнота, владение языком, сложность, словарный запас и связность текста, языковые конструкции) по семибалльной шкале; 3) оценить по пятибалльной шкале свою уверенность в выставленных рейтингах. В результате качество эссе, сгенерированных ChatGPT, было оценено выше, чем качество работ, написанных человеком. Студенты получили наихудшие оценки, ChatGPT-3 средние, а ChatGPT-4 преподаватели поставили самую высокую оценку.

Стоит отметить, что в подобных исследованиях существует чёткое разграничение ***экспертов*** и ***непрофессионалов.*** Например, в упомянутом выше исследовании так и обозначено: «Сосредоточение внимания только на этих экспертах позволяет нам получить значимые результаты, поскольку эти участники имеют большой опыт в оценке письменных работ учащихся». В противовес этому внимание прочих исследователей сосредоточено на другом объекте – обычных людях, которые не получали филологического образования, не являются преподавателями и не работают регулярно с текстами. Вероятно, такой подход выбирают, когда хотят узнать, как на сгенерированные тексты реагирует широкая, неподготовленная аудитория. Так, Элизабет Кларк, Тал Август, София Серрано и другие учёные из школы компьютерных наук и Института Аллена [14] выяснили, что люди, чья профессиональная деятельность не связана с текстами, смогли определить контент, сгенерированный GPT-2, с точностью 57,9 %, а тексты, сгенерированные GPT-3, – с точностью всего 49,9 %.

Было проведено исследование [4], в котором участвовали эксперты (опытные пользователи ChatGPT) и непрофессионалы (никогда не слышали о ChatGPT). Как объясняют исследователи, такая выборка была связана с тем, что люди, знакомые с ChatGPT, возможно, запомнили некоторые паттерны его поведения и это помогает им легко определять, кто автор текста. Сравнивая результаты, исследователи пришли к выводу, что точность экспертов намного выше, чем у непрофессионалов.

Особенно интересны случаи, когда респондентов или информантов перед проведением исследования предварительно обучали. Такая подготовка проводилась с разными целями: либо чтобы получить экспертов, которые могут максимально точно оценить качество текста [13]; либо для того, чтобы проверить, каким образом такое обучение влияет на способность человека определять авторство текстов. Например, учёные из Пенсильванского университета [15] отобрали для своего исследования аспирантов и студентов старших курсов из двух секций курса по искусственному интеллекту. Само исследование проходило в формате ***обучающей веб-игры.*** Игра начиналась с образца текста, написанного человеком. Затем игроку показывали по одному предложению. Он должен был решить, сгенерировано это предложение или написано человеком, и объяснить причину своего выбора. Благодаря этому исследованию было обнаружено, что при определённых условиях и с течением времени люди могут демонстрировать улучшение результатов. Игроки, которые могли получать баллы за достижения в игре, читать дополнительные инструкции и руководство с советами, лучше

справлялись с определением текстов. Исследование показало, что обнаружение сгенерированных текстов – это навык, который можно развивать и которому можно обучать.

А. В. Громова, С. Н. Логинова и Е. С. Китаева [16] провели ***эксперимент,*** чтобы диагностировать ошибки, допускаемые системами искусственного интеллекта. Для эксперимента были привлечены чат-боты «pBot», «Алиса», «Маруся» и «Cleverbot», а также специалисты-филологи. На протяжении месяца группа экспертов вела переписку с чат-ботами, а по завершении исследования выделила признаки, характерные для сгенерированных текстов. Например, на уровне синтаксиса в таких текстах наблюдались простые нераспространенные или малораспространённые конструкции и отсутствовали высказывания с прямой речью или парцеллирование конструкций. Также были выявлены отличия на лексическом, морфологическом и графическом уровнях, подробный анализ которых выходит за рамки нашего обзора. С помощью данного эксперимента исследователи, с одной стороны, выявили слабости генеративных систем, а с другой стороны, показали, что люди с экспертным знанием могут охарактеризовать особенности сгенерированных текстов, используя определённые критерии.

Не менее важным методом в исследовании данной проблематики является ***контент-анализ,*** который также применяется для выявления отличительных признаков сгенерированных текстов и демонстрирует способность человека (в данном случае исследователя) успешно заниматься подобным определением. Т. А. Безуглый и М. Е. Ершова [17] провели сравнительный анализ двух текстов: один был сгенерирован в ChatGPT с помощью запроса «Научный текст о том, что такое диабет», другой – взят из статьи с сайта Всемирной организации здравоохранения. В результате сравнения были обнаружены следующие особенности сгенерированного текста: нарушение лексической сочетаемости, обилие однотипных синтаксических конструкций, повтор предлогов, тавтология, большое количество деепричастных и причастных оборотов и придаточных частей, неправильная постановка запятых.

Джон Хьюстон [18] провёл в рамках своего исследования ***фокус-группу.*** В ней участвовало 11 добровольцев, которых попросили прочитать два образца небольшого художественного текста объёмом около 1300 знаков. Первый образец H1 был полностью написан человеком. Некоторые фрагменты второго образца G3 были дописаны в ChatGPT, который по своему усмотрению изменил развитие истории. Участникам фокус-группы нужно было ответить на вопросы: 1) какой образец лучше погружает вас в обстановку рассказа? 2) какой образец лучше всего связывает вас с главным героем? 3) какой образец имеет более хороший ритм? 4) есть ли у вас какие-либо наблюдения? 54,5 % читателей посчитали, что сгенерированный текст погрузил их в историю лучше, чем образец, написанный человеком, а 72,7 % сказали, что, читая текст, сгенерированный ИИ, они почувствовали более сильную связь с главным героем. Хьюстон также провёл ***экспертное интервью.*** Если до этого мы говорили об экспертах как о тех, кто выделяет конкретные критерии для различения текстов, то в данном исследовании эксперты выступали как специалисты по искусственному интеллекту. Они помогали понять, как компьютеры обучаются с помощью алгоритмов, говорили об этической стороне использования ИИ, об эффективных способах применения ChatGPT или о чат-ботах и их отношении к языку. Такой метод помогает исследователю заполнить пробелы в новой для него теме и сделать свое исследование междисциплинарным.

Мы видим, что для исследования сгенерированных текстов нередко применяется ***смешанная стратегия,*** которая включает как качественный, так и количественный подходы. Учёные из Вьетнама и Сингапура [19] пытались выяснить, могут ли преподаватели корректно оценивать работы студентов, содержащие сгенерированные тексты. Преподаватели должны были проверить студенческие работы и указать на те, которые были созданы GPT-4. Помимо этого, все тексты проверялись в программе обнаружения сгенерированного контента Turnitin AI. Результаты показали, что из 22 сгенерированных текстов преподаватели смогли определить 12. Это чуть более половины (54,5 %). Turnitin AI обнаружила 54,8 %, а 91 % выделила как содержащий некоторое количество сгенерированного контента. Средний балл оценок настоящих работ студентов соста-

вил 54,4, а для текстов, созданных ИИ, – 52,3. Это показало, что обнаружение сгенерированного текста несущественно повлияло на выставленные оценки. На следующем этапе исследования использовалось качественное интервью, когда преподаватели комментировали свой выбор. Сгенерированные тексты воспринимались по-разному. Некоторые эксперты высоко оценили работы: «Много идей, которые, возможно, стоит развить»; «[Это] хорошее исследование и ясное мышление». Другие отметили, что материалу не хватает глубины и целенаправленности. Информанты отмечали, что у сгенерированных текстов «запутанный» стиль, отсутствует «индивидуальность и визуализация». Работы критиковали также за «очень длинное введение» и «проблемы с источниками».

Итак, если раньше проблема авторства текстов решалась довольно простым тестом Тьюринга, то современные реалии побуждают исследователей постоянно искать новые и более сложные исследовательские методы, а также комбинировать их. Как мы увидели, количественные исследования могут быть довольно перспективными, а использование смешанного подхода с элементами качественной и количественной стратегии помогает получить более исчерпывающие результаты. Однако, на наш взгляд, качественная стратегия достойна большего внимания и может выступать не только как вспомогательная, но и как основная. Именно качественные исследования, которые предполагают более развёрнутые комментарии информантов, помогут заглянуть глубже и увидеть, как люди размышляют, какими критериями руководствуются, когда определяют, является текст сгенерированным или нет.

Литература

1. Фишман Р., Кузнецов Д. Картотека: что такое тест Тьюринга? 2003 [Электронный ресурс]. URL: https://iq.hse.ru/news/874939177.html (дата обращения: 08.05.2024).

2. Теплыгина И. М. Естественность языковых данных // Язык и культура в глобальном мире. 2024. № 2. С. 513–517.

3. Естественность языковых данных / А. Ю. Красняров, М. А. Аргузова, Ж. А. Хужамурадов [и др.] // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 6: Языкознание. 2022. С. 41–49.

4. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection / B. Guo, X. Zhang, Z. Wang [et al.] // CoRR. abs/2301.07597. 2023. doi.org/10.48550/arXiv.2301.07597.

5. Tang R., Chuang Y.-N., Hu X. The Science of Detecting LLM-Generated Text [Electronic resource]. URL: https://cacm.acm.org/research/the-science-of-detecting-llm-generated-text/#B28 (accessed: 07.05.2024).

6. Testing of detection tools for AI-generated text / D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba [et al.] // International Journal for Educational Integrity. 2023. Vol. 19. № 26. doi.org/10.1007/s40979-023-00146-z.

7. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation / N. Anderson, D. L. Belavy, S. M. Perle [et al.] // BMJ Open Sport Exerc Med. 2023. № 9 (1). doi:10.1136/bmjsem-2023-001568.

8. Heikkiläarchive M. How to spot AI-generated text [Electronic resource]. URL: https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/ (accessed: 07.05.2024).

9. Automatic Detection of Generated Text is Easiest when Humans are Fooled / D. Ippolito, D. Duckworth, C. Callison-Burch [et al.] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 1808–1822.

10. Human Detection of Machine Manipulated Media / M. Groh, Z. Epstein, N. Obradovich [et al.] // Communications of the Acm. 2021. Vol. 64. № 10. P. 40–47.

11. Crothers E. N., Japkowicz N., Viktor H. L. Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods // IEEE Access. 2023. Vol. 11. P. 70977–71002.

12. Elkhatat A. M., Elsaid K., Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text // International Journal for Educational Integrity. 2023. Vol. 19. № 17. doi.org/10.1007/s40979-023-00140-5.

13. A large-scale comparison of human-written versus ChatGPT-generated essays / S. Herbold, A. Hautli-Janisz, U. Heuer [et al.] //

Scientific Reports. 2023. Vol. 13. № 18617. doi.org/10.1038/s41598-023-45644-9.

14. All That's `Human' Is Not Gold: Evaluating Human Evaluation of Generated Text / E. Clark, T. August, S. Serrano [et al.] // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. Vol. 1. P. 7282–7296.

15. Real or Fake Text?: Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text / L. Dugan, D. Ippolito, A. Kirubarajan [et al.] // CoRR. abs/2212.12672. 2022. https://doi.org/10.48550/arXiv.2212.12672

16. Громова А. В., Логинова С. Н., Китаева Е. С. К вопросу о выявлении признаков искусственной генерации текстов при проведении авторoведческого исследования // Фундаментальная лингвистика и проблемы судебной экспертизы: социальные сети как объект научного и экспертного анализа. 2021. С. 11–16.

17. Безуглый Т. А., Ершова М. Е. Использование текстовых нейросетей и искусственного интеллекта в учебных работах студентов // Проблемы современного образования. 2023. Т. 5. С. 206–2016.

18. Huston J. Artificial Intelligence as a Content Creator in the Publishing Industry // Book Publishing Final Research Paper. 2022. № 65 [Electronic resource]. URL: https://archives.pdx.edu/ds/psu/37861 (accessed: 07.05.2024).

19. Game of Tones: Faculty Detection of GPT-4 Generated Content in University Assessments / M. Perkins, J. Roe, D. Postma [et al.] // CoRR. abs/2305.18081. 2022. doi.org/10.48550/arXiv.2305.18081.

HUMAN OR AI: ON THE QUESTION OF AUTHORSHIP

A. S. Nikitina

The article is dedicated to the problem of identifying texts generated by artificial intelligence (AI). Various methods of detecting generated texts are considered, namely machine detectors and human thinking. The conclusion is drawn that detectors are severely limited and cannot be used to verify academic integrity. Therefore, special attention is paid to the works of researchers who propose collaboration between humans and machines. Additionally, the article provides examples of recent research and methods through which scientists determine whether humans can identify the authorship of texts. In conclusion, it is suggested to use both qualitative and quantitative approaches of sociological research to seek a more reliable detection of authorship in the context of rapidly evolving technologies.

Key words: artificial intelligence; AI-detector; Turing test; academic integrity; research methods; text analysis.

Статья поступила в редакцию 28.05.2024 г.

Nikitina Anna Sergeevna (nikitinanna.s@yandex.ru),
1st year master student of the Faculty of Sociology of Samara University,
443086, Russia, Samara, Moskovskoye shosse, 34.