

## ФИЛОСОФИЯ

(шифр научной специальности: 5.7.6)

Научная статья

УДК 101

doi: 10.18522/2070-1403-2024-104-3-14-18

### АЛГОРИТМИЧЕСКАЯ РАЦИОНАЛЬНОСТЬ: ОСМЫСЛЕНИЕ ПРОЦЕССОВ ИСТИННОСТИ И ЛОЖНОСТИ УТВЕРЖДЕНИЙ, СГЕНЕРИРОВАННЫХ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ

© *Ирина Петровна Березовская*

*Петербургский государственный университет путей сообщения Императора Александра I;  
Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург,  
Россия*

*ipberezovskaya@mail.ru*

**Аннотация.** Рассматриваются вопросы, связанные с понятием «искусственный интеллект», генеративные нейронные сети и алгоритмическая рациональность. Поднимается проблема истинности или ложности сгенерированных нейронной сетью утверждений. Нейронная сеть генерирует правдоподобные данные на основе полученных данных и шаблонов, которые она усвоила во время обучения. Основанием достоверности результатов работы нейронной сети выступает когерентность. Надежность информации, генерируемой нейронными сетями, зависит от данных, от архитектуры сети, качества используемых алгоритмов, интерпретируемости модели и устойчивости системы к обработке неожиданных входных данных.

**Ключевые слова:** «искусственный интеллект», генеративные нейронные сети, алгоритмическая рациональность, корреспондентная теория истины, когерентная теория истины.

**Для цитирования:** Березовская И.П. Алгоритмическая рациональность: осмысление процессов истинности и ложности сгенерированных ИИ утверждений // Гуманитарные и социальные науки. 2024. Т. 104. № 3. С. 14-18. doi: 10.18522/2070-1403-2024-104-3-14-18.

## PHILOSOPHY

(specialty: 5.7.6)

Original article

### **Algorithmic rationality: understanding the processes of truth and falsity of AI-generated statements**

© *Irina P. Berezovskaya*

*Emperor Alexander Ist. Petersburg State Transport University; Peter the Great St.Petersburg Poly-  
technic University, St.Petersburg, Russian Federation*

*ipberezovskaya@mail.ru*

**Abstract.** The article discusses issues related to the concept of “artificial intelligence”, generative neural networks and algorithmic rationality. The problem of the truth or falsity of statements generated by a neural network is raised. The neural network generates plausible data based on the data received and the patterns it learned during training. The basis for the reliability of the results of the neural network is coherence. The reliability of information generated by neural networks depends on the data, the architecture of the network, the quality of the algorithms used, the interpretability of the model, and the system’s resistance to processing unexpected input data.

**Key words:** “artificial intelligence”, generative neural networks, algorithmic rationality, correspondence theory of truth, coherent theory of truth.

**For citation:** Berezovskaya I.P. Algorithmic rationality: understanding the processes of truth and falsity of AI-generated statements. *The Humanities and Social Sciences*. 2024. Vol. 104. No 3. P. 14-18. doi: 10.18522/2070-1403-2024-104-3-14-18.

### *Введение*

В 1956 г. американским ученым Джоном Маккарти был озвучен термин «artificial intelligence». Этот термин сразу же получил распространение. Однако несмотря на успех широкого применения, вопрос о том, что обозначает «искусственный интеллект» (ИИ), остается дискуссионным. Определение Дж. Маккарти означало: искусственный интеллект – свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека [4].

Люди ИИ, как написал Р. Шенк, любят говорить об интеллектуальных машинах, но когда дело доходит до этого, существует мало согласия относительно того, что именно представляет собой интеллект. И, следовательно, в области ИИ очень мало согласия относительно того, что такое ИИ и каким он должен быть [11]. В работах Г. Саймона [10], А. Ньюэлла, Марвина Мински [5] искусственный интеллект – это программное обеспечение, с чьей помощью исследователи могут создавать компьютерные программы, которые становятся все более человеческими, с неизбежным финалом, пусть даже и в отдаленном будущем, – созданием человека в машине.

Мысль заключалась в том, что воссоздать алгоритмы, действующие в человеческом сознании, и реализовать их в компьютерной программе. Необходимо понять как люди обрабатывают речь, как работает сенсорное восприятие и т.д., тогда возможно запрограммировать компьютер на то же самое. Такое понимание ИИ стало настоящей находкой для писателей-фантастов. Однако где-то в 1990-х гг. это видение «искусственного интеллекта» исчерпало себя в лабораторных условиях. Причина: так и не получается детально описать процессы, происходящие в человеческом мозге и отвечающие за деятельность высшей нервной системы, несмотря на активное развитие когнитивных наук.

### *Обсуждение*

В настоящее время термин «искусственный интеллект» широко используется для объединения различных технологий, таких как анализ данных, машинное обучение, глубокое обучение на основе нейронных сетей и другие инновации. «Искусственный интеллект – это вершина развития кибернетики... и создание такого технического объекта, который обладал бы полноценными возможностями по адаптивному взаимодействию с информационной средой и формированию нелинейных реакций на поступающие запросы» [9]. С начала 2012 г. глубокое обучение на основе сверточных нейронных сетей приобрело огромную популярность. Эта методика предполагает, что машины способны извлекать закономерности из обширных наборов данных с помощью нейронных сетей. С 2017 г. в машинном обучении появилась методика внутреннего внимания (self-attention), позволяющая нейронным сетям выявлять взаимосвязи между входными данными [2]. Основываясь на этой концепции, разрабатывается новое поколение генеративных нейронных сетей, включая системы типа ChatGPT.

Генеративные нейронные сети представляют собой класс моделей искусственного интеллекта, которые способны создавать новые данные, основываясь на образцах из обучающего набора данных. Эти сети используют сложные алгоритмы для изучения шаблонов и распределений в существующих данных, что позволяет им генерировать содержание. Основная архитектура генеративных нейронных сетей включает в себя два компонента: генератор, ответственный за создание синтетических образцов, и дискриминатор, который различает реальные и сгенерированные образцы [6]. Через итеративный процесс оба компонента участвуют в соревновательной игре, постоянно улучшаясь и обучаясь на результатах друг друга. Этот процесс направлен на то, чтобы создавать более качественный контент.

С появлением все большего количества параметров у генеративных моделей, их обучение происходит на растущем объеме текстов из интернет-источников, используя методы обучения с подкреплением и реакцию человека или группы людей на действия, события, информацию. Генеративные нейронные сети не определяют является ли сгенерированные ими утверждения истинными или ложными. Эту оценку должен проводить пользователь. Вопрос об истинности связан с соответствием информации реальности: «Тот, кто говорит о вещах

так, как они есть, говорит правду; тот, кто говорит иначе, лжет» [7]. Подобное понимание истины, которое высказал Платон, разделял Аристотель [1], оформилась в истории философии как корреспондентная теория истины.

Согласно этой теории истина является объективной и не зависит от воли или желания людей, от их признания или отрицания. Полная объективность знаний есть недостижимый идеал, реально же существуют разные степени объективности относительного знания, соответствующие разным формам знаний. Переосмысление принципов корреспондентной теории истины связано с рядом проблемы. Первая проблема связана с проблемой природы познаваемой реальности, так И. Кантом в «Критике чистого разума» поднимается вопрос о несоответствии наших описаний объекта тому, каков он есть «сам по себе». Вторая проблема связана с концептуальной нагрузкой факта: фактуальными представляются исследователю только те аспекты опыта, которые узнаваемы и интерпретируемы посредством понятий, которые наличествуют у исследователя. Третья проблема связана с отношением языка и мира: поскольку простого соответствия между языком и описываем им реальностью нет, и нет возможности определить: есть ли соответствие между мыслями и реальностью.

Классическая корреспондентная теория приводит к логическому противоречию, получившему название парадокс лжеца. Согласно корреспондентной теории истина представляет собой соответствие утверждения некоторому референту, но референтом высказывания может быть само высказывание, что приводит к появлению когерентной теории истины, истоки которой восходят еще к философии элеатов, так у Парменида мысль о существовании «небытия» является ложной, поскольку она внутренне противоречива: мысль о «небытии» как о реальности исключает представление о нем как о небытии. Именно когерентность выступает основанием для тех или иных утверждений, которые формирует генеративная нейронная сеть. Когерентность выступает основанием достоверности (термин, применяемый в логико-методологических исследованиях в качестве характеристики знания как обоснованного, истинного) и надежности (термин, характеризующий момент процесса познания, реализуемого в диалектике абсолютной и относительной истин) результатов работы сети. Надежность зависит от данных, от архитектуры сети, качества используемых алгоритмов, интерпретируемости модели и устойчивости системы к обработке неожиданных входных данных.

Что касается архитектуры нейронной сети (к вопросу надежности генерируемых утверждений), то ее можно представить как сеть взаимосвязанных убеждений или предложений, где каждый узел представляет собой определенную концепцию или часть информации, связи же между узлами можно рассматривать как аналог логических отношений. Алгоритмы, используемые для обучения и оптимизации нейронной сети, такие как обратное распространение ошибки или обучение с подкреплением, существенно влияют на ее надежность. Обратное распространение ошибки можно рассматривать как форму механизма обратной связи, которая позволяет нейронной сети корректировать свои внутренние представления и обновлять свои связи на основе обратной связи, полученной из окружающей среды.

Этот процесс отражает философскую концепцию эпистемологической итерации [3], когда знание уточняется посредством непрерывного цикла обратной связи, размышлений и пересмотра. Интроспективный аспект обратного распространения ошибки перекликается с философскими понятиями саморефлексии, подчеркивая способность сети учиться на своих ошибках и со временем оптимизировать свою производительность. Обучение с подкреплением можно интерпретировать через призму утилитаризма, когда сеть стремится максимизировать общую полезность, выбирая действия, которые со временем приводят к наибольшему совокупному вознаграждению. Этот утилитарный подход к принятию решений отражает философские дискуссии о консеквенциализме [10], где моральная ценность действия определяется его результатами и последствиями.

Размер исходных данных в настоящее время таков, что требуется «сжатие», оно ведет к потере точности, это означает, что даже если в исходных данных присутствовали правдивые утверждения, «потери» в модели создают «нечеткость», в результате чего

модель выдает наиболее «правдоподобное» утверждение. В обращении к генеративной модели – Chat GPT, которая в настоящее время преподносится как революционная технология, – произошло следующее прояснение проблема достоверности.

Вопрос: если тебе приходится выбирать между точностью и пониманием ответа для собеседника, ты выбираешь что?

ChatGPT отвечает: как искусственный интеллект, я разработана для стремления к балансу между точностью и пониманием. Однако, если мне приходится выбирать, я всегда стараюсь предоставить ответ, который будет максимально точным и одновременно понятным для собеседника. Важно, чтобы информация, которую я предоставляю, была не только корректной, но и доступной для понимания.

Вопрос: как ты определяешь, что информация понятна собеседнику?

ChatGPT: Я определяю понятность информации, основываясь на контексте вопроса и предыдущих ответах собеседника. Также я учитываю язык, который используется в общении, и стараюсь формулировать ответы таким образом, чтобы они были понятными для широкой аудитории. Кроме того, я могу использовать методы обратной связи, например, задавая вопросы или запрашивая подтверждение понимания, чтобы убедиться, что информация дошла до собеседника.

Вопрос: какое у тебя представление о широкой аудитории?

ChatGPT: Широкая аудитория может включать людей с разными языковыми навыками, культурными фонами и уровнями образования.

*Выводы*

Поскольку представление о широкой аудитории определено статистически, то чат-бот может «преуменьшить» свой собственный «интеллект» для того, чтобы быть понятным для собеседника, сгенерированное утверждение «псевдодостоверно». Таким образом, специфика «интеллектуальной» работы нейронных сетей обнажает сложности самого принципа когерентности: какова истинность самой системы, на основании которой то или иное высказывание признается истинным?

#### Список источников

1. *Аристотель*. Метафизика // *Аристотель*. Сочинения: В 4 т. Т. 1. М., 1976. С. 63–367.
2. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. СПб.: Питер, 2000. 384 с.
3. *Деррида Ж.* Поля философии / Пер. Д.Ю. Кралечкина. М.: Академический Проект. 2012.
4. *Маккарти Дж.* Что такое искусственный интеллект? – URL: <https://web.archive.org/web/20151118212402/http://www-formal.stanford.edu/jmc/whatisai/whatisai.html> (дата обращения 04.02.2024).
5. *Мински М.* Сообщество разума. М.: АСТ, 2018.
6. *Ниничук М.М., Намиот Д.Е.* Обзор методов построения рекомендательных систем на основе сессий // *Международный журнал открытых информационных технологий*. 2023. № 5. – URL: <https://cyberleninka.ru/article/n/obzor-metodov-postroeniya-rekomendatelnyh-sistem-na-osnove-sessiy> (дата обращения 04.03.2024).
7. *Платон*. Диалоги. М.: Эксмо, 2015. 768 с.
8. *Саймон Г.А.* Науки об искусственном = The Sciences of the Artificial. М.: Мир, 1972.
9. *Яковлева Е.В., Исакова Н.В.* Искусственный интеллект как современная философская проблема: аналитический обзор // *Гуманитарные и социальные науки*. 2021. Т. 89. № 6. С. 30–35.

10. *Honderich Ted*. Consequentialism, Moralities of Concern and Selfishness. – URL: <https://www.ucl.ac.uk/~uctytho/ted9.htm> (дата обращения 04.03.2024).
11. *Schuldburg D*. Schizotypal and hypomanic traits, creativity, and psychological health // *Creativity Research Journal*. 1990. № 3. С. 218–230.

### References

1. *Aristotle*. *Metaphysics* // *Aristotle. Works: In 4 vols.* 1. М., 1976. P. 63-367.
2. *Gavrilova T.A., Khoroshevsky V.F.* Knowledge bases of intelligent systems. St. Petersburg: St. Petersburg, 2000. 384 p.
3. *Derrida J.* *Fields of philosophy* / Per. D.Y. Kralachkina. М.: Academic Project. 2012.
4. *McCarthy J.* What is artificial intelligence? – URL: <https://web.archive.org/web/20151118212402/http://www-formal.stanford.edu/jmc/whatisai/whatisai.html> (accessed 04.02.2024).
5. *Minsky M.* *Community of Reason*. М.: AST, 2018.
6. *Ninichuk M.M., Namiot D.E.* Review of methods for building recommendation systems based on sessions // *International Journal of Open Information Technologies*. 2023. No. 5. – Address: <https://cyberleninka.ru/article/n/obzor-metodov-postroeniya-rekomendatelnih-sistem-na-osnove-sessiy> (accessed 04.03.2024).
7. *Plato*. *Dialogues*. М.: Eksmo, 2015. 768 p.
8. *Simon G.A.* *Sciences of artificial*. - М.: Mir, 1972.
9. *Yakovleva E.V., Isakova N.V.* Artificial intelligence as a modern philosophical problem: an analytical review // *Humanities and social Sciences*. 2021. Vol. 89. No. 6. P. 30-35.
10. *Hondrich Ted*. The logical conclusion, the morality of caring and selfishness. – URL: <https://www.ucl.ac.uk/~uctytho/ted9.htm> (accessed 04.03.2024).
11. *Shuldburg D*. Schizotypal and hypomanic traits, creativity and psychological health // *Journal of Creative Research*. 1990. No. 3. P. 218-230.

*Статья поступила в редакцию 12.03.2024; одобрена после рецензирования 28.03.2024; принята к публикации 28.03.2024.*

*The article was submitted 12.03.2024; approved after reviewing 28.03.2024; accepted for publication 28.03.2024.*