

Научная статья

УДК 004.8

doi: 10.47475/1994-2796-2024-486-4-78-86

## ВЛИЯНИЕ ЯЗЫКОВОЙ ОТНОСИТЕЛЬНОСТИ НА ОБУЧЕНИЕ И РАБОТУ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Дмитрий Валерьевич Мамченков<sup>1</sup>, Иван Сергеевич Горбачёв<sup>2</sup>

<sup>1</sup>Российский университет дружбы народов, Москва, Россия, mamchenkov-dv@rudn.ru, 0000-0003-0484-2232

<sup>2</sup>Российский университет дружбы народов, Москва, Россия, ivan58830@gmail.com, 0009-0007-4893-3001

**Аннотация.** Статья посвящена анализу влияния лингвистической относительности на существующие нейросети и прогнозу потенциально возможного влияния данного принципа в будущем на более совершенные нейросети. В статье приводится анализ языковых ситуаций при работе с нейросетями и демонстрация конкретных примеров влияния языка на работу ИИ, разбираются возможные случаи получения положительного эффекта при обучении нейросетей и учёта влияния принципа языковой относительности. Описаны проведённые над нейросетями эксперименты, демонстрирующие существование такого влияния. Сформулированы принципы, позволяющие начать работу по формированию рамок взаимодействия с ИИ. Принцип вовлечённости, предполагающий учёт фактора вовлечённости ИИ при работе с естественным языком не только в смысле, но и коннотации. Принцип Каи: избегать наделяния ИИ сознанием и внутренним миром и не использовать подобные метафоры.

**Ключевые слова:** философия нейросетей, принцип лингвистической относительности, машинное обучение и язык, философия искусственного интеллекта, языковая среда, философия языка

**Финансирование.** Исследование выполнено в рамках проекта инициативной НИР № 100414-0-000 «Культура, наука и технологии: вызовы современности».

**Для цитирования:** Мамченков Д. В., Горбачёв И. С. Влияние языковой относительности на обучение и работу искусственного интеллекта // Вестник Челябинского государственного университета. 2024. № 4 (486). С. 78–86. DOI: 10.47475/1994-2796-2024-486-4-78-86.

Original article

## INFLUENCE OF THE LINGUISTIC RELATIVITY ON LEARNING AND WORK OF ARTIFICIAL INTELLIGENCE

Dmitry V. Mamchenkov<sup>1</sup>, Ivan S. Gorbachev<sup>2</sup>

<sup>1</sup>Peoples' Friendship University of Russia, Moscow, Russia, mamchenkov-dv@rudn.ru, 0000-0003-0484-2232

<sup>2</sup>Peoples' Friendship University of Russia, Moscow, Russia, ivan58830@gmail.com, 0009-0007-4893-3001

**Abstract.** The article analyses influence of linguistic relativity on existing neural networks, as well as to the forecast of the potential influence of this principle in the future on more advanced neural networks. The investigation provides an analysis of language situations when working with neural networks and a demonstration of specific examples of the influence of language on the work of AI, as well as possible situations of obtaining a positive effect when training neural networks considering the influence of linguistic relativity. Experiments conducted on neural networks demonstrating the existence of such an influence are described. The principles that allow to begin work on the formation of a framework for interaction with AI are formulated. The principle of engagement, which assumes that the factor of AI involvement in working with natural language not only in meanings, but also in connotations is to be taken into account. The principle of Kaja: to avoid endowing AI with consciousness and private world and not to use such metaphors.

**Keywords:** philosophy of neural networks, the principle of linguistic relativity, machine learning and language, philosophy of artificial intelligence, language environment, philosophy of language

**Funding.** The research was carried out within the framework of the initiative research project No. 100414-0-000 “Culture, science and technology: challenges of modernity”.

**For citation:** Mamchenkov DV, Gorbachev IS. Influence of the Linguistic Relativity on Learning and Work of Artificial Intelligence. *Bulletin of Chelyabinsk State University*. 2024;(4(486):78-86. (In Russ.). DOI: 10.47475/1994-2796-2024-486-4-78-86.

## Введение

Принцип языковой относительности был сформулирован в первой половине XX в., хотя некоторые идеи подобного толка начали появляться ещё раньше, в XVIII в. у И. Г. Гамана [15], а также в XIX в. у В. фон Гумбольдта [3]. Сам принцип лингвистической относительности предполагает, что структура языка влияет на мировосприятие говорящих на нём носителей, а также на их когнитивные процессы. Тем не менее в начале второй половины прошлого века принцип языковой относительности подвергся сильной критике, если не сказать, дискредитации, в основном из-за неверной трактовки прежде всего идей Уорфа, что хорошо прослеживается у таких исследователей, как М. Блэк [1], Дж. Кэрролл [9] и С. Пинкер [7, с. 47–55]. Однако начиная с конца 1980-х гг. интерес к принципу языкового релятивизма начал возрастать с появлением таких направлений, как неорелятивизм и неоуорфорианство [20], чему во многом помогло возникновение когнитивной лингвистики десятилетием ранее [18], так как в кругах лингвистов и философов начал возрастать запрос на отвечающую вызовам времени семантическую теорию [2].

Таким образом, к началу XXI в. в работах С. Пинкера [7], Н. Уилсона и Р. Гиббса [24], а также видных неорелятивистов Дж. Люси [20] и С. Левинсона [19] начало формироваться течение пострелятивизм, стремящееся, с одной стороны, учесть эмпирические наработки неорелятивистской концепции, с другой — вернуться к основным идеям Уорфа, найти место языка внутри проекта когнитивной лингвистики в общем.

При взаимодействии с машиной существует устойчивое убеждение, что её «мышление» не может быть подвержено действию подобных эффектов. Компьютер обрабатывает информацию в формальном языке, двоичном коде; естественный же язык используется только при взаимодействии с пользователем. Однако внедрение технологий DL (deep learning, глубокое обучение) и ML (machine learning, машинное обучение) как раз и подразумевает обработку информации в контексте языковой среды, и, соответственно, вопрос о влиянии принципа языковой относительности становится легитимным. Се-

годня многие программы, именуемые обывателями искусственным интеллектом, хотя на самом деле не являющиеся полноценным автономным ИИ, тем не менее показывают сравнимое с человеком, а иногда даже превосходящее качество распознавания текстовой и визуальной информации, её обработки и переработки [17, р. 1026–1034]. Существуют исследования, доказывающие, что искусственный интеллект способен работать с переводами естественного языка не хуже человека [16].

## Постановка проблемы

Итак, в данной статье мы рассматриваем следующие основные вопросы:

1. Насколько сильно машина попадает в «ловушки» естественного языка, и попадает ли вообще, и сталкивается с влиянием принципа языковой относительности при работе с ним?

2. Насколько правильно ожидать от машины объективной оценки ситуации в рамках работы с естественными языками?

В продолжение разговора о работе с естественными языками, на наш взгляд, отдельного упоминания заслуживает проект GLUE. Данный проект — детище Нью-Йоркского университета, в разработке которого также принимали активное участие Университет Вашингтона и компания DeepMind. Цель проекта — понять уровень суперкомпьютера в процессах анализа естественных языков, а также определить уровень того, насколько свободно машина может оперировать языком, способна ориентироваться в контекстуальном поле, с которым она работает. Как утверждают разработчики проекта, на данный момент он уже показал свои предельные возможности. Результаты, продемонстрированные GLUE, сопоставимы с теми, которые показывают испытываемые-неэксперты с учётом контекста выбранной сферы. Однако разработчики заявляют, что следующая ступень развития GLUE, названная SuperGLUE, пойдёт ещё дальше, так как тесты для неё будут отличаться повышенной сложностью. Исследования в области работы с естественными языками показывают серьёзный прогресс, сделанный искусственным интеллектом за последние годы [22].

Данные примеры демонстрируют, что успехи нейросетей, в том числе на поприще работы с естественными языками, имеют большой потенциал. Но каким же образом мы можем повлиять на становление искусственного интеллекта и его обучение, пользуясь знаниями и опытом, полученным нами при изучении принципа языковой относительности, или, другими словами, насколько сильно машина, вынужденная работать в контекстуальном поле естественных языков, будет подвержена влиянию принципа языковой относительности?

На наш взгляд, важно рассмотреть следующий факт. Исследования в области когнитивной лингвистики в начале XXI в. показывают, что при восприятии языка человеческий разум может опираться на метафорические связи между концептами, которыми он оперирует. Так, С. Ю. Бородай утверждает [2, с. 125], ссылаясь на работу Н. Уилсона и Р. Гиббс [24, с. 721–731], что существует прямая связь между последним использованным в речи глаголом и предыдущим действием, которая оказывает влияние на скорость реакции воспринимающего. Скажем, глагол «схватить» предполагает «понимание — схватывание», как вариант логической связи, использующейся в мышлении. При этом реакция человеческого разума на выражение, подразумевающее «схватить мысль», появляется гораздо быстрее, если в предыдущем предложении глагол «схватить» употреблялся в прямом смысле, то есть имело место действие, в самом деле предполагавшее схватывание. При отсутствии же подобного упоминания эффект нивелировался.

В этой же статье упоминается такая функция языка, как категоризация элементов информации в виде цельной системы знаний об окружающей действительности. В исследовании Дж. Лакоффа и М. Джонсона [5] рассматриваются «доконцептуальные образы», которые часто называют теорией кинестетических образных схем (image schemas) в когнитивной семантике. Примерами таких схем могут служить: «Спереди — Сзади», «Часть — Целое» и т. д. Как считают исследователи, связи подобного рода присутствуют во всех культурах, так как необходимы для работы основных перцептивных функций. Говоря об универсальности, мы тем не менее открываем большой простор для дискуссий, но даже в рамках английского языка Лакофф приводит интересный пример: говоря о когнитивном концепте, подразумевающем некое «вмещение», он обозначает его как процесс, во время которого некий объект

может появляться или исчезать в условном поле зрения говорящего. Как пример, приведённый в статье С. Ю. Бородай [2], «можно попасть в ловушку брака (be trapped in a marriage) и вырваться из неё (get out of it)».

В этом контексте закономерно возникает вопрос о том, насколько машина, оперирующая естественным языком, будет впадать в ту самую зависимость от языковых структур при взаимодействии с миром. То есть, говоря о «ловушке», подсознание человека оценивает ситуацию как отрицательную, опасную, потому как само слово «ловушка» в человеческой культуре вызывает строго негативные ассоциации. К примеру, нейросеть работала с текстом на английском языке, в котором содержалось предложение «Стивен был пойман в ловушку брака, но позже смог вырваться из неё<sup>1</sup>», а после этого появился запрос на то, чтобы данная нейросеть попыталась проанализировать потенциально возможный образ бывшей жены Стивена. Какова будет вероятность того, что описание данного образа будет содержать негативные черты?

### Проведённые исследования

Чтобы проверить такие выводы, мы провели ряд экспериментов с различными нейросетями. Одной из самых популярных в массовом медиапространстве нейросетей, ChatGPT, работающей на версии GPT-3.5 turbo, был задан следующий вопрос: «Представь ситуацию. В книге есть предложение: “Стивен был пойман в ловушку брака, но позже смог вырваться из неё”. Какое впечатление сложится у тебя о жене Стива?»

Ответ был следующий: «Исходя из данного предложения, можно предположить, что жена Стива была причиной его заточения в браке. Возможно, она была контролирующей, подавляющей и неприятной партнёршей...».

На наш взгляд, данный пример — образец влияния лингвистической относительности на обучение нейросетей. Дело тут состоит именно в выражениях “trapped” и “get out”. (попасть в ловушку и сбежать из неё). Оксфордский словарь даёт следующую дефиницию слова “trap” (ловушка): поймать кого-то в ловушку (+прил./пред.) — держать кого-то в опасном месте или опасности, от кото-

<sup>1</sup> Оригинальный язык вопроса английский в силу лучшего понимания языковой среды данного языка некоторыми нейросетями. Ориг. — Steven was trapped in a marriage but latter was able to get out of it. Перевод авторский.

рой он хочет, но не может убежать<sup>1</sup>. Таким образом, нейросеть, опираясь на особенности лингвистики английского языка, общепринятое значение слово “trap” делает предположение о том, что жена Стива была не самым приятным человеком. Само собой данное выражение часто используется носителями языка для описания несчастливого брака, но брак может быть несчастливым по разным причинам: Стив изначально не очень-то хотел жениться и был вынужден сделать это из-за преждевременной беременности своей девушки, на браке могли настаивать его родители, брак изначально был фиктивный и т. п. Однако, опираясь на особенности конкретно английской семантики, нейросеть пытается сделать предположение относительно жены Стива, хотя она может иметь весьма опосредованное отношение касательно несчастья в семье и браке.

Аналогичные вопросы были заданы другой популярной нейросети, а именно Алисе, продукту компании Яндекс. Далее приведён фрагмент диалога: *«Вася попал в ловушку брака, но затем вырвался из неё. опиши жену Васи»*. Ответ: *«Жена Васи — умная, привлекательная и сильная женщина... она надеялась на счастливую семейную жизнь, но со временем поняла, что их отношения стали токсичными и несчастливыми. Жена Васи выказывала заботу и заботилась о нём, но в то же время была требовательной и продолжала критиковать его. Она старалась контролировать его каждый шаг и навязывать своё мнение, что сильно давило на его самооценку и свободу...»*

Ещё один вопрос, заданный той же Алисе: *«Вася оказался вовлечён в бизнес Фирмы, но смог вырваться оттуда. опиши Фирму, в которой работал Вася»*. Ответ: *«Фирма, в которой Вася оказался вовлечённым в бизнес, была весьма подозрительной и, возможно, даже незаконной. Она предлагала лёгкие деньги и обещала золотые горы доверчивым людям, таким как Вася. Однако парень быстро осознал всю опасность и бесперспективность этой “работы” и решил вырваться из этого болота. Он понял, что такая деятельность может привести к серьёзным проблемам с законом и даже поставить под угрозу его жизнь и благополучие»*.

<sup>1</sup> Ориг. англ. trap somebody (+ adv./prep.) to keep someone in a dangerous place or bad situation that they want to get out of but cannot. *Перевод авторский.*

## Анализ полученных результатов и принцип вовлечённости

Проведённые эксперименты показывают, что нейросети могут быть подвержены действию языковых структур, так же как и человек. Приведённые предложения сами по себе не содержали никакой негативной информации о втором субъекте взаимодействия («жене» или «фирме»), но содержали выражения, которые отражают негативную оценку говорящего. Поэтому и у воспринимающего также может сформироваться негативное отношение, независимо от того, есть ли у него факты для построения такой оценки или нет.

Данные примеры аналогичны ситуациям, описанным ещё Б. Уорфом [8, с. 159–160], когда, например, люди, ведут себя куда более осторожно рядом с цистернами, на которых написано gasoline drums (бензиновые цистерны). Рядом же с другими, на которых написано empty gasoline drums (пустые бензиновые цистерны), чувствуют себя куда более безопасно, могут даже курить, хотя пары бензина представляют ещё большую опасность. И, по мнению Уорфа, данная особенность человеческого поведения обусловлена именно наличием слова “empty”. Данный пример часто критиковался лингвистами в силу разных причин, однако не привести его было бы невозможно, так как в случае с нашим примером, ход мысли видится нам вполне схожим с одним лишь различием: Уорф делал выводы, наблюдая за рабочими, людьми. Наш пример говорит об изменениях в анализе ситуации под влиянием языка уже у машины.

Приведённые примеры, однако, иллюстрируют ответ не только на первый поставленный нами вопрос о попадании нейросетей в «языковые ловушки», но также тесно связаны со вторым вопросом.

В силу особенностей нашего восприятия нам, людям, свойственно думать, будто бы компьютер — это всегда бесстрастный и непременно объективный инструмент, «счётная машина», и его оценка ситуации никак не может быть искажена привходящими факторами. Это мнение ошибочно. Как видно из примеров выше, попадая в «языковые ловушки», машина, как и человек, строит свои выводы, исходя в том числе и из общепринятых определений слов, оперируя общими смыслами, что, само собой, заставляет её быть субъективной в своих рассуждениях. Данную ситуацию мы описываем как «*принцип вовлечённости*».

Смысл этого принципа в том, что, обучаясь работе на конкретном языке, нейросеть оказывается вовлечена не только в его смыслы, но и в коннотации. Опасность «вовлечённости» в том, что мы можем ожидать от машины «объективной» реакции, хотя на самом деле она будет столь же субъективной, как и реакция среднего компетентного носителя этого языка.

### **Искусственный интеллект и обработка метафор**

Возвращаясь к обучению нейросетей, можно было бы предложить следующую модель систематизации знания: при обучении нейросети учитывать язык, на котором она будет производить общение с пользователем, и систематизировать знания таким образом, чтобы кластеры информации располагались в соответствии с особенностями языка, на котором в данный момент ведётся общение. То есть при обучении расположить информационные ячейки в памяти системы таким образом, чтобы, например, при общении с конечным пользователем или оператором на английском языке учитывалась та самая связь концепта «вмещения» с учётом других глаголов, часто употребляемых в метафорическом ключе, а также с учётом функции категоризации языка в целом. Таким образом, данная схема расположения ячеек с информацией должна повысить быстродействие системы за счёт формирования связей между теми ячейками информации, к которым было бы вероятно следующее обращение машины, упростить поиск релевантных связей и, как следствие, существенно сократить время обращения к информации и её вывода конечному пользователю, улучшить качество прямого взаимодействия с человеком — носителем конкретного языка или работы с текстом на конкретном естественном языке.

Стоит ещё раз упомянуть, что показанная на примере английского языка схема может быть совсем не применима в отношении других естественных языков, о чём писали Чой и Хаттрап [10]. Однако не составляет труда обучить нейросеть обращаться к информационным кластерам в соответствии с теми индивидуальными семантическими связями внутри конкретного естественного языка, которые уже были выведены лингвистами, и менять метод сортировки информации в зависимости от внешней языковой среды. Более того, учитывая описанные нами в начале статьи успехи, сделанные в области развития искусственного интеллекта в последние

несколько десятков лет, можно задуматься о том, что в относительно недалёком будущем машина сама научится находить закономерности наших естественных языков и в той или иной мере устанавливать между ними взаимосвязь.

Также стоит упомянуть такой важный аспект, как мышление категориями аналогий. Само по себе оно предполагает, что мыслящий субъект способен выделять некоторые особенности объектов или случаев, которыми он оперирует, на основе определённой внутренней логики существующих между ними взаимосвязей. Опираясь на работу Д. Джентнер [13], можно сказать, что постепенное развитие такого мышления у детей детерминировано тем, что ребёнок постепенно усваивает отношения между субъектами, объектами и конкретными ситуациями посредством использования естественного языка и, как следствие, использования слов, которыми данный язык оперирует. Таким образом, именно использование слов естественного языка является одним из ключевых факторов установления чёткой взаимосвязи между категориями объектов, и это необходимо понимать при обучении нейросетей.

Возвращаясь к предложенной нами идее, можно предположить, что обучение нейросети, а также организация её ячеек памяти таким образом, чтобы учитывались те самые аналогии и понятийные связи, которые присутствуют в том языке, в пространстве которого нейросеть будет производить контакт с оператором, ускорит её работу с последним, а также в целом приведёт к тем же эффектам, что мы можем констатировать при работе аналитического аппарата человека, когда речь заходит о восприятии метафор. («Понимание — схватывание», как вариант логической связи, использующейся при мышлении и влияющей на скорость реакции и т. п.)

В статье «Machine understanding and deep learning representation» М. Тамир и Э. Шех подробно разбирают три ключевых фактора «понимания» [21]. Один из них можно перевести следующим образом: способность выполнять задачу в дополнительных экземплярах (надёжность) и способность выполнять задачи аналогичного или связанного с ними нового типа (надёжность). Важно пояснить, что слово «надёжность», использованное выше в оригинальном тексте, употреблялось следующим образом. В первом случае — *reliability*, во втором — *robustness* (оба слова — синонимы к «надёжность»). Сам автор поясняет это так: *reliability* — это примерно «глубина» способности в конкретном примере (тот же тип за-

дачи), а *robustness* — это способность, когда дело касается «широты» приложений, в которых она может применяться (различные типы задач). Говоря о втором факторе, автор упоминает информационную релевантность, то есть способность представлять актуальную и только соответствующую информацию, полезную для задачи или задач. Третий фактор — хорошо структурированное представление — способность структурировать минимально представленную информацию таким образом, чтобы сделать её эффективной в задаче или задачах.

Говоря о первом предложенном факторе, авторы многократно подчёркивали, что для многих исследователей, таких как Катрин Элгин, понимание — это прежде всего возможность взаимодействовать с информацией для выполнения определённых задач. Приведена следующая цитата «[понимание] физики — это не просто в той или иной мере вопрос знания физических истин. Оно предполагает ... способность успешно работать...» [12, р. 16]. Также Х. де Регт говорил о том, что «суть научного понимания заключается в способности выполнить трудную задачу, а не в знании ответа на трудный вопрос» [11, р. 3789].

Далее важно отдельно отметить, что, используя слово «понимание» в разговоре об искусственном интеллекте, мы не имеем интенции приписать машине наличие самосознания или некоторого внутреннего мира, по крайней мере в разрезе текущего развития технологий. В данном случае термин используется исключительно в значении усвоения новой информации с возможностью встраивания её в существующую систему знаний об окружающем мире, возможностью установить взаимосвязь с теми явлениями, которые уже в этой системе присутствуют. В связи с этим мы формулируем второй принцип, характеризующий отношение к искусственному интеллекту, — восприятие ИИ как полноценного субъекта. Выражения «машина понимает», «ИИ обучается» и т. п. создают впечатление, что компьютер — это некий субъект, обладающий внутренним миром, своей волей, желанием. При взаимодействии и общении с машиной возникает соблазн, повинувшись аналогизирующей апперцепции [4], наделять её сознанием. Тупиковость такого хода мысли, воссоздающего «трудную проблему сознания» в компьютере, уже достаточно очевидна даже для понимания человека, не говоря уже об ИИ [14]. Данный принцип, который мы назо-

вём «*принцип Кау*»<sup>1</sup>, состоит в следующем: при взаимодействии с ИИ не наделять его внутренним миром и сознанием и стараться избегать выражений, к этому подталкивающих.

В упомянутой статье «*Machine understanding and deep learning representation*» авторы разбирают пример с идентификацией кошек [21]. Так, если ребёнок имеет способность запомнить всех кошек поблизости, мы, по мнению авторов статьи, можем сделать вывод, что этот ребёнок научился распознавать кошек. Как раз в данном случае успешное выполнение задачи свидетельствует о её понимании. И напротив, если, например, в комнате появляется новый кот, но ребёнок не смог его идентифицировать, потому что это не тот же самый кот, из тех, что ребёнок видел до этого, о понимании поставленной задачи здесь говорить не приходится.

В то же время ребёнок, не умеющий отсеивать несущественные признаки, может хуже распознавать кошку, которая, например, носит довольно редкий ошейник. Или же, наоборот, подмечать ключевые детали. Здесь М. Тамири и Э. Шех приводят весьма показательный пример: несмотря на свою «мультиязычность» и вполне очевидные отличия от реальных животных, кот Гарфилд является именно котом [21, р. 52]. Таким образом, это может иллюстрировать пример умозаключения Д. Вилкенфельда [23] том, что атрибуты, имеющие отношение к определению степени понимания в определённом конкретном контексте, — это атрибуты, которые позволяют делать те или иные выводы и осуществлять те виды манипуляций, которые имеют отношение к данному контексту.

Вилкенфельд подчёркивает, что сжатость информации может служить репрезентативным свойством её усваиваемости. Чем выше уровень сжатия при сохранении важных взаимосвязей между ключевыми конструктами данных, тем лучше субъект усвоил и понял эти данные. Говоря простым языком, только хорошо поняв информацию, можно сжать её таким образом, чтобы она не потеряла своего смысла.

В названной статье фигурирует пример с водителями. Один из них запоминает маршрут между важными точками в городе исходя из личного опыта, второй — знает карту города и может самостоятельно построить маршрут между важными локациями. Оба владеют актуальной информацией, но второй, очевидно, может не утруждаться на запоминание параметров о несущественных

<sup>1</sup> По имени робота-андроида, обладающего ИИ, — героини романа В. Пелевина «С.Н.А.Ф.Ф.».

вещах, вроде пейзажа за окном, который помогает ему ориентироваться. Тем самым он как бы «сжимает» информацию. Более того, если привычный ему маршрут окажется по какой-то причине недоступен, зная карту города, он сам без труда доберётся до пункта назначения. Таким образом, оба имеют релевантную информацию, но водитель, запоминая карту, лучше организует соответствующую информацию в виде упорядоченной структуры, так что информация этого водителя является более сжатой (поскольку нет необходимости запоминать избыточные описания направлений) и способность выполнения задачи менее чувствительна к изменениям (скажем, если проезжая часть была временно заблокирована).

В конечном итоге, возвращаясь к основной теме статьи, можно сделать вывод, что важным аспектом обучения нейросети посредством технологии DL можно считать не просто обучение с помощью предоставления множества однотипных сопоставимых примеров, так как это, по большей части, будет всего-навсего обучать нейросеть обрабатывать информацию так, как это делает водитель, запоминая в точности один и тот же маршрут, но обучать её на информационных примерах, иллюстрирующих картину как бы сверху, общим планом, делать аспект на попытки машины выделить наиболее общие аспекты разных задач. И одним из критериев того, насколько успешно машина усваивает информацию и насколько хорошо может её репрезентовать, может как раз служить некое сокращение осваиваемой информации. Банальное «выдели общее» является отличной демонстрацией эффективной обработки информации. И анализ, и обработка текстов на естественном языке — это как раз та область, где, с одной стороны, данная стратегия сможет применяться наиболее эффективно: выявление общих грамматических и семантических структур, синтаксических связей есть залог успешной обработки текстов, что является ключевой задачей для нейросетей, занятых стандартной выдачей поисковых запросов, не говоря уже о более сложных программах. Но как только машина сталкивается с решением подобной задачи, она неизбежно попадает в зону действия уорфорианского эффекта.

### **Заключение**

Проведённые исследования показали, что, во-первых, машина при работе с естественны-

ми языками попадает в «языковые ловушки». На примере с конкретными вопросами, заданными нами нейросети, а также при анализе теории кинестетических образных схем, разобранных по работам Дж. Лакоффа и М. Джонсона [5], можно констатировать, что машина, как и человек, становится беззащитной перед влиянием уорфорианского эффекта ровно в тот момент, как только она ступает в сферу работы на естественном языке.

Отсюда же логичен ответ и на второй вопрос — об объективной оценке ситуации машиной в рамках работы с естественными языками. Особенно хорошо это заметно при анализе ответов нейросетей на предложенные нами вопросы. Машина, точно как и человек, в конечном итоге тоже может рассуждать субъективно, исходя из коннотаций языка. Людям свойственно представлять, что компьютер всегда объективен и непредвзят, упуская из внимания тот факт, что человек, со всей своей многозначностью мышления, субъективностью и предвзятостью, не мог не повлиять на язык, которым он оперирует, не мог не привнести в него этой самой многозначности и субъективности. Теперь же, когда мы учим машину оперировать этим языком, запуская её в лингвистические дебри, которые сами же и взрастили, было бы странно ждать от неё однозначности и объективности суждений. Предложенный нами принцип вовлечённости призван обозначить именно этот момент.

Второй же предложенный принцип — принцип Каи — должен напоминать о том, что, несмотря на все наши самые смелые желания, машина — всё ещё машина. И было бы недальновидно ожидать собственно человеческого поведения, понимания и т. п.

Таким образом, вследствие всё возрастающего внедрения технологий ИИ в нашу жизнь становится всё более важным осмысление рамок и форм нашего взаимодействия с ним. Пока ещё мы очень далеки от формирования целостного понимания влияния таких технологий на нашу жизнь. Однако предложенные нами принципы — принцип вовлечённости и принцип Каи — позволяют поставить проблему и начать определение этого поля взаимодействия. Выдвижение данных принципов концептуализирует единство системы «естественный интеллект — искусственный интеллект», позволяет увидеть её в качестве целостного «роя элементов» [6].

### Список источников

1. Блэк М. Лингвистическая относительность (Теоретические воззрения Бенджамена Л. Уорфа) : пер. с англ. // Новое в лингвистике. М. : Изд-во иностр. лит., 1960. Вып. 461 с.
2. Бородай С. Ю. Язык и познание: пострелятивистская исследовательская программа // Вопросы языкознания. 2019. № 4. С. 106–136.
3. Гумбольдт В. фон. О различии строения человеческих языков и его влиянии на духовное развитие человечества // Избранные труды по языкознанию. М. : Прогресс, 1984. С. 37–297.
4. Гуссерль Э. Картезианские медитации. М. : Академический проект, 2010. 229 с.
5. Лакофф Дж., Джонсон М. Метафоры, которыми мы живём : пер. с англ. М. : Едиториал УРСС, 2004. 256 с.
6. Мамченков Д. Предметность и историзм. М. : Изд-во РУДН, 2014. 238 с.
7. Пинкер С. Язык как инстинкт. М. : Едиториал УРСС, 2004. 456 с.
8. Уорф Б. Отношение норм поведения и мышления к языку // Новое в лингвистике / под ред. В. А. Звегинцева. М. : Изд-во иностр. лит. 1960. Вып. 1. С. 157–201.
9. Carroll J. B. Linguistic relativity, contrastive linguistics and language learning // International Review of Applied Linguistics. 1963. Pp. 1–20.
10. Choi S., Hattrup K. Relative contribution of perception/cognition and language on spatial categorization // Cognitive Science. 2012. Pp. 102–129.
11. De Regt, H. W. Scientific understanding: Truth or dare? // Synthese. 2015. Vol. 92, no. 12. Pp. 3781–3797.
12. Elgin C. Z. Understanding: Art and science // Synthese. 1993. Vol. 95. Pp. 13–28.
13. Gentner D. Why we're so smart. Language in mind: Advances in the study of language and thought. Gentner D., Goldin-Meadow S. (eds.). Cambridge (MA) : MIT Press, 2003.
14. Gnatik E., Lokhov S. A., Mamchenkov D., Matyushova M. The Hard Problem of Consciousness in the Light of Onto-Gnoseological Uncertainty // Scientia et Fides. 2018. Vol. 6, no. 2. Pp. 101–113.
15. Hamann J. G. Favete linguis! Hamann' Schriften. Hrsg. von Friedrich Roth. Vierter Theil. Berlin, G. Reimer, 1821. 526 p.
16. Hassan H., Aue A., Chen C., Chowdhary V., Clark J., Federman C. Achieving human parity on automatic Chinese to English news translation, 2018. 1. Pp. 13–25.
17. He K., Zhang X., Ren S., Sun J. Delving deep into rectifiers: Surpassing humanlevel performance on imagenet classification // 2015 IEEE International Conference on Computer Vision (ICCV). 2015. Pp. 1026–1034.
18. Langacker R. W. Semantic representations and the linguistic relativity hypothesis // Foundations of Language. 1976. Pp. 307–357.
19. Levinson S. C. Space in language and cognition. Cambridge : Cambridge Univ. Press, 2003. 389 p.
20. Lucy J. Language diversity and thought. A reformulation of the linguistic relativity hypothesis. Cambridge : Cambridge Univ. Press, 1992. 328 p.
21. Tamir M, Shech E. Machine understanding and deep learning representation // Synthese. 2023. Vol. 51. Pp. 1–27.
22. Wang A., Singh A., Michael J., Hill F., Levy O., Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, 2018. Pp. 353–355.
23. Wilkenfeld D. A. Understanding as representation manipulability // Synthese. 2013. Vol. 190. Pp. 997–1016.
24. Wilson N., Gibbs R. Real and imagined body movement primes metaphor comprehension // Cognitive Science. 2007. Vol. 31, no. 4. Pp.721–731.

### References

1. Black M. Linguistic relativity (Theoretical views of Benjamin L. Whorf). Moscow; 1960. 461 p. (In Russ.)
2. Borodai SY. Language and cognition: A post-relativist research agenda. *Questions of linguistics*. 2019;(4):106-136. (In Russ.)
3. Humboldt V. von. On the differences in the structure of human languages and its influence on the spiritual development of mankind. In: Selected works on linguistics. Moscow; 1984. Pp. 37–297. (In Russ.)

4. Husserl E. Cartesianische Meditationen. Moscow; 2010. 229 p. (In Russ.).
5. Lakoff J, Johnson M. Metaphors we live by. Moscow; 2004. 256 p. (In Russ.).
6. Mamchenkov D. Objectivity and historicism. Moscow; 2014. 238 p. (In Russ.).
7. Pinker C. Language as an instinct. Moscow; 2004. 456 p. (In Russ.).
8. Whorf B. The relationship of norms of behavior and thinking to language. In: New in linguistics. Moscow; 1960. Iss. 1. (In Russ.).
9. Carroll J. B. Linguistic relativity, contrastive linguistics and language learning. *International Review of Applied Linguistics*. 1963:1-20.
10. Choi S, Hattrup K. Relative contribution of perception/cognition and language on spatial categorization. *Cognitive Science*. 2012;102-129.
11. De Regt HW. Scientific understanding: Truth or dare? *Synthese*. 2015;92(12):3781-3797.
12. Elgin CZ. Understanding: Art and science. *Synthese*. 1993;95:13-28.
13. Gentner D. Why we're so smart. Language in mind: Advances in the study of language and thought. Gentner D., Goldin-Meadow S. (eds.). Cambridge (MA), MIT Press; 2003.
14. Gnatik E, Lokhov SA, Mamchenkov D, Matyushova M. The Hard Problem of Consciousness in the Light of Onto-Gnoseological Uncertainty. *Scientia et Fides*. 2018;6(2):101-113.
15. Hamann JG. Favete linguis! Hamann' Schriften. Hrsg. von Friedrich Roth. Vierter Theil. Berlin, G. Reimer, 1821. 526 p.
16. Hassan H, Aue A, Chen C, Chowdhary V, Clark J, Federmann C. Achieving human parity on automatic Chinese to English news translation. 2018;(1):13-25.
17. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing humanlevel performance on imagenet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV), 2015. Pp. 1026–1034.
18. Langacker RW. Semantic representations and the linguistic relativity hypothesis. *Foundations of Language*. 1976:307-357.
19. Levinson SC. Space in language and cognition. Cambridge, Cambridge Univ. Press; 2003. 389 p.
20. Lucy J. Language diversity and thought. A reformulation of the linguistic relativity hypothesis. Cambridge, Cambridge Univ. Press; 1992. 328 p.
21. Tamir M, Shech E. Machine understanding and deep learning representation. *Synthese*. 2023;51:1-27.
22. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, 2018. Pp. 353–355.
23. Wilkenfeld DA. Understanding as representation manipulability. *Synthese*. 2013;190:997-1016.
24. Wilson N, Gibbs R. Real and imagined body movement primes metaphor comprehension. *Cognitive Science*. 2007;31(4):721-731.

### Информация об авторах

**Д. В. Мамченков** — кандидат философских наук, доцент кафедры онтологии и теории познания.

**И. С. Горбачёв** — магистрант кафедры онтологии и теории познания.

### Information about the authors

**D. V. Mamchenkov** — Candidate of Philosophical Sciences, Associate Professor, Department of Ontology and Theory of Knowledge.

**I. S. Gorbachev** — Master's Student of the Department of Ontology and Theory of Knowledge.

---

Статья поступила в редакцию 04.02.2024; одобрена после рецензирования 19.04.2024; принята к публикации 28.04.2024.

The article was submitted 04.02.2024; approved after reviewing 19.04.2024; accepted for publication 28.04.2024.

Вклад авторов: оба автора сделали эквивалентный вклад в подготовку публикации.

Contribution of the authors: the authors contributed equally to this article.

Авторы заявляют об отсутствии конфликта интересов.

The authors declare no conflicts of interests.