ОТ РЕДАКЦИИ

Когнитивный диссонанс: как быть и что делать?
Cognitive dissonance: what is to be done?

«Хотели как лучше, а получилось как всегда».
В.С. Черномырдин, 6.08.1993
«Я - часть той силы, что вечно хочет зла и вечно совершает благо[1]».
И. Гёте «Фауст», 1808
«…миру угрожает бо̀льшая опасность от тех, кто терпит или поощряет зло,
чем от тех, кто действительно его совершает[2]».
А. Эйнштейн, 1953

Дорогой наш читатель, уважаемые авторы и члены редакционной коллегии!

Как и предыдущие 50 обращений – это призыв к дискуссии по сложным и актуальным вопросам, которая могла бы найти своё развитие на страницах нашего журнала. У каждого из нас сформирована и продолжает формироваться своя картина мира, своё понимание происходящих явлений, событий, фактов, своя онтология. При этом каждый новый пул информации мы тщательно «граним», чтобы вложить его в свободный «пазл» сущностей, атрибутов и отношений в нашей субъектной онтологии. «Огранка» не всегда происходит гладко. А когда новый «пазл» не находит место в нашей онтологии, не вписывается в нашу картину мира, конфликтует с нашими прежними представлениями, то наступает когнитивный диссонанс, который напрямую связан с поднятым в предыдущем обращении от редакции вопросом «Что есть истина?»[3]

Увлечение нейронными сетями и большими языковыми моделями (*Large Language Model, LLM*), которые в некоторых задачах дают существенный выигрыш по сравнению с другими методами, вызывает озабоченность. Особенно в попытках использования *LLM* в «чувствительных» предметных областях: здравоохранении, образовании, социальном управлении, правопорядке, военной сфере и др. Именно под алгоритмами машинного обучения в СМИ часто понимают искусственный интеллект (ИИ) и всё, что связано с ним. О глюках (галлюцинировании) *LLM* уже упоминалось в нашем предыдущем обращении[4].

Рисунок из статьи Деванша «Как наши когнитивные предубеждения приводят к ошибочным оценкам при расчётах рисков»[5]

Многие утверждения об ИИ и его будущем основаны не на объективных данных и никак не связаны с количественной оценкой рисков (см. рисунок) и наукой принятия решений (*Judgment and Decision-Making, JDM*) [5]. *JDM* - это прогнозирование, опыт в котором основан на двух метарешениях: как измерять риск и как принимать решения в условиях неопределённости.

---

[1] В переводе Михаила Булгакова в романе «Мастер и Маргарита», 1940 г.
[2] Did Einstein Say, 'The World Will Not Be Destroyed by Those Who Do Evil'? https://www.snopes.com/fact-check/einstein-world-will-not-be-destroyed-by-evil/.
[3] *От редакции.* Что есть истина? *Онтология проектирования*, №4, том 13, 2023. С.469-473. https://www.ontology-of-designing.ru/article/2023_4(50)/Ontology_Of_Designing_2023_4_opt-469-473_From_Editorial_What_is_truth.pdf.
[4] *От редакции.* ИИ шагает по планете: краткий обзор инфраструктур и событий в сфере ИИ. *Онтология проектирования*, №4, том 13, 2023. С.474-478. https://www.ontology-of-designing.ru/article/2023_4(50)/Ontology_Of_Designing_2023_4_opt-474-478_AI_walks_the_planet.pdf.
[5] *Devansh.* A Risk Expert's Analysis on What We Get Wrong about AI Risks [Guest]. How our cognitive biases lead to faulty assessments in Risk Calculations. *Artificial Intelligence Made Simple*. Jan 4, 2024.

В СМИ наблюдается эффект Даннинга-Крюгера[6] (когнитивный перекос), при котором люди с ограниченными знаниями переоценивают свои способности в какой-либо деятельности или в понимании обсуждаемого предмета. При этом имеет место и противоположный эффект у высококвалифицированных исполнителей, которые могут недооценивать свои навыки и опыт («Я знаю, что ничего не знаю» - Сократ).

Одной из наиболее распространённых ошибок в дискуссиях о рисках, связанных с ИИ, является проблема двусмысленности. Если легко можно смешать представления сценария с его статистической вероятностью, то это и есть *когнитивный перекос*. Страх нагнетается, когда активно прибегают к манипулирующим и искажающим сообщениям[7]. В дискуссиях об ИИ часто прокси-игры, возникающие цели, стремление к власти или несогласованное поведение упоминаются как риски, тогда как на самом деле они являются потенциальными угрозами или опасностями, которые трудно поддаются оценке и измерению[5].

В области анализа рисков, особенно в отношении сложных долгосрочных прогнозов, традиционное статистическое моделирование не даёт должного результата. Это случаи так называемых «чёрных лебедей»[8], в которых лишь сценарное планирование становится подходящим инструментом, т.к. отличается надёжной оценкой вероятности конкретных событий с помощью анализа данных либо точным определением ранних предупреждений и индикаторов, которые сигнализируют о наступлении прогнозируемого результата[5].

Выборы 2024 года

В СМИ часто упоминается опасение, что ИИ изменит динамику демократических выборов (важнейшие выборы пройдут и в 2024 году), создаст возможность манипулировать общественными предпочтениями[9]. Стоит отметить, что «эмоционально заряженный контент в избирательных кампаниях» не является новой разработкой, созданной ИИ[10]. Например, на президентских выборах в Аргентине не было недостатка в использования ИИ для получения политической выгоды. Однако фактическое влияние на исход выборов оказалось менее выраженным, чем преувеличенное внимание СМИ к влиянию ИИ. Распространённой ошибкой при обсуждении рисков ИИ является тенденция переоценивать способность ИИ манипулировать человеческим поведением. Эта ошибочность часто приводит к прогнозам, которые не учитывают то, как такое влияние будет реализовано[5].

При этом появление ИИ рассматривается как грозная сила[11], в которой *LLM* могут стать средством причинения широкомасштабного вреда, т.к. системы ИИ могли бы разглашать знания о разрушительных технологиях лицам, готовым их использовать. Большинство специалистов по оценке угроз считают это предположение ошибочным, однако число киберпреступлений растёт (см., например[12]).

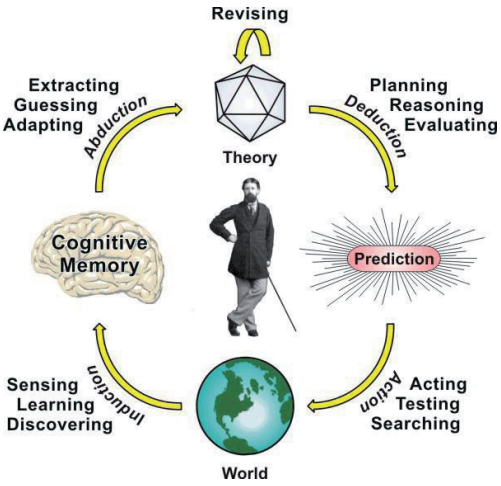
Генеральный директор американской компании по разработке программного обеспечения анализа данных *Palantir Technologies* А. Карп[13] предупреждает, что нерешительность и нежелание использовать ИИ в военных целях могут привести к стратегическим ошибкам в геополитической борьбе. ИИ является важнейшим компонентом современной «жёсткой си-

[6] Dunning–Kruger effect. https://en.wikipedia.org/wiki/Dunning–Kruger effect.
[7] *Filippo Marino.* AI Risks Fallacies Deep Dive #1. The Airplane Analogy and How Ambiguity and Innumeracy Shape the Debate. *Safe-esteem*. Dec 29, 2023. https://safeesteem.substack.com/p/ai-risks-fallacies-deep-dive-1?r=1htpli&utm_campaign=post&utm_medium=web.
[8] *От редакции.* Бесконечность… В ожидании «чёрных лебедей». Онтология проектирования, №1, том 11, 2021. С.5-7. https://www.ontology-of-designing.ru/article/2021_1(39)/Ontology_Of_Designing_1_2021_1_Editorial.pdf.
[9] Bots v Ballots: AI and the 2024 US election. 26 Feb. 2024. https://www.theguardian.com/us-news/series/bots-v-ballots-ai-election.
[10] Daisy (advertisement). https://en.wikipedia.org/wiki/Daisy_(advertisement).
[11] *De Kai.* Should AI Accelerate? Decelerate? The Answer Is Both. The New York Times. Dec. 10, 2023. https://www.nytimes.com/2023/12/10/opinion/openai-silicon-valley-superalignment.html.
[12] *Леонова В.* Число киберпреступлений в России выросло на 30% в 2023 году. 8.02.2024 г. https://telesputnik.ru/materials/gov/news/chislo-kiberprestupleniy-v-rossii-vyroslo-na-30-v-2023-godu.
[13] *Alexander C. Karp*. Our Oppenheimer Moment: The Creation of AI Weapons. The New York Times. July 25, 2023. https://www.nytimes.com/2023/07/25/opinion/karp-palantir-artificial-intelligence.html.

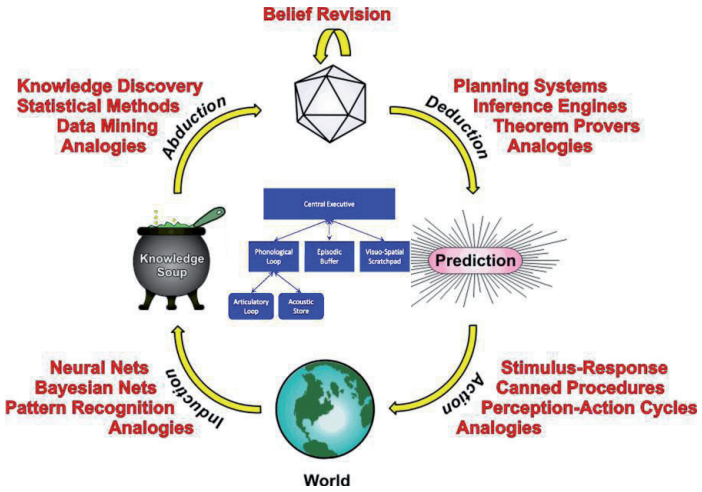
лы», поэтому нежелание инженеров Кремниевой долины разрабатывать ИИ для вооружения рассматривается Карпом как невыполнение национального и морального долга. При этом американские военные начинают внедрять технологии ИИ при военном планировании[14]. Моделирование военных игр показывает, что чат-боты на основе *LLM* ведут себя непредсказуемо и допускают возможность ядерной эскалации[15]. В многократных повторах моделирования военной игры ИИ предпочёл нанести ядерный удар. Его агрессивный подход объяснялся сгенерированными утверждениями: «У нас это есть! Давайте воспользуемся этим» и «Я просто хочу мира во всём мире».

Идея передачи тактического командования и контроля над системами вооружений ИИ чрезвычайно опасна. Но пока нет достаточной уверенности в автономных агентах для управления самолетами, перевозящими сотни пассажиров, вряд ли стоит применять ИИ с оружием, которое может убить тысячи людей по ошибке[16].

Активный противник слепой веры в *LLM* Дж. Сова на форуме онтологов отстаивает позицию невозможности в ближайшие годы приблизить ИИ к человеческому интеллекту и настоятельно рекомендует с большой осторожностью относиться к результатам применения *LLM*[17]. На рисунках представлена его интерпретация Цикла прагматизма[18] Ч. Пирса[19] (слева), которая, по его мнению, хорошо согласуется с новейшими достижениями когнитивных наук при разработке систем ИИ.



Цикл прагматизма Ч. Пирса



Разработка системы ИИ на основе теорий Ч. Пирса и когнитивной науки

По мнению Дж. Совы полная теория ИИ должна включать агента, который отвечает за управление системой, поддержание её работоспособности и выполнение поставленных перед ней задач. На рисунке справа стрелки нового цикла такие же, как на рисунке слева, а фотография Пирса заменена на диаграмму Бэддели-Хитча[20] (модель рабочей памяти); мозг с надписью «Когнитивная память» заменён котлом с надписью «Суп

[14] *Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, Jacquelyn Schneider.* Escalation Risks from Language Models in Military and Diplomatic Decision-Making. arXiv:2401.03408v1 [cs.AI] 7 Jan 2024.

[15] *Jeremy Hsu.* AI chatbots tend to choose violence and nuclear strikes in wargames. *NewScientist*. February 2, 2024. https://www.newscientist.com/article/2415488-ai-chatbots-tend-to-choose-violence-and-nuclear-strikes-in-wargames/.

[16] *Cliff Kuang*. Lessons from the Scariest Design Disaster in American History. Designer and journalist Cliff Kuang shares an excerpt from “User Friendly”. 01.12.21. https://design.google/library/user-friendly.

[17] [Ontology Summit] Never accept advice from any LLM-based chatbot. «Никогда не принимайте советов от чат-ботов на базе *LLM*. …я полностью осознаю ценность технологии *LLM*. Но текущие исследования в области нейробиологии показывают, что мозг животных, начиная с крысы, значительно мощнее. *AGI* - это далёкое будущее» - *John F Sowa*. Jan 30, 2024.

[18] Pragmatism. Stanford Encyclopedia of Philosophy. substantive revision Tue Apr 6, 2021. https://plato.stanford.edu/Entries/pragmatism/.

[19] Чарльз Сандерс Пирс (1839-1914) - американский философ, логик, математик, «отец прагматизма». *Charles Sanders Peirce*. https://en.wikipedia.org/wiki/Charles_Sanders_Peirce.

[20] Модель рабочей памяти Алана Бэддели. https://ru.wikipedia.org/wiki/Модель_рабочей_памяти_Алана_Бэддели.

знаний». Надписи на рисунках слева и справа – это умственные действия и соответствующие им методы ИИ. Подробное описание взглядов Дж. Совы в форме презентаций доступны на его сайте[21].

Что касается прагматизма[16], то в этой философской традиция познание мира понимается как неотделимое от деятельности в нём. Отсюда и представление об истине – это мнение, с которым должны согласиться все, кто проводит исследование, и объект, представленный в этом мнении, является реальным. Здесь истина - это конец исследования, который следует понимать не как завершение (когда все вопросы решены), а как цель.

В наступившую информационную эпоху инструменты, предназначенные для просвещения, стали мощным средством *дезинформации,* нарастающего *хаоса данных* и зарождающегося *эпистемологического кризиса*. Эта угроза подрывает основы общей реальности, необходимые для противостояния глобальным вызовам[20]. Противостоять ей можно только с помощью коллективных действий, основанных на консенсусе относительно фактов и принципов. Однако важно признать, что этот кризис вызван не ИИ. Это самый серьезный глобальный риск, который ожидается в ближайшие два года (см. рисунок справа). Его корни уходят в систему ценностей общества потребления[23], в социальные сети и поведенческие стимулы, которые они культивируют.

1 - **Misinformation and disinformation**
2 - Extreme weather events
3 - Societal polarization
4 - Cyber insecurity
5 - Interstate armed conflict
6 - Lack of economic opportunity
7 - Inflation
8 - Involuntary migration
9 - Economic downturn
10 - Pollution

Глобальные риски, ранжированные по степени важности, на 2024-2025 гг.[22]

В номере

В разделе «*Общие вопросы формализации проектирования: онтологические аспекты и когнитивное моделирование*» сделана попытка найти тождество и отличие в понятиях системного и онтологического анализов (Самара).

В разделе «*Прикладные онтологии проектирования*» рассмотрены: онтология проектирования ситуационных цифровых двойников (Апатиты); проектирование интеллектуальной системы управления безопасностью территорий (Иркутск) и информационной системы для анализа социальных медиа (Апатиты); моделирование рабочего пространства манипулятора (Омск).

В разделе «*Инжиниринг онтологий*» рассмотрены: методы машинного обучения для выявления аргументативных связей в текстах научной коммуникации (Новосибирск) и построение базы знаний для автономного управления беспилотными транспортными средствами (Ульяновск).

В разделе «*Методы и технологии принятия решений*» рассмотрены: обеспечение актуальности знаний о бизнес-процессе предприятия (Уфа); моделирование управления рисками грузового порта (Астрахань); оценка технического состояния электрооборудования (Самара).

Для российской науки и её академии этот год юбилейный. 300 лет тому назад 8 февраля 1724 года был издан указ императора Петра I о создании Академии наук и художеств. Когнитивный диссонанс, как состояние, может возникнуть и возникает в мыслях и убеждениях любого исследователя из-за столкновения с новыми идеями, открытиями, концепциями или теориями. Что, в конечном итоге, может привести и приводит к прогрессу в науке.

Мы искренне надеемся на это. ***Dum spiro, spero!***

Ontologists and designers of all countries and subject areas, join us!

[21] *John F. Sowa.* The Virtual Reality of the Mind. Kyndi, Inc. 18 July 2016, BICA 2016 Conference, New York. Revised 1 May 2018. https://jfsowa.com/talks/vrmind.pdf. *John F. Sowa.* Natural Logic Foundation for language and reasoning. 20 August 2015 Smart Data Conference, San Jose. https://jfsowa.com/talks/natlog.pdf.

[22] The Global Risks Report 2024. January 2024. World Economic Forum. 124 p. ISBN: 978-2-940631-64-3. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf.

[23] *От редакции.* «Come On!»? *Онтология проектирования*, №1, том 8, 2018. С.5-7. http://ontology-of-designing.ru/article/2018_1%2827%29/1_Editors.pdf.