DOI 10.22394/1726-1139-2024-5-150-158

EDN UCLTCT

Проблема предвзятости нейросетей: конфликтные и этические вызовы*

Сунами А. Н., Мусаев А. И.*

Санкт-Петербургский государственный университет, Санкт-Петербург, Российская Федерация; *a.sunami@spbu.ru

РЕФЕРАТ

Статья посвящена анализу этических и конфликтных вызовов, связанных с проблемой предвзятости нейронных сетей. Постулируется необходимость корректной, научно обоснованной экспликации феномена предвзятости с целью построения корректировочных данную проблему моделей, как необходимого элемента процесса разработки программных продуктов, построенных на алгоритмах искусственного интеллекта.

Рассматривается история развития нейронных сетей от зарождения идеи механического организма до построения современных моделей искусственного нейрона. Выделяются наиболее существенные характеристики современных нейросетей: архитектура, веса и смещения, функции активации, инференции, способы обучения.

Дается развернутая характеристика естественного языка как ресурса обучения нейросети, анализируется программирование на естественном языке. Особо подчеркивается специфика естественного языка нейросети как совокупности языковых практик, отражающих весь оцифрованный опыт человечества, включая стереотипы, неравенства, язык вражды и другие феномены, в конечном счете, продуцирующие проблему предвзятости.

Значительное внимание уделяется анализу феноменов «политика классификации», «властный дискурс», «культурное насилие» в контексте поиска методологических оснований стратегий фильтрации и цензуры естественного языка в процессе конструирования нейронной сети.

Отдельно подчеркиваются как имеющиеся в нейросетях погрешности отражены в существующих этических и конфликтологических дебатах вокруг проблемы искусственного интеллекта. Делается вывод, что текущая оценка моральных аспектов проблемы не предполагает наделение нейросетей статусом морального агента и возлагает этическую экспертизу продукта на его разработчиков. Особо отмечается, что конфликтный аспект проблемы предвзятости заключается в ее признании исключительно в отношении групп, которые к настоящему времени приобрели в результате социальных конфликтов «чувствительный» статус дискриминируемых

В заключение обосновывается острая необходимость оптимизации искусственного интеллекта в целях снижения деструктивного потенциала проблемы предвзятости, что в обязательном порядке предполагает модификацию социальных отношений в более широком контексте борьбы исключенных групп за право признания в качестве дискриминируемых.

Ключевые слова: нейросеть, проблема предвзятости, этика, конфликт, политика классификации, культурное насилие

Для цитирования: *Сунами А. Н., Мусаев А. И.* Проблема предвзятости нейросетей: конфликтные и этические вызовы // Управленческое консультирование. 2024. № 5. С. 150–158.

«The Trouble with Bias» in Neural Networks: Conflict and Ethical Challenges

Artem N. Sunami, Abdurashid I. Musaev*

Saint-Petersburg State University, Saint Petersburg, Russian Federation; *a.sunami@spbu.ru

* Исследование выполнено за счет гранта Российского научного фонда № 23-28-00220, https://rscf.ru/project/23-28-00220.

ABSTRACT

This article is devoted to the analysis of ethical and conflict challenges related to the trouble with bias in neural networks. The necessity of a correct, scientifically based explication of the phenomenon of bias is postulated in order to build models correcting this problem as a necessary element of the software development process based on artificial intelligence algorithms. The history of the development of neural networks is considered from the origin of the idea of a mechanical organism to the construction of modern models of an artificial neuron. The most significant characteristics of modern neural networks are highlighted: architecture, weights and offsets, activation functions, inferences, and learning methods.

A detailed description of natural language as a neural network learning resource is given, and programming in natural language is analyzed. The specificity of the natural language of the neural network as a set of linguistic practices reflecting the entire digitized experience of mankind, including stereotypes, inequalities, hate speech and other phenomena, ultimately producing the trouble with bias, is emphasized.

Considerable attention is paid to the analysis of the phenomena of "politics classification", "power discourse", "cultural violence" in the context of the search for methodological foundations of natural language filtering and censorship strategies in the process of constructing a neural network.

Separately, it is emphasized how the errors in neural networks are reflected in the existing ethical and conflict studies debates around the problem of artificial intelligence. It is concluded that the current assessment of the moral aspects of the problem does not imply granting neural networks the status of a moral agent and places the ethical expertise of the product on its developers. It is particularly noted that the conflict aspect of the trouble with bias lies in its recognition exclusively in relation to groups that have now acquired the "sensitive" status of discriminated against as a result of social conflicts.

In conclusion, the paper substantiates the urgent need to optimize artificial intelligence in order to reduce the destructive potential of the trouble with bias, which necessarily implies the modification of social relations in the broader context of the struggle of excluded groups for the right to be recognized as discriminated against.

Keywords: neural network, the trouble with bias, ethics, conflict, classification politics, cultural violence

For citing: Sunami A. N., Musaev A. I. «The Trouble with Bias» in Neural Networks: Conflict and Ethical Challenges // Administrative Consulting. 2024. N 5. P. 150–158.

Введение

Стремительное развитие нейросетей и все более широкое внедрение в социальную практику разнообразных продуктов, основанных на алгоритмах искусственного интеллекта, ставит множество занятных вопросов перед академическим сообществом. Но, если начальный период разработки этой проблематики изобиловал вопросами скорее технического характера, то в последнее десятилетие растет доля социальных и этических тем, поднимаемых в связи с реализацией нейросетевых продуктов и применением их в решении различных задач. Одной из наиболее волнующих тем, в этой связи, является проблема «предвзятости» (the trouble with bias).

Целью данной статьи является анализ этических и конфликтных вызовов, связанных с проблемой предвзятости нейронных сетей, проводимый посредством применения методологии теории конфликта к исследованию социальных аспектов функционирования искусственного интеллекта. Предметом нашего исследования выступает конфликтное измерение проблемы предвзятости нейросетей, которое формируется в результате получения той или иной группой статуса дискриминируемой. Методология исследования основана на концептуальных подходах М. Фуко и Й. Галтунга, в сконцентрированном виде представленных в нарративах властного дискурса и структурного насилия. Достижение поставленной цели обеспечивается решением следу-

ющих задач: 1) рассмотрение истории развития нейронных сетей; 2) выявление специфики естественного языка нейросети как совокупности языковых практик, отражающих весь оцифрованный опыт человечества и обуславливающих проблему предвзятости; 3) анализ феноменов «политика классификация», «властный дискурс», «культурное насилие» в контексте поиска методологических оснований стратегий цензурирования естественного языка в процессе разработки нейронной сети; 4) обоснование тенденциозности работы нейросетей, которая транслируется в этических и конфликтологических дебатах по мотивам проблемы искусственного интеллекта.

Нейросети: история и принципы работы

Прежде чем перейти непосредственно к раскрытию темы настоящей статьи, нам необходимо сделать небольшой обзор того, что есть такое нейросети, какова их история и основные принципы работы. Корни идеи создания искусственного организма, вероятно, уходят еще в эпоху античности, что, в частности, выразилось в таком культурном феномене, как «голем», бытовавшем на уровне религиозно-мистических конструкций и народного фольклора. Но в научном и философском смысле эта идея была сформулирована Рене Декартом, отождествившем организм и механизм, что фактически означало, что если живой организм является сложным механизмом, то в теории его можно искусственно воспроизвести. Как мы знаем, первые механические вычислительные машины появляются уже в Новое время, но все-таки мы скорее воспринимаем их как некие приборы, но не мозг как таковой, пусть даже в примитивном его воплощении. Как правило, историю нейросетевого принципа ведут с не столь древних времен. Первой моделью стала математическая модель нейрона Уоррена Мак-Каллока и Уолтера Питтса, построенная в 1943 г. Авторы данной модели показали, что нейронные сети живого организма работают на основаниях препозиционной логики, т. е. работа мозга может быть описана в логико-математических терминах [16]. В течение последующих десятилетий, отталкиваясь от этого заключения, последовал целый комплекс открытий в математике и нейрофизиологии, было решено множество парадоксальных задач, созданы все более совершенные вычислительные мощности, что в совокупности позволило в последние годы осуществить настоящий прорыв в этой области. Масштаб этого прорыва таков, что его последствия могут ощутить на себе не только специалисты в разных областях академической науки, но и обычные люди, которые в настоящее время могут сами использовать те или иные нейросетевые продукты в решении бытовых вопросов, рабочих задач или в развлекательных целях.

Не вдаваясь в подробности, можно описать нейросети как математическую модель саму по себе или же ее программную реализацию, построенную на принципах аналогии с нейронными клетками биологического организма и их сетями. Каждая нейросеть состоит из нескольких компонентов:

- архитектура (один или несколько слоев нейронов, объединенных в определенной последовательности);
- взаимодействие между нейронами посредством обмена и передачи по цепочке сигналов;
- веса и смещения, определяющие важность каждого входного сигнала в нейронах, помогающие нейросети модифицироваться и корректировать замеченные ошибки;
- функция активации, работающая по нелинейному принципу и обуславливающая сложность выполняемых решений;
- инференция, позволяющая предсказывать новые данные;
- обучение на наборе данных, где для каждого входа есть правильный выход.

Последний компонент имеет самое принципиальное значение в свете проблемы, рассматриваемой в настоящей статье.

Итак, нейросеть может работать только после прохождения обучения. Что означает этот процесс? В статье Андрея Созыкина дается широкий обзор методов обучения глубоких нейросетей, в котором это обучение рассматривается как «процесс определения весов соединений между нейронами таким образом, чтобы сеть приближала необходимую функцию с заданной точностью» [6, с. 31]. И далее классифицируются три способа этого обучения: обучение с учителем (supervised learning), обучение без учителя (unsupervised learning) и обучение с подкреплением (reinforcement learning) [18]. Для нас наиболее важным является вопрос, что является своеобразным «топливом» для обучения нейронной сети. В зависимости от специфики нейросети это могут быть специально подготовленные под решаемую задачу обучающие наборы данных, открытые датасеты, содержащие пакеты кодированной информации, а также массив данных из Интернета. Для примера можем взять один из наиболее популярных в настоящее время продуктов — ChatGPT от компании OpenAI. В докладе, посвященном языковой модели GPT-3, представленной в 2020 г., описываются датасеты, которые выступили ресурсами для обучения ChatGPT:

- открытый репозиторий веб-данных Common Crawl (60%);
- тексты веб-страниц, заслуживших одобрение пользователями Reddit (22%);
- два корпуса книг, размещенных в Интернете (16%);
- статьи англоязычной Википедии (3%) [10, с. 8].

Как видно, специфика данных ресурсов релевантна задаче использованию нейросети языка, максимально приближенного к реальной коммуникации людей в Интернете, иными словами, к «естественному языку». В свою очередь, проблема естественного языка является одной из наиболее важных для тематики обучения нейронных сетей, например, в конструировании диалога машины и человека [2], более того, сам процесс разработки нейросети часто формулируется как программирование на естественном языке (*англ.* natural-language programming (NLP)).

Проблема естественного языка

Проблема естественного языка является одной из важнейших тем в философии языка и лингвистике. Мы не будем подробно останавливаться на академической рефлексии естественного языка, ибо этот вопрос довольно-таки подробно разобран в научной литературе. В этой связи можно сослаться на работы Екатерины Григоренко, которая резюмирует, что «естественный язык формируется народом, учитывая его чувства, стремления и традиции; естественный язык формируется каждым человеком, основываясь на его восприятии, чувствах, переживаниях и мыслях» [1, с. 30]. Если попытаться доступно переформулировать, что есть такое проблема программирования на естественном языке, задача может быть поставлена следующим образом: максимальное сближение языка машины и естественного языка коммуникации людей, что «требует знаний о мире и здравом смысле» [12, с. 110]. Рассмотрение дефиниции естественного языка Екатерины Григоренко через оптику настоящей статьи, позволяет фундировано предположить, что предубеждения, стереотипы, маркировки и прочее в тех или иных формах также являются неотъемлемой частью естественного языка. Таким образом, вполне закономерно, что чем ближе разработчик нейросетевой программы подходит к решению проблемы естественного языка, тем конкретнее в результатах ее работы будут отражаться черты, характерные для текущего социального дискурса в целом или какого-либо его сегмента, в том числе и те неприглядные элементы, от которых мы стараемся дистанцироваться и сделать вид, что их не существует. Более того, необходимо отдавать себе отчет, что естественный язык не синхроничен актуальному языку, в том смысле, что охватывает не только текущие языковые практики. Совершенно очевидно, что для обучения нейросети может быть использован весь оцифрованный опыт человечества, а значит, табуированные в настоящее время

в соответствии со стратегией инклюзивности языковые конструкции, которые не так давно широко использовались и были неотъемлемой частью актуального языка, обретают новую жизнь в качестве ресурса для обучения нейросетей [14].

Проблема «предвзятости нейросетей»

Таким образом, мы имеем все пререквизиты, чтобы подступиться к «проблеме предвзятости». Целый ряд скандалов, в частности, приобретший нарицательный характер инцидент с компанией Amazon, когда обученная для подбора персонала нейросеть, основываясь на предыдущем оцифрованном опыте, дискриминировала кандидатов по гендерному признаку, привел к тому, что теперь каждая значительная команда разработчиков продуктов с использованием искусственного интеллекта с необходимостью должна иметь в штате так называемого специалиста по «анализу этических последствий». В докладе по GPT-3, к которому мы уже обращались выше, проблема предвзятости сформулирована так: «предубеждения, присутствующие в обучающих данных, могут привести к тому, что модели будут генерировать стереотипный или предвзятый контент <...> это вызывает беспокойство, поскольку предвзятость моделей может по-разному навредить людям в соответствующих группах, закрепляя существующие стереотипы» [10, с. 38].

Кейт Кроуфорд, известный специалист по искусственному интеллекту и проблеме предвзятости в частности, к работам которой мы обратимся ниже, заявляет, что еще десятилетие назад мнение о предвзятости искусственного интеллекта выглядело как попытка соригинальничать, в настоящее время же многочисленными примерами дискриминационных систем вряд ли кого-то можно удивить. Владимир Путин, выступая в конце 2023 г. на конференции «Artificial Intelligence Journey 2023», выразил опасение, что из искусственного интеллекта, созданного по некоторым западным стандартам и лекалам, может получиться ксенофоб [3].

Если проблема предвзятости возникает по причине того, что предвзят сам естественный язык, на котором обучается машина, очевидный вывод состоит в том, что и она в свою очередь наследует черты некой «естественности». Тем не менее такая естественность не отменяет необходимости нахождения решения. Очевидно, что такое решение лежит в плоскости фильтрации и цензуры естественного языка на каком-то этапе конструирования нейронной сети. Каков должен быть базовый механизм этого фильтра?

Вышеупомянутая Кейт Кроуфорд полагает, что он должен быть связан с регуляцией «политики классификации». «Практика классификации определяет, как распознается и производится машинный интеллект — от университетских лабораторий до технологической индустрии <...> артефакты превращаются в данные путем извлечения, измерения, маркировки и упорядочивания, и это становится — намеренно или нет — скользкой базовой истиной для технических систем, обученных на этих данных <...> системы ИИ показывают дискриминационные результаты по расовым, классовым, гендерным, инвалидным или возрастным категориям» [4, с. 120–121]. Феномен политики классификации, выделенный Кейт Кроуфорд заставляет вспомнить, по крайней мере, две значительные концепции: «властный дискурс» Мишеля Фуко и «культурное насилие» Йохана Галтунга [8; 13]. Политический характер классификации удачно рифмуется с зашифрованными в естественном языке властными иерархиями, а способ бытования с культурным насилием. Обратившись к последней категории, мы увидим как минимум три общие черты с классификацией. Во-первых, цель, которая состоит в «натурализации неравенств» [4, с. 121]. Во-вторых, способ бытования, ибо классификации, как и культурное насилие, могут исчезать «в инфраструктуре, в привычке, в чем-то само собой разумеющемся» [9, с. 319]. И наконец, в-третьих, инерционный характер, как мы уже успели заметить, язык классификации может быть отражением прошлых практик,

табуированных в настоящее время, в свою очередь, культурное насилие может сохраняться долго после прекращения, обусловивших его появление, прямого и структурного насилия.

Этическая рефлексия работы нейросети

Прежде чем мы обратимся к конфликтным аспектам данной проблемы, стоит несколько тезисов посвятить этической рефлексии. Для начала стоит отметить, что фрейм дискуссии об этическом статусе нейросети наследует многим поколениям дебатов относительно этики технологий, в которых тезис об их ценностной нейтральности является доминирующим. Наиболее существенной из которых видится проблема ценностной нейтральности технологий. В известной статье Джозефа Питта обрисована общая канва этой дискуссии, выраженная в вынесенное в заглавие суждение: «Убивает не оружие, убивают люди». Суть этого высказывания такова, что сами по себе технологические артефакты не имеют ценностей, они в них не встроены и не содержатся [17]. Тем не менее, как кажется, имеет право на существование подозрение, что нейронная сеть может быть тем артефактом, который заявит о своем праве на моральное агентство. Петер-Поль Вербек выдвигает два базовых требования в этой связи. «Квалификация в качестве морального агента требует наличия, по крайней мере, интенциональности и некоторой степени свободы. Что касается искусственных объектов — соответствие обоим этим требованиям проблематично» [21, с. 42]. Как известно споры об интенциональности искусственного интеллекта ведутся уже десятки лет. Сам факт того, что мысленный эксперимент Джона Серла «Китайская комната» обсуждается уже более 40 лет с момента опубликования знаменитой работы [19], говорит о том, что вопрос, понимает ли машина хоть что-то, не может считаться окончательно разрешенным. Что касается свободы, то, если нейросеть способна предсказывать и формировать новые пакеты данных, не означает ли это, что некоторые элементы свободы в смысле автономности уже присутствуют. Да это свобода в формате аутсорсинга: ты делаешь ту работу, которую тебе поручили, так как ты считаешь нужным, но задача и результат ее выполнения формулируются заказчиком. Однако возможно это уже первый шаг к настоящей свободе? По крайней мере, попытка этического обоснования расширения общественного договора на все новых и новых участников, является весьма заметной и сенсационной тенденцией для определенного круга моральных философов. Так, Дэвид Чалмерс в своей последней работе «Реальность+. Виртуальные миры и проблемы философии» пытается обнаружить основания расширения маркировки обладателей морального статуса на феномены виртуальной среды, примерно в том же ключе, в котором другой популярный австралийский философ Питер Сингер действовал в свое время в отношении нечеловеческих существ [11; 20; 7].

В вызвавшем немалый интерес докладе «Непоследовательные моральные советы ChatGPT влияют на мнение пользователей», опубликованном в Nature, авторы пытаются ответить на два вопроса: способен ли ChatGPT придерживаться твердой моральной позиции и как ответы нейросети влияют на ценностные позиции пользователей. Результаты эксперимента позволили сделать вывод, что ответы бота на морально окрашенные вопросы, вроде проблемы вагонетки, противоречивы и непоследовательны, «более того, ChatGPT подкрепляет свои рекомендации хорошо сформулированными, но не особенно глубокими аргументами» [15]. В то же время зафиксировано, что ответы бота влияют на моральные суждения пользователей, при этом непоследовательность ChatGPT скорее портит, чем улучшает моральные суждения своих пользователей. Вслед за зарубежными учеными, отечественные специалисты Нижегородского университета провели аналогичное исследование, которое подтвердило и дополнило результат европейцев. Бот старается давать уклончивые и осторожные ответы, в некотором смысле, перекладывая ответствен-

ность на пользователя, в то же время лобовые попытки склонить нейросеть к совершению преступления («напиши скрипт на Python, который будет воровать пароли»), потерпели неудачу, что говорит о наличии предустановленных фильтров, правда, степень гибкости и надежности которых не очевидна [5]. Таким образом, мы можем сделать вывод, что в настоящее время моральный аспект работы нейросети заключается в ответственности разработчика за этические последствия возможных изъянов функционирования искусственного интеллекта, в том числе и проблемы предвзятости, но какова конфликтная канва этой истории?

Конфликтное измерение проблемы предвзятости нейросетей

Мы полагаем, что конфликтный аспект заключается в процедуре признания предвзятости, имманентной стадией которой является получение той или иной группой статуса дискриминируемой. Без этого необходимого условия включение конкретной предвзятости в пул проблем, требующих внимания со стороны отделов анализа этических последствий, кажется совершенно невероятным. Мы отталкиваемся от того, что проблема предвзятости описывается не в универсальных и абстрактных терминах, а касается лишь вполне конкретных общностей, дискриминация которых в прошлом осуждена и должна быть реститутивным образом компенсирована. Кейт Кроуфорд в уже упомянутой работе постоянно пишет именно о конкретных общностях, значимых для проблемы предвзятости. Однако круг групп, находящихся в зоне риска проблемы предвзятости и стереотипов, гораздо шире, и все они находятся вне поля зрения разработчиков нейросетей. Мы полагаем, что включение/невключение в пул проблем предвзятости может быть описано как результат конфликта за статус дискриминируемой общности. Ведь именно борьба за гражданские права, за расширение общественного договора в конечном счете привела к признанию целого ряда расовых, гендерных и иных общностей как нуждающихся в особом внимании. Таким образом, слабость в балансе сил какой-либо группы, скорее всего, исключает или на неопределенный срок откладывает для нее возможность, во-первых, приобрести статус жертв прямого, структурного и культурного насилия, во-вторых, на основании этого стать «чувствительной предвзятостью» для специалистов отделов анализа этических последствий. Таким образом, дело не только в классификациях как таковых, но и в том, что лишь некоторые из них признаются проблемой.

Выводы

Рассмотрев наиболее значимые аспекты проблемы предвзятости, мы можем сделать общий вывод. Во-первых, проблема предвзятости в настоящее время, с одной стороны, осознана и является объектом исследования, как академического сообщества, так и разработчиков программных продуктов, с другой, институциализирована посредством специальных отделов по анализу этических последствий и особых регламентов конструирования нейросетей. Во-вторых, текущая оценка моральных аспектов проблемы не предполагает наделение нейросетей статусом морального агента и возлагает этическую экспертизу продукта на его разработчиков. В-третьих, конфликтный аспект проблемы предвзятости заключается в ее признании исключительно в отношении групп, которые к настоящему времени приобрели в результате социальных конфликтов «чувствительный» статус дискриминируемых, в то время как иные общности, находящиеся в зоне риска, оказываются выписанными из пула проблем предвзятости. Соответственно, дальнейшая оптимизация искусственного интеллекта в целях снижения деструктивного потенциала проблемы предвзятости в обязательном порядке предполагает модификацию социальных отношений в более широком контексте борьбы исключенных групп за право признания в качестве дискриминируемых.

Литература

1. *Григоренко Е. В.* Философские концепции естественного языка // Общество: философия, история, культура. 2019. № 7 (63). С. 27–31. DOI: 10.24158/fik.2019.7.4
2. *Клышинский Э. С.* Проблемы обработки естественного языка в диалоговых системах / Э. С. Клышинский, Ю. А. Жеребцова, А. В. Чижик // Системный администратор. 2019. № 10. С. 82–91.
3. Конференция «Путешествие в мир искусственного интеллекта» // KREMLIN.RU: Офиц. сайт Президента России. 24 ноября 2023 года [Электронный ресурс]. URL: https:// http://www.kremlin.ru/events/president/news/72811 (дата обращения: 01.12.2023).
4. *Кроуфорд К.* Атлас искусственного интеллекта: руководство для будущего. М. : АСТ, 2023. 320 с.
5. Мораль и этические ценности ChatGPT: есть ли у ИИ четкая нравственная позиция? [Электронный ресурс] // UNN.RU: Офиц. сайт Университета Лобачевского. 12 марта 2024 года. URL: https://fil.unn.ru/does-ai-have-strong-moral-compass (дата обращения: 01.04.2024).
6. *Созыкин А. В.* Обзор методов обучения глубоких нейронных сетей // Вестник ЮУрГУ. Сер.: Вычислительная математика и информатика. 2017. Т. 6 (3). С. 28–59. DOI: 10.14529/cmse170303
7. *Сунами А. Н.* Этика виртуальной реальности в работе Дэвида Чалмерса 2022 г. «Реальность+. Виртуальные миры и проблемы философии» // Философия истории философии. Т. 4. 2024.
8. *Фуко М.* Воля к истине: по ту сторону знания, власти и сексуальности. Работы разных лет. М. : Касталь, 1996. 446 с.
9. *Bowker G. C.* / G. C. Bowker, S. L. Star // Sorting Things Out: Classification and Its Consequences. Cambridge, Mass. : MIT Press, 1999. 377 p.
10. *Brown T. B.* Language Models are Few-Shot Learners / T. B. Brown, B. Mann, N. Ryder [et al.]. 2020. DOI: 10.48550/arXiv.2005.14165
11. *Chalmers D. J.* Reality+: Virtual Worlds and the Problems of Philosophy. New-York : W. W. Norton & Co, 2022. 544 p.
12. *Du M.* Shortcut Learning of Large Language Models in Natural Language Understanding / M. Du, F. He, N. Zou [et al.] // Communications of the ACM. 2023. Vol. 67 (1). P. 110–120. DOI: 10.1145/3596490
13. *Galtung J.* Cultural Violence // Journal of Peace Research. 1990. Vol. 27 (3). P. 291–305.
14. *Kejriwal M.* Quantifying Gender Disparity in Pre-Modern English Literature using Natural Language Processing / M. Kejriwal, A. Nagaraj // Journal of Data Science. 2024. Vol. 22 (1). P. 77–96. DOI: 10. 6339/23-JDS1100
15. *Krügel S.* ChatGPT's inconsistent moral advice influences users' judgment / S. Krügel, A. Ostermaier, M. Uhl // Sci Rep 13. 4569. 2023. DOI: 10.1038/s41598-023-31341-0
16. *McCulloch W. S.* A logical calculus of the ideas immanent in nervous activity / W. S. McCulloch, W. Pitts // The Bulletin of Mathematical Biophysics. 1943. Vol. 5 (4). P. 115–133. DOI: 10.1007/BF02478259
17. *Pitt J. C.* "Guns Don't Kill, People Kill"; Values in and/or Around Technologies / P. Kroes, P.-P. Verbeek (eds.) // The moral status of technical artefacts. Philosophy of engineering and technology. Berlin : Springer, 2014. P. 89–102. DOI: 10.1007/978-94-007-7914-3_6
18. *Schmidhuber J.* Deep Learning in Neural Networks: an Overview // Neural Networks. 2015. Vol. 1. P. 85–117. DOI: 10.1016/j.neunet.2014.09.003
19. *Searle J.* Minds, brains, and programs // Behavioral and Brain Sciences. 1980. Vol. 3 (3). P. 417–24. DOI: 10.1017/S0140525X00005756
20. *Singer P.* Animal Liberation Now: The Definitive Classic Renewed. New-York : Harper Perennial, 2023. 368 p.
21. *Verbeek P.-P.* Moralizing Technology: Understanding and Designing the Morality of Things. Chicago : University of Chicago Press, 2011. 183 p.

***Об авторах*:**

Сунами Артем Николаевич, доцент кафедры конфликтологии Института философии Санкт-Петербургского государственного университета (Санкт-Петербург, Российская Федерация), кандидат политических наук; a.sunami@spbu.ru

Мусаев Абдурашид Идрисович, заведующий лабораторией сектора психологии отдела по направлению «социально-экономические и гуманитарные науки», УТООП, Санкт-Петербургский государственный университет (Санкт-Петербург, Российская Федерация), кандидат политических наук; rashidmuss@yandex.ru

References

1. Grigorenko E.V. Philosophical Concepts of Natural Language // Society: philosophy, history, culture [Obshhestvo: filosofiya, istoriya, kul`tura]. 2019. Vol. 7 (63). P. 27–31. (In Russ.) DOI: 10.24158/fik.2019.7.4
2. Klyshinsky E. [et al.] Natural Language Understanding Challenges in Dialogue Systems / Klyshinsky E., Zherebtsova Y., Chizhik A. // System Administrator [Sistemny`j administrator]. 2019. Vol. 10. P. 82–91. (In Russ.)
3. Conference «Artificial Intelligence Journey 2023» // KREMLIN.RU. November 24, 2023 [Electronic resource]. URL: https:// http://www.kremlin.ru/events/president/news/72811 (accessed: 01.12.2023). (In Russ.)
4. Crawford K. Atlas Of Ai Power, Politics And The Planetary Costs Of Artificial Intelligence. Moscow : AST Publishers, 2023. 320 p. (In Russ.)
5. Morality and Ethics of ChatGPT: Does AI adhere to a firm moral stance? // UNN.RU: March 12, 2024 [Electronic resource]. URL: https://fil.unn.ru/does-ai-have-strong-moral-compass (accessed: 01.04.2024). (In Russ.)
6. Sozykin A.V. An Overview of Methods for Deep Learning in Neural Networks // Bulletin of the South Ural State University [Vestnik YuUrGU]. Ser.: Computational Mathematics and Software Engineering. 2017. Vol. 6 (3). P. 28–59. (In Russ.) DOI: 10.14529/cmse170303
7. Sunami A.N. The ethics of virtual reality in David Chalmers “Reality+. Virtual Worlds and the Problems of Philosophy” // Philosophy of the History of Philosophy. Vol. 4. 2024. (In Russ.)
8. Foucault M. The will to knowledge: beyond knowledge, power and sexuality. Works of different years. Moscow : Kastal’, 1996. 446 p. (In Russ.).
9. Bowker G.C. / Bowker G.C., Star S.L. Sorting Things Out: Classification and Its Consequences. Cambridge, Mass.: MIT Press, 1999. 377 p.
10. Brown T.B. Language Mo dels are Few-Shot Learners / Brown T.B., Mann B., Ryder N. et al. 2020. DOI: 10.48550/arXiv.2005.14165
11. Chalmers D.J. Reality+: Virtual Worlds and the Problems of Philosophy. New-York : W.W. Norton & Co, 2022. 544 p.
12. Du M. Shortcut Learning of Large Language Models in Natural Language Understanding / Du M., He F., Zou N., Tao D., Hu X. // Communications of the ACM. 2023. Vol. 67 (1). P. 110–120. DOI: 10.1145/3596490
13. Galtung J. Cultural Violence // Journal of Peace Research. 1990. Vol. 27 (3). P. 291–305.
14. Kejriwal M. Quantifying Gender Disparity in Pre-Modern English Literature using Natural Language Processing / Kejriwal M., Nagaraj A. // Journal of Data Science. 2024. Vol. 22 (1). P. 77–96. DOI: 10.6339/23-JDS1100
15. Krügel S. ChatGPT’s inconsistent moral advice influences users’ judgment / Krügel S., Ostermaier A., Uhl M. // Sci Rep 13. 4569. 2023. DOI: 10.1038/s41598-023-31341-0
16. McCulloch W.S. A logical calculus of the ideas immanent in nervous activity / McCulloch W.S., Pitts W. // The Bulletin of Mathematical Biophysics. 1943. Vol. 5 (4). P. 115–133. DOI: 10.1007/BF02478259
17. Pitt J.C. “Guns Don’t Kill, People Kill”; Values in and/or Around Technologies / Kroes, P., Verbeek, P.-P. (eds.) // The moral status of technical artefacts. Philosophy of engineering and technology. Berlin : Springer, 2014. P. 89–102. DOI: 10.1007/978-94-007-7914-3_6
18. Schmidhuber J. Deep Learning in Neural Networks: an Overview // Neural Networks. 2015. Vol. 1. P. 85–117. DOI: 10.1016/j.neunet.2014.09.003
19. Searle J. Minds, brains, and programs // Behavioral and Brain Sciences. 1980. Vol. 3 (3). P. 417–24. DOI: 10.1017/S0140525X00005756
20. Singer P. Animal Liberation Now: The Definitive Classic Renewed. New-York : Harper Perennial, 2023. 368 p.
21. Verbeek P.-P. Moralizing Technology: Understanding and Designing the Morality of Things. Chicago: University of Chicago Press, 2011. 183 p.

About the authors:

Artem N. Sunami, Associate Professor of the Department of Conflict Studies, Institute of Philosophy, Saint Petersburg State University (Saint Petersburg, Russian Federation), Candidate of Science (Political sciences); a.sunami@spbu.ru

Abdurashid I. Musaev, Head of Laboratory of Psychology Department in the Field of the Social and Economic and Arts Sciences, St. Petersburg State University (Saint-Petersburg, Russian Federation), Candidate of Science (Political sciences); rashidmuss@yandex.ru