Секция «32.8 Социология коммуникаций»

Кризис критериев аутентичности в эпоху ИИ-генерированного контента: когнитивные стратегии и лингвистические маркеры верификации в академической среде

Научный руководитель – Юрасова Мария Владимировна

Скапцова Софья Алексеевна
Студент (бакалавр)
Московский государственный университет имени М.В.Ломоносова, Социологический факультет, Кафедра социологии организаций и менеджмента, Москва, Россия
E-mail: sskaptsova@mail.ru

Генеративные языковые модели последних поколений производят текст, который реципиент квалифицирует как человеческий с вероятностью, статистически неотличимой от случайного угадывания. В серии из шести экспериментов (N = 4600) была зафиксирована точность атрибуции авторства на уровне 50–52 %, причём ни денежное стимулирование, ни самоотчётная экспертиза не повышали результат [1]. Сопоставимые данные получены в смежных контекстах: при слепой оценке перефразированных, обобщённых и аргументативных текстов участники краудсорсинговых платформ ошибались в 43–50% случаев, а при манипуляции метками «Сделано человеком» / «Сгенерировано ИИ» предпочтение сдвигалось на 30 % в сторону якобы человеческого текста вне зависимости от реального источника [2]. В исследовании, посвящённом сравнению школьных преподавателей и учащихся, показано, что преподаватели корректно идентифицировали эссе ChatGPT лишь в 70 % парных сравнений, а ученики — в 62 %, при этом самооценка уверенности не коррелировала с точностью [3]. Эти результаты ставят под вопрос не столько возможности детекции, сколько природу когнитивных оснований, на которые опирается суждение об «авторском» или «машинном» происхождении текста.

Центральным механизмом ошибок выступает набор интуитивных, но дисфункциональных эвристик. Качественное кодирование самоотчётов в сочетании с регрессионным моделированием позволило реконструировать перечень подсказок, которым следуют респонденты: местоимения первого лица, контракции, упоминание семьи и неформальный тон устойчиво ассоциируются с «человечностью», тогда как грамматические аномалии, редкие биграммы и длинные лексемы маркируются как признак алгоритма [1]. На практике эти ассоциации оказываются инвертированы: грамматические дефекты на 15 % чаще встречаются в текстах, написанных людьми, а контракции — на 13 % чаще в машинных [1]. Единственные функциональные индикаторы — бессмысленность фрагмента и повторяемость — обеспечивают около 59 % точности, однако их эффект нивелируется дисфункциональными подсказками [1]. Дополнительно показано, что флюидный интеллект является значимым предиктором различительной способности, тогда как эмпатия и исполнительные функции статистически значимого вклада не вносят; интенсивное использование смартфонов и социальных сетей, вопреки ожиданиям, повышает вероятность ошибочной атрибуции ИИ-текста как человеческого [4]. Указанный паттерн типологически близок к «теории истины по умолчанию» (truth-default theory), согласно которому коммуникативный партнёр изначально воспринимается как правдивый, и аналогичный порог доверия переносится на оценку авторства.

Помимо интуитивных ярлыков действует и более осознанная система стереотипов о том, «как должен выглядеть машинный текст». Экспериментально показано на выборке 254 чешских носителей, что участники без обратной связи опираются на ожидание статичности, подготовленности, высокой когезии и политематичности ИИ-генерированного текста — характеристик, соответствующих формально-научному регистру [5]. Когда текст

отвечал этим предубеждениям, его верно атрибутировали как машинный; когда нет — ошибки возрастали, причём максимальная самоуверенность сопровождала наименьшую точность. Группа, получавшая пошаговую обратную связь, достигала 65,1 % точности против 55,4 % у контрольной, что указывает на обучаемость навыка и корректируемость ложных стереотипов [5]. Введено также понятие «импостор-предвзятости» — систематической склонности ставить под сомнение подлинность любого цифрового контента на фоне осведомлённости о генеративных возможностях ИИ, — которая, в отличие от здорового скептицизма, действует индискриминантно и ведёт к ложноположительным маркировкам аутентичного материала [6; 7]. Одновременно фиксируется «автоматизационная предвзятость»: участники склонны менять первоначальное суждение в пользу алгоритмической подсказки, даже когда она ошибочна, причём обе предвзятости действуют относительно независимо [7].

Дополнительный пласт эвристик связан с метками авторства: показано, что один лишь ярлык «ИИ-автор» при идентичном тексте снижает воспринимаемую достоверность сообщения ($d = 0{,}36$) и источника ($d = 0{,}24$), а также антропоморфизм ($d = 0{,}67$) и воспринимаемый интеллект автора ($d = 0{,}41$) [8]. Иначе говоря, атрибуция авторства запускает каскад оценочных суждений, не связанных с лингвистическим содержанием, — своеобразную «эвристику нечеловечности», при которой знание об алгоритмическом происхождении автоматически активирует негативную схему восприятия.

Существенный пробел в литературе проявляется при попытке перенести эти выводы на академическое сообщество. Подавляющее большинство экспериментов проведено на краудсорсинговых выборках (Amazon Mechanical Turk, Prolific), не дифференцированных по профессиональной принадлежности [1; 2]. Между тем академическая аудитория обладает специфическими ресурсами — навыком нормированного научного письма, чувствительностью к конвенциям цитирования и аргументации, институциональной мотивацией к верификации. Согласно имеющимся данным, преподаватели лишь незначительно превосходят студентов в детекции, а их субъективная уверенность систематически расходится с реальной точностью [3]. Остаётся открытым вопрос: формирует ли академическая среда качественно иные эвристики — например, реагирование на избыточную гладкость изложения, отсутствие авторского голоса, шаблонность логических связок — или воспроизводит те же дисфункциональные паттерны, но с более высоким порогом уверенности. Выявление этих специфических механизмов может стать основой для проектирования образовательных программ по ИИ-грамотности, ориентированных не на алгоритмические детекторы, а на развитие рефлексивного критического чтения.

Источники и литература

1) 1. Human heuristics for AI-generated language are flawed [Электронный ресурс] // Proceedings of the National Academy of Sciences. 2023. Vol. 120, № 11. Art. e2208839120. URL: https://www.researchgate.net/publication/369065224_Human_heuristics_for_AI-generated_language_are_flawed (дата обращения: 28.02.2026).
2) 2. Human Bias in the Face of AI: Examining Human Judgment Against Text Labeled as AI Generated [Электронный ресурс] // arXiv. 2024. URL: https://arxiv.org/abs/2410.03723 (дата обращения: 28.02.2026).
3) 3. Testing the Ability of Teachers and Students to Differentiate between Essays Generated by ChatGPT and High School Students [Электронный ресурс] // Human Behavior and Emerging Technologies. 2023. Art. 1923981. DOI: 10.1155/2023/1923981. URL: https://www.researchgate.net/publication/371894228_Testing_the_Ability_of_Teachers_and_Students_to_Differentiate_between_Essays_Generated_by_ChatGPT_and_High_School_Students (дата обращения: 28.02.2026).

4) 4. Human intelligence can safeguard against artificial intelligence: individual differences in the discernment of human from AI texts [Электронный ресурс] // Scienti fic Reports. 2024. Vol. 14. Art. 25989. DOI: 10.1038/s41598-024-76218-y. URL: https://www.researchgate.net/publication/385355032_Human_intelligence_can_safeguard_against_artificial_intelligence_individual_differences_in_the_discernment_of_human_from_AI_texts (дата обращения: 28.02.2026).
5) 5. Learning to detect AI texts and learning the limits [Электронный ресурс] // PLoS ONE. 2025. Vol. 20, № 10. Art. e0333007. DOI: 10.1371/journal.pone.0333007. URL: https://www.researchgate.net/publication/396513939_Learning_to_detect_AI_texts_and_learning_the_limits (дата обращения: 28.02.2026).
6) 6. GenAI Mirage: The Impostor Bias and the Deepfake Detection Challenge in the Era of Artificial Illusions [Электронный ресурс] // arXiv. 2023. arXiv:2312.16220. URL: https://arxiv.org/abs/2312.16220 (дата обращения: 28.02.2026).
7) 7. A (Mid)journey Through Reality: Assessing Accuracy, Impostor Bias, and Automation Bias in Human Detection of AI-Generated Images [Электронный ресурс] // Human Behavior and Emerging Technologies. 2025. DOI: 10.1155/hbe2/9977058. URL: https://www.researchgate.net/publication/395431908_A_Midjourney_Through_Reality_Assessing_Accuracy_Impostor_Bias_and_Automation_Bias_in_Human_Detection_of_AI-Generated_Images (дата обращения: 28.02.2026).
8) 8. The Effects of Assumed AI vs. Human Authorship on the Perception of a GPT-Generated Text [Электронный ресурс] // Journalism and Media. 2024. Vol. 5, № 3. P. 1085–1097. DOI: 10.3390/journalmedia5030069. URL: https://www.mdpi.com/2673-5172/5/3/69 (дата обращения: 28.02.2026).