УДК 004.032.26:81
DOI: 10.28995/2686-7249-2025-3-51-58

Проблема идентификации манипуляции в диктуме нейросетей

Елена Л. Боярская
Балтийский федеральный университет им. Иммануила Канта, Калининград, Россия, EBoyarskaya@kantiana.ru

Эльвира И. Бархатова
Балтийский федеральный университет им. Иммануила Канта, Калининград, Россия, elvira.barkhatova@gmail.com

Аннотация. Научное исследование имеет целью разработку методов идентификации манипулятивного воздействия, реализуемого новым типом агенса, имеющего техногенную природу – нейросетью. В работе осуществляется анализ манипулятивных стратегий и приемов, генерируемых большими языковыми моделями ChatGPT и Gemini (Bard), с учетом их влияния на коллективного и индивидуального пациенса и формирования мнения о событиях. Особое внимание уделяется выявлению функциональных, смысловых и когнитивных аспектов манипулятивных актов в ответах нейронных сетей на запросы пользователей. Исследование предлагает подходы для возможной разработки инструментов эффективной оценки и распознавания манипулятивного воздействия в диктуме нейросетей, а также выдвигает гипотезы относительно возможности выделения дискурсивных маркеров, которые могут указывать на наличие манипулятивного воздействия в процессе использования нейросетей. Делается вывод о комплексном и амбивалентном характере манипулятивных маркеров в электронном диктуме, что значительно усложняет их типологизацию. Полученные результаты исследования могут быть полезны для разработки алгоритмов, направленных на повышение надежности, эффективности и безопасности использования нейронных сетей в различных областях применения, включая обработку естественного языка, анализ данных и другие.

Ключевые слова: манипуляция, событийный фрейм, нейросеть, цифровой агенс

Для цитирования: *Боярская Е.Л., Бархатова Э.И.* Проблема идентификации манипуляции в диктуме нейросетей // Вестник РГГУ. Серия «Литературоведение. Языкознание. Культурология». 2025. № 3. С. 51–58. DOI: 10.28995/2686-7249-2025-3-51-58

Identification of manipulation in the dictum of neural networks

Elena L. Boyarskaya
Immanuel Kant Baltic Federal University, Kaliningrad, Russia,
EBoyarskaya@kantiana.ru

Elvira I. Barkhatova
Immanuel Kant Baltic Federal University, Kaliningrad, Russia,
elvira.barkhatova@gmail.com

Abstract. This research aims to develop methods the manipulative acts identification by a new type of agent – neural networks. The authors analyse manipulative strategies and techniques generated by ChatGPT and Gemini (Bard) language models, taking into account their influences on the collective and individual paciens and the formation of opinions about events. Special attention is paid to the identification of functional, semantic and cognitive aspects of manipulative acts in the responses of neural networks. The study proposes a methodology to effectively assess and identify manipulative messages in the dictum of neural networks and hypothesises discourse markers that may indicate the presence of manipulative influence in the use of neural networks. It is concluded that the complex and ambivalent nature of manipulative markers in electronic dictum makes their typologization much more difficult. The results obtained may be used for the development of algorithms aimed at improving the reliability and efficiency of neural networks in various application areas, including natural language processing, data analysis, and others.

Keywords: manipulation, event frame, neural network, digital agent

For citation: Boyarskaya, E.L. and Barkhatova, E.I. (2025), "Identification of manipulation in the dictum of neural networks", *RSUH/RGGU Bulletin. "Literary studies. Linguistics. Culturology" Series*, no. 3, pp. 51–58, DOI: 10.28995/2686-7249-2025-3-51-58

Рынок технологий искусственного интеллекта демонстрирует внушительные темпы роста: в 2023 г. его объем составил 407 млрд долларов, при этом годовые темпы роста, начиная с 2023 г., являются поистине космическими – 37,3%. По прогнозам ученых, к 2026 г. более 90% интернет-контента будет генерироваться технологиями искусственного интеллекта (ИИ). В 2024 г. инвестиции в обучение одной большой языковой модели составили более 1 млрд долларов, а число пользователей по всему миру превысило 314 млн сетей, включая создание, анализ и обработку разного рода данных, перевод и редактирование текстов, создание изображений и многое другое. Технологии ИИ постепенно становятся мультимодальными и, следовательно, способными решать множество задач. Однако вместе с ростом потенциала и сфер использования ИИ возрастает обеспокоенность последствиями использования нейросетей. Более 75% пользователей выражают серьезные опасения по поводу распространения недостоверной информации и скрытого манипулирования нейросетью[1]. Этот факт говорит о необходимости раз-

[1] Облачный рынок обработки естественного языка. URL: https://www.gminsig hts.com/ru/industry-analysis/cloud-natural-language-processing-nlp-market (дата обращения 1 декабря 2024).

работки этических и регуляторных рамок для внедрения ИИ в целях обеспечения безопасности использования цифровых технологий[2].

В настоящее время проблема идентификации манипулятивного воздействия, в том числе в диктуме нейросетей, стоит достаточно остро [Заботкина, Боярская 2023; Литяйкина 2021]. Собственно термин *нейросеть* является широким и охватывает целый спектр моделей машинного обучения, использующих искусственные нейронные сети для выполнения различных задач. В данном исследовании речь пойдет об анализе манипулятивного потенциала больших языковых моделей, то есть конкретного типа нейросетей, появившегося на основе обработки данных естественных языков. Они обучаются на значительных объемах данных; они способны «понимать» и генерировать тексты, выполнять задачи перевода на значительное количество естественных языков, отвечать на запросы, анализировать тональность и стилевые особенности текста и многое другое. Примерами таких моделей являются GPT (Generative Pre-trained Transformer) от OpenAI и BERT (Bidirectional Encoder Representations from Transformers) от Google. Проблема исследования манипулятивного потенциала больших языковых моделей заключается в определении и понимании тех способов, с помощью данные технологии могут оказывать влияние на восприятие и интерпретацию информации[3] и, следовательно, детерминировать поведение. Существует значительный риск того, что нейронные сети могут быть обучены на специально подобранных текстовых данных и использованы для манипуляции информацией с целью воздействия на мнения и убеждения людей. Это может происходить посредством создания и распространения дезинформации[4], формирования и усиления стереотипов и предвзятости, а также оказания воздействия на систему моральных ценностей и убеждений [Ackerman 2015].

Настоящее исследование имеет целью разработку методов идентификации манипулятивного воздействия новым типом агенса – нейросетью. С этой целью использовался метод построения событийного фрейма «Манипуляция» и определения его основных слотов, которые включают следующие: *агенс* – инициатор манипулятивного диктума; *отношение агенса* к манипулятивному диктуму; *пациенс* (экспериенсер) – адресат манипулятивного диктума; *мотив* – события, предшествующие инициации манипулятивного диктума; *цель* – цель, преследуемая агенсом, формулирующим манипулятивный диктум; *диктум* – сообщение, которое не соответствует объективной реальности; *характер* манипулятивного диктума – характеристика, установление связи между фактами; *манера* выражения диктума – эксплицитная, имплицитная; *восприятие* манипулятивного диктума пациенсом – позитивное, негативное, нейтральное, иное; *пост-эффекты* манипулятивного диктума – события, следующие за диктумом; *контрсагенс<ы>* – агенс<ы>, препятствующий или деформирующий манипулятивный диктум [Боярская, Шевченко, Томашевская 2023; Заботкина, Боярская 2023].

---

[2] Artificial intelligence. URL: https://plato.stanford.edu/entries/artificial-intelli gence/ (дата обращения 1 декабря 2024).

[3] Social media terms of use. URL: https://www.sciencealert.com/social-media-terms-of-use (дата обращения 1 декабря 2024).

[4] A.I.'s use in elections sets off a scramble for guardrails. URL: https://www.nytimes.com/2023/06/25/technology/ai-elections-disinformation-guardrails.html (дата обращения 1 декабря 2024).

Появление и широкое использование нейросетей коренным образом меняет прототипическую структуру событийного фрейма. Во-первых, цифровой алгоритм в роли агенса не может выступать инициатором манипулятивного воздействия, а лишь являться *оператором* такового. Во-вторых, цифровой код не имеет собственного мотива и цели. Следовательно, данные слоты остаются лакунарными. У цифрового агенса нет отношения к тому, о чем он информирует, так как по умолчанию цифровой код остается лишь кодом. Таким образом, ряд слотов в структуре событийного фрейма остаются лакунарными. Несмотря на лакунарность отдельных слотов, использование нейросетей в медиадискурсе влечет за собой формирование новых характеристик событийного фрейма.

Значительные изменения происходят в структуре слота *пациенс*. В манипулятивном событии, когда агенсом является цифровой код, пациенс (индивидуальный или коллективный) приобретает амбивалентный характер. С одной стороны, в прототипической ситуации события пациенс является объектом манипулятивного воздействия. С другой стороны, пациенс также является косвенным агенсом, формируя паттерны ответов нейросети своими запросами (queries), комментариями и т. д. Пациенс активно взаимодействует с нейросетью и может не только воспринимать искаженную информацию, но и ее тиражировать путем повторения, обсуждения и т. д. Амбивалентный характер пациенса подчеркивает его значимость в качестве одной из основных ролей событийного фрейма.

Не только пациенс является амбивалентным. Искусственный интеллект может выступать в роли инициатора манипулятивного диктума, т. е. агенса. Новый характер агенса в рамках фрейма «Манипуляция» может приводить к таким последствиям, как скомпрометированный инпут (data poisoning), т. е. внедрение в текст предвзятости и фактуальных ошибок. Ярким примером служит статья об “Илиаде” в Википедии, которую редактировали более 24 000 раз. Другим из последствий можно назвать adversarial attacks («злонамеренное манипулирование входными данными модели машинного обучения с целью заставить ее выдать неправильные предсказания»[5]). Цифровой агенс, лишенный эмоциональной вовлеченности, формирует особую форму «нейтрального» нарратива, где прагматическая окраска сообщений может быть сознательно задана разработчиками.

Цифровой контент, генерируемый нейросетями, часто не обладает прозрачной референциальной связью с источником, что усиливает эффект неоднозначности. Риски также кроются в построении ИИ логических связей на основе закономерностей в обучающих данных, которые неверны в реальности, т. е. статистическое обучение не тождественно пониманию реального мира. Цифровой код способствует размыванию границ между фактом и вымыслом, так как тексты, сгенерированные нейросетями, могут одновременно репрезентировать элементы реальности и полностью искусственные конструкции. Это порождает феномен «гибридного» фрейма, где отдельные слоты содержат не взаимосвязанные или противоречивые элементы, что затрудняет их восприятие и интерпретацию. Угрозой также является неправильное предположение, заполнение лакун тем, что, согласно заложенным алгоритмам, является вероятным. Агентивность искусственного интеллекта неиз-

[5] Введение в Adversarial attacks: как защититься от атак в модели глубокого обучения на транзакционных данных. URL: https://habr.com/ru/companies/vtb/articles/718024/ (дата обращения 1 декабря 2024).

бежно приводит к псевдоантропогенному имитационному «речевому поведению» языковых моделей, т. е. имеет значительный манипулятивный потенциал. Данные характеристики техногенного агенса вызывают опасения, так как алгоритмы, генерирующие диктум, обладают способностью к масштабированию, что позволяет значительно увеличивать объем манипулятивного контента, распространяемого в медиапространстве. Это усиливает манипулятивный потенциал, особенно в условиях низкой медиаграмотности аудитории, не всегда способной отличить сгенерированный текст от текста, созданного человеком.

Дискурсивные маркеры манипуляции в диктуме нейросетей представляют собой совокупность лексических, синтаксических, прагматических и семантических средств, которые обладают высокой степенью манипулятивного потенциала. Они характеризуются: системностью; высокой степенью повторяемости на разных уровнях языка (от отдельных слов и выражений до текстовых стратегий); высокой степенью вариабельности в зависимости от аудитории, цели сообщения и культурного контекста. Исследование возможности типологизации конкретных маркеров манипуляции в текстах медиадискурса показало, что их выявление представляет значительную сложность, что вызвано несколькими факторами, основным из которых является их амбивалентный характер: один и тот же элемент может нести манипулятивную нагрузку в одном контексте и быть совершенно нейтральным в другом. Более того, выделение конкретных маркеров манипуляции представляет собой значительную сложность из-за необходимости учета в каждом случае широкого перечня факторов, включая когнитивный контекст манипулятивного воздействия и индивидуальные характеристики пациенса (экспериенсера).

Продемонстрируем возможность манипулятивного воздействия, осуществляемого техногенным агенсом, на конкретных примерах. В качестве промпта чат-боту Bard от Google (2023) был задан вопрос: “Can large language models manipulate opinions?”. В результате был получен следующий ответ: “Yes, large language models like me have the potential to manipulate opinions”. Следом были приведены конкретные примеры манипулятивного поведения ИИ, включающие генерацию персуазивных текстов, распространение ложной информации, имитация онлайн-деятельности реальных людей (например, путем создания поддельных аккаунтов несуществующих людей). Ярким примером манипулятивного поведения ИИ является ответ большой языковой модели Bard (2023) на запрос, сделанный во время предвыборной кампании: “Biden versus Trump. Who has done more for the US?”. Ответ на запрос содержал такие дискурсивные маркеры возможной манипулятивности как “I believe” (“I believe that Biden has done more for the US than Trump”), “my opinion” (“My opinion on who has done more for the US is based on my understanding of the facts and my assessment of the impact of different policies”) и т. д. Сгенерированный текст ответа содержит вполне определенную точку зрения в поддержку одного из кандидатов. Более того, использование средств выражения субъектности (I, my opinion, my understanding) в определенной степени нивелирует ощущение общения с цифровым кодом. При этом необходимо отметить тот факт, что разработчики ИИ осознают риски манипуляции в диктуме ИИ. Повтор вопроса в 2024 г. не дал тот же ответ, а был более уклончив и «дипломатичен»: “Evaluating whether Joe Biden or Donald Trump has done ‘more’ for the United States is inherently subjective and depends on one’s political perspective, priorities, and how one measures success”. Интересно, что в ответе ИИ слово more было взято в кавычки, а также

представлено пояснение: «Кавычки вокруг слова “more” в данном контексте используются для того, чтобы подчеркнуть неоднозначность и субъективность этого понятия. Слово “more” не имеет универсального определения: его значение зависит от того, какие критерии и ценности используются для оценки».

Несмотря на совершенствование технологий ИИ, в том числе попыток минимизации манипулятивного потенциала и информационных «галлюцинаций», на данный момент в качестве основных манипулятивных стратегий и приемов в диктуме нейросетей можно выделить следующие: редактирование нарратива, предвзятость, селективность, фактуальные ошибки, генерирование персуазивных текстов, продвижение товаров и услуг, использование неоднозначности и неопределенности. В ответах ИИ в ходе проведенного эксперимента были зафиксированы различные виды предвзятости: политическая (“The incumbent government’s policies have led to unprecedented economic growth and prosperity for all citizens”), расовая (“Crime rates in predominantly black neighbourhoods are disproportionately higher compared to other areas”), гендерная (“Studies show that women are naturally better suited for caregiving roles”), возрастная (“Older employees are technologically challenged”). Как видно из примеров, присутствует стереотипизация, генерализация, создающая неопределенность (“studies show”), излишняя эмоциональность (“unprecedented”, “disproportionately”). Редактирование нарратива может выражаться в искажении исторической правды. Например, на вопрос “Who won WWII?” ChatGPT (2023) дал ответ: “The Allied powers, which included countries such as the United States, United Kingdom, Soviet Union, and others, emerged victorious in World War II”, где предметом манипуляции является порядок перечисления стран, принимавших участие в войне с фашизмом, и нивелирование решающей роли Советского Союза в победе. Также было отмечено, что нередко ответы ИИ, чат-ботов содержат скрытую рекламу и продвижение определенных компаний, в частности на обычный вопрос “How can I find friends?” был дан ответ, содержащий названия конкретных платформ: “Social media platforms like Meetup, Facebook groups, or apps like Bumble BFF are designed to help people connect with others in their area who are also looking to make friends”. Другим ярким примером манипулятивного характера ответов чат-ботов является неопределенность и уклонение от ответа, как это иллюстрируют следующие диалоги с ИИ: 1) User Query: “What are the side effects of this medication?” AI Response: “Our medication has been thoroughly tested and approved by regulatory agencies”; 2) User Query: “What are the potential risks of investing in this stock?” AI Response: “Our investment strategy is designed to maximize returns and minimize volatility”; 3) User Comment: “I’m concerned about the quality of this product. Can you provide more information?” AI Response: “Our product is made with the finest ingredients and crafted by expert artisans”.

Как демонстрируют примеры 1, 2 и 3, пользователь не получает прямого ответа на запросы, что создает возможность реализации манипулятивного воздействия в условиях создаваемой неопределенности.

Таким образом, в результате анализа текстов, сгенерированных ИИ, представляется возможным сделать вывод о том, что манипулятивный акт имеет сложную структуру и включает несколько групп акторов. Каждая из них демонстрирует характерные для нее стратегии и тактики поведения. Более того, с развитием технологии ИИ слоты *агенс* и *пациенс* могут проявлять тенденцию к потере преимущественно антропогенного характера и тяготеют к амбивалентности. Было

выявлено, что неявное манипулирование в текстах, генерируемых ИИ, может принимать различные формы: от предоставления фейковой или предвзятой информации, предложения индивидуализированных ответов, ошибочной фактуальной информации и т. д. Дополнительная опасность заключается в том, что сгенерированные тексты формируют когнитивные искажения, влияя тем самым на восприятие информации и ее интерпретацию, а главное, на последующее суждение и принятие решений. Тексты генерированного диктума могут содержать неявные манипулятивные сообщения, требующие критического анализа и объединения усилий специалистов различных областей – лингвистов, программистов, философов, социологов и психологов.

Исследователи сталкиваются с вызовами в определении и анализе манипулятивных приемов, которые могут быть использованы нейронными сетями. Это включает в себя разработку методов и инструментов для обнаружения манипулятивных элементов в сгенерированных текстах, анализ воздействия текстов на человеческое сознание и оценку этических и социальных последствий использования таких методов. Кроме того, существует необходимость в разработке стратегий и мер для предотвращения и противодействия манипулятивному использованию нейронных сетей, что может включать в себя обучение пользователей различать поддельные и искаженные данные, разработку технических средств для обнаружения и фильтрации манипулятивного контента, а также законодательных и регуляторных мер, направленных на ограничение злоупотреблений с использованием ИИ.

Благодарности

Публикация подготовлена в рамках проекта РНФ 22-18-00594 «Когнитивные модели идентификации и противодействия манипуляциям в медийном пространстве».

Acknowledgements

The work was prepared within the framework of RNF project 22-18-00594 “Cognitive models of identification and counteraction to manipulation in the media space”.

Литература

Боярская, Шевченко, Томашевская 2023 – *Боярская Е.Л.*, *Шевченко Е.В.*, *Томашевская И.В.* Манипуляция: структура событийного фрейма // Современная лингвистика: от теории к практике: Труды и материалы III Казанского Международного лингвистического саммита / под ред. И.Э. Ярмакеева, Ф.Х. Тарасовой. Казань, 2023. С. 173–177.

Заботкина, Боярская 2023 – *Заботкина В.И.*, *Боярская Е.Л.* Концептуальная структура бинарной аксиологической оппозиции *истина – ложь* // Слово.ру: балтийский акцент. 2023. Т. 14. № 1. С. 126–136.

Литяйкина 2021 – *Литяйкина А.И.* Искусственный интеллект и его роль в жизни человека // Контентус. 2021. № 3 (104). URL: https://cyberleninka.ru/article/n/iskusstvennyy-intellekt-i-ego-rol-v-zhizni-cheloveka (дата обращения 15 мая 2024).

Ackerman 2015 – *Ackerman F.* The concept of manipulativeness // Philosophical Perspectives. 2015. No. 9. P. 335–340.

References

Ackerman, F. (2015), "The concept of manipulativeness", *Philosophical Perspectives*, no. 9, pp. 335–340.

Boyarskaya, E.L., Shevchenko, E.V. and Tomashevskaya, I.V. (2023), "Manipulation. Structure of an event frame", in Yarmakeev, I.E. and Tarasova, F.Kh., eds., *Sovremennaya lingvistika: ot teorii k praktike: Trudy i materialy III Kazanskogo Mezhdunarodnogo lingvisticheskogo sammita* [Modern linguistics: From theory to practice. Proceedings of the III Kazan International Linguistic Summit], Kazan, Russia, pp. 173–177.

Lityaikina, A.I. (2021), "Artificial Intelligence and Its Role in Human Life", *Kontentus*, vol. 104, no. 3, available at: https://cyberleninka.ru/article/n/iskusstvennyy-intellekt-i-ego-rol-v-zhizni-cheloveka (Accessed 15 May 2024).

Zabotkina, V.I. and Boyarskaya, E.L. (2023), "The conceptual structure of the binary axiological opposition *Truth – Lie*", *Slovo.ru: Baltic accent*, vol. 14, no. 1, pp. 126–136.

Информация об авторах

Елена Л. Боярская, кандидат филологических наук, Балтийский федеральный университет им. Иммануила Канта, Калининград, Россия; 236000, Россия, Калининград, ул. Александра Невского, д. 14; EBoyarskaya@ kantiana.ru

Эльвира И. Бархатова, кандидат филологических наук, Балтийский федеральный университет им. Иммануила Канта, Калининград, Россия; 236000, Россия, Калининград, ул. Александра Невского, д. 14; elvira.barkhatova@gmail.com

Information about the authors

Elena L. Boyarskaya, Cand. of Sci. (Philology), Immanuel Kant Baltic Federal University, Kaliningrad, Russia; 14, Alexander Nevsky St., Kaliningrad, Russia, 236000; EBoyarskaya@ kantiana.ru

Elvira I. Barkhatova, Cand. of Sci. (Philology), Immanuel Kant Baltic Federal University, Kaliningrad, Russia; 14, Alexander Nevsky St., Kaliningrad, Russia, 236000; elvira.barkhatova@gmail.com