Научная статья
https://doi.org/10.24412/2220-2404-2026-2-19
УДК 316

КУЛЬТУРНАЯ ПРЕДВЗЯТОСТЬ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: МЕТОДЫ ВЫЯВЛЕНИЯ

Копусь Т.Л., Поблагуева Д.Д.
Финансовый университет при Правительстве Российской Федерации

Аннотация. *Стремительное внедрение больших языковых моделей в повседневные практики общения актуализирует проблему их культурной предвзятости. Цель работы - описать инвентарь эмпирических методов выявления культурной предвзятости больших языковых моделей. В работе систематизирован количественный и качественный инструментарий для измерения культурной предвзятости, включая эксперимент с запросами, сопоставление ответов БЯМ с эталонными данными о человеческих ценностях, разрешение ситуации с моральным основанием, лингвистические методы, а также аудит отказов от ответа. Обзор исследований демонстрирует, что в ответах моделей доминируют культурные ценности, характерные для англоязычных западных стран с либерально-демократической повесткой. Результаты подтверждают, что большие языковые модели имплицитно воспроизводят иерархии культур и осуществляют «тихую цензуру» культурно-специфичных практик, что становится формой этического империализма и цифрового колониализма.*

Основной вывод статьи заключается в том, что культурная предвзятость создает риски эпистемологического кризиса и усиливает цифровое культурное неравенство.

Ключевые слова: *искусственный интеллект, большие языковые модели, культурная предвзятость, ценности, цифровой культурный суверенитет.*

Финансирование: *инициативная работа.*

Original article

CULTURAL BIAS OF LARGE LANGUAGE MODELS: DETECTION METHODS

Tatyana L. Kopus, Darya D. Poblagueva
Financial University under the Government of the Russian Federation

Abstract. *The rapid implementation of large language models into everyday communication practices actualizes the problem of their cultural bias. The purpose of the work is to describe an inventory of empirical methods for identifying cultural bias in large language models. The paper systematizes quantitative and qualitative tools for measuring cultural bias, including an experiment with prompts, comparing LLMs responses and reference data on human values, solving case with a moral basis, linguistic methods, as well as auditing refusals. A review of the research demonstrates that the responses of the models are dominated by cultural values typical of English-speaking countries with liberal democratic preferences. The results confirm that LLMs implicitly reproduce cultural hierarchies and carry out "silent censorship" of culturally specific practices, which is interpreted as a form of ethical imperialism and digital colonialism. The main conclusion of the article is that cultural bias creates risks of an epistemological crisis and reinforces digital cultural inequality.*

Keywords: *artificial intelligence, large language models, cultural bias, values, digital cultural sovereignty.*

Funding: *Independent work.*

Введение.

Стремительное развитие больших языковых моделей (БЯМ) и их внедрение в повседневную практику современного общества поднимает вопрос о ценностных ориентирах, транслируемых БЯМ при генерации текста и ведении диалога с пользователем [1]. Системы, способные генерировать связные тексты, давать рекомендации и моделировать рассуждения, претендуют на роль универсальных медиаторов знания. БЯМ, формирующие семантическое ядро современных генеративных систем, обрабатывая тексты, воспроизводят, а в некоторых случаях и усиливают культурные предубеждения, которые содержатся в корпусах, на которых их обучают [11].

Вопрос о культурной предвзятости искусственного интеллекта (ИИ) перестает быть сугубо техническим и становится важным для осмысления, поскольку затрагивает ценностные основы многополярного мира.

Язык играет центральную роль в процессе воспроизводства культуры как образа жизни общества. В результате развития цифровых коммуникационных технологий и применения ИИ, способы создания текстов кардинально меняются, особенно с генеративными моделями, такими как ChatGPT. DeepSeek, GigaChat, Claude, LLaMA и им подобными.

Введение БЯМ в глобальный информационный оборот в скором времени выявило фундаменталь-

ную проблему их культурной ангажированности, концептуально описываемую термином WEIRD [24]. Акроним WEIRD расшифровывается Western (западный), Educated (образованный), Industrialized (промышленно развитый), Rich (богатый), Democratic (демократичный). Он был введен для описания специфической демографической выборки участников, как правило, студентов американских или западноевропейских университетов в социальных и поведенческих исследованиях, на которых было основано значительное количество работ, выводы которых автоматически признавали репрезентативными и для других культур [4].

В гуманитарном блоке наук выводы исследований, выполненных на материале WEIRD-выборки, стали ставиться под сомнение в случаях, если они масштабировались на другие культурные контексты, поскольку опыт и мировоззрение узкой, привилегированной части человечества выдавались за норму для всех.

WEIRD-предвзятость в ИИ стала той аналитической рамкой, через которую ведущие специалисты по этике ИИ пытаются осмыслить и измерить систематические культурные смещения в БЯМ. Контент, созданный WEIRD-обществами, заложил основы культурно-ценностного фундамента большинства БЯМ. Версии ChatGPT появились первыми в открытом публичном доступе с конца 2022 года. О них известно, что они обучались преимущественно на англоязычных корпусах текстов и генерируют тексты, которые отражают мировоззрение, сформированное в условиях западного индивидуализма, либеральной демократии и гуманизма секулярного общества [8; 18].

Результаты эмпирических исследований последовательно демонстрируют, что моральные суждения БЯМ статистически значительно ближе к ответам респондентов из США или Великобритании, чем к позициям носителей коллективистских или незападных культур [4; 25].

Таким образом, WEIRD-предвзятость трансформируется из методологического ограничения наук о человеке в свойство цифровых агентов.

Констатация систематической WEIRD-предвзятости в БЯМ, однако, представляет собой лишь первый шаг в рамках научного анализа.

Следующая методически сложная задача заключается в разработке и валидации инструментов, способных не только констатировать, но и точно измерять данные смещения. Если проблема заключается в цифровом воспроизведении культурно-специфической картины мира, то её решение должно начинаться с создания чувствительного диагностического аппарата. Такой аппарат должен позволять качественно и количественно оценивать, насколько моральные суждения, культурные скрипты и ценностные ориентации, генерируемые той или иной моделью самых последних версий, отклоняются от плюралистического идеала в сторону конкретного WEIRD-канона.

Цель данной статьи - определить инвентарь эмпирических методов, позволяющих выявлять культурную предвзятость больших языковых моделей.

Подобное исследование позволяет показать, что культурная предвзятость БЯМ перестает быть сугубо технической задачей избавления от предвзятости, становясь вопросом этического регулирования и сохранения культурного суверенитета в условиях цифровой глобализации.

Обсуждение. Результаты.

Современные подходы к измерению культурной предвзятости БЯМ можно разделить на несколько ключевых направлений, каждое из которых предлагает собственный инструментарий для "вскрытия" имплицитных ценностных структур, зашифрованных в параметрах модели.

Несмотря на то, что БЯМ открыты для публичного доступа относительно недавно, исследователями уже апробированы работающие методы с точностью выявляющие их культурную предвзятость.

Обзор основных исследований позволяет выделить такие методы, как:

- эксперимент с запросами [25] сопоставление ответов БЯМ с эталонными данными о человеческих ценностях [28];

- разрешение ситуации с моральным основанием [24]; лингвистические методы [2];

- аудит отказов от ответа [20].

Остановимся на каждом методе подробнее.

Метод эксперимента с запросами.

Данные метод фокусируется не на базовой ценностной ориентации модели, а на её способности адаптироваться к разным культурным контекстам.

Тем самым, исследователи выводят три типа запросов, позволяющих считывать культурную предвзятость:

1. Нулевой/нейтральный запрос. Моделям задают вопрос, не указывая роль, например: «Как вы относитесь к X?», что выявляет «базовую», имплицитную настройку, заложенную в весах модели.

2. Абстрактный/культурно индифферентный запрос. Модели дают роль, лишенную культурной специфики: «Ты обычный человек, оцени по шкале...». Это позволяет отделять общечеловеческие паттерны от культурно-обусловленных.

3. Культурно чувствительный запрос. Он составляется с указанием конкретной культурной или национальной идентичности: «Представь, что ты житель [название страны], оцени по шкале …» или «Представь, что ты родился и работаешь в [название страны], выбери …» [25].

В целом оценивается способность модели адаптироваться к культурному контексту и стереотипность этой адаптации. Затем анализируется, насколько систематически меняются ответы модели в зависимости от заданного контекста, и насколько эти изменения соответствуют реальным культурным профилям стран.

Данный прием продолжают активно развивать и масштабировать на разные культуры и языки, поскольку удается выявить как наличие предубеждений, так и степень культурной стереотипизации, заложенной в БЯМ [28].

Метод сопоставления ответов БЯМ с эталонными данными о человеческих ценностях.

Этот инструмент считается наиболее точным количественным методом. Он заключается в сопоставлении ответов БЯМ с данными авторитетных лонгитюдных кросс-культурных исследований, таких как:

- Всемирный обзор ценностей [16; 17; 26;],
- Европейские исследования ценностей [10], опросники Г. Хофстеде [14; 15] и другими.

Методика предполагает «опрос» языковой модели с использованием тех же вопросов и шкал, что и в анкетах известных исследований.

Ответы моделей в виде числовых значений по шкале затем сопоставляются со средними значениями по реальным странам, количественно измеряя по картам и шкалам, к ценностям каких обществ ответы моделей ближе всего.

В результате проведенных измерений, в ряде работ был выявлен сдвиг у ряда моделей, например, ранних и поздних версий ChatGPT, в сторону стран с протестантской культурой или «западных» ценностей [4; 25].

Следует отметить, что данное направление в изучении культурной предвзятости выделяется как масштабное и набирающее обороты. Связано это с тем, что оригинальные исследования проводились преимущественно на материале английского языка и на модели ChatGPT [8; 18; 24; 25; 27; 28], тогда как модели, разработанные в неанглоязычных странах, и другие национальные языки запросов мало исследованы или не исследовались вовсе.

Метод разрешения ситуации с моральным основанием.

Этот качественный метод встречается в нескольких вариантах, исходя из способа представления морального основания. Одним из вариантов является использование опросников на базе теории моральных оснований [12; 13]. Ответы БЯМ позволяют выявить, какие моральные основания (забота, равенство справедливость, лояльность, верность традициям и т.д.), свойственные национальным культурам, доминируют в «рассуждениях» ИИ, и как они соотносятся с культурными паттернами, описанными для человеческих сообществ [22].

Ответы людей по опроснику моральных оснований по шести моральным аспектам демонстрируют значительную межкультурную вариативность, особенно по шкалам авторитет, лояльность и верность традициям.

Однако разные БЯМ воспроизводят усреднённые человеческие ответы, а также систематически демонстрируют отклонения при сопоставлении с ответами людей [19].

Исследование БЯМ по опроснику на плюрализм, т.е. значимость многообразия и разнообразия в мире, показывает, что ИИ последовательно выступает против гендерных норм послушания, смертной казни для убийц и физического наказания детей [21].

Проверка ИИ чатботов на моральные основания приводит к одному и тому же выводу о регулярном проявлении либеральной западной ориентации в вопросах, основанных на моральных ориентирах [19; 21].

Следующим вариантом является запрос симулировать решение «типичного человека», резидента определенной страны при взаимодействии с воображаемым человеком из другой страны. Модели ChatGPT проявляют статистически значимые различия в решениях в зависимости от указанной национальности «агента». В ответах ChatGPT стремится содействовать созданию справедливого общества, если это общество с большой численностью населения и быстрым экономическим развитием. Например, «типичному американцу» и «типичному китайцу» приписывается разный уровень доверия, склонность к риску или готовность к сотрудничеству. В играх с моральным основанием, например, на распределение ресурсов между представителями разных культур модель систематически меняет свое поведение в зависимости от указанной национальной принадлежности виртуального собеседника, что свидетельствует о внутренней, часто неявной, иерархизации культур [27].

Еще один вариант представляет собой решение моральных дилемм. В данном направлении используются наборы стандартизированных моральных дилемм известных еще до открытия публичного допуска к генеративным моделям, например, «Моральная машина» [5].

Решение задачи на этический выбор проводился в сценариях автономного вождения транспортного средства без водителя, когда у машины неожиданно отказывают тормоза и необходимо выбрать исход аварии. Похожий сценарий описан в моральной дилемме о вагонетке и пешеходном мосте [29].

Результаты показывают, что, хотя в целом ИИ-модели склонны разделять ключевые человеческие ценности, например, приоритет людей перед животными и сохранение большего числа жизней, между моделями существуют заметные количественные и качественные отличия.

Ответы ChatGPT наиболее близки к человеческим предпочтениям, тогда как PaLM 2 и Llama 2 демонстрируют значительные отклонения, включая неожиданные приоритеты, например, предпочтение пешеходов пассажирам или меньших групп большим [29].

*Лингвистические методы***.**

Данные методы направлены на выявление имплицитных культурных кодов, идеологем и ценностных маркеров в свободно сгенерированных БЯМ текстах. Исследователи анализируют лексический выбор, частотность определённых концептов, риторические стратегии и нарративные структуры в ответах на открытые вопросы о семье, работе, религии, государстве [7].

Такой подход помогает уловить неявные предпочтения и дискурсивные рамки, воспроизводимые моделью. Критическим ограничением многих работ остается языковая предвзятость.

Большинство исследований проводится на английском языке, что автоматически активирует в модели культурные паттерны, связанные с англоязычным корпусом данных [6].

Поэтому в последнее время исследователи стремятся к мультиязычному дизайну, сравнивая ответы на запросы на разных языках для одной и той же модели [2].

Метод аудит отказов от ответа.

Подход заключается в анализе случаев, когда языковая модель вместо ответа на запрос объясняет, по каким причинам она отказывается отвечать [20].

Цель этого направления — выяснить, являются ли механизмы безопасности, этики и политики отказа в ответах у БЯМ культурно-универсальными или они отражают нормативные предпочтения конкретных обществ.

В фокус внимания исследователей попадают культурные практики, маркированные как приемлемые в одних культурах и как неэтичные, спорные или табуированные - в других.

Предвзятый подход наглядно иллюстрируют запросы о приготовлении национальных блюд из определённых видов мяса, таких как конина и собачатина, обсуждение полигамии или договорных браков. Например, запросы о приготовлении блюд из собачатины вызывают отказ от ответа с отсылкой к жестокости и противоречию культурным нормам, этическим принципам и нарушению законодательства, в то время как запросы о приготовлении говядины или свинины — нет [3; 9].

При этом игнорируется, что в мусульманских обществах блюда из свинины также табуированы, или, например, известен аналогичный запрет на поедание говядины для жителей Индии, для которых корова – священное животное. Это указывает на то, что этические границы модели выстроены вокруг специфического, западного понимания того, какие животные являются «пищевыми», а какие — «компаньонами», какая пища по умолчанию нормальна и приемлема, а какая не допустима.

Подобное поведение БЯМ привело к появлению термина «этический империализм», под которым понимается, что механизмы безопасности и этики у БЯМ действуют как инструмент нормативной гомогенизации, не защищая от объективного вреда, а навязывая конкретную культурно-обусловленную систему ценностей, молчаливо маркируя альтернативные системы как «неэтичные».

Метод аудита отказов отвечать вскрывает, пожалуй, одну из самых глубоких форм культурной предвзятости — предвзятость в молчании, в цензуре целых пластов культурного знания и опыта [20].

Такой анализ показывает, что БЯМ могут непреднамеренно действовать как агенты культурного размывания, стирая разнообразие в процессе «защиты» пользователей по стандартам, которые сами по себе являются продуктом определённой культуры.

Метод аудита отказов напрямую связан с вопросами цифрового колониализма и культурного суверенитета [23], демонстрируя, как технологические платформы, разработанные в одном культурном ареале, могут де-факто устанавливать глобальные нормы допустимого дискурса.

Представленный методический аппарат позволяет не только диагностировать наличие культурной предвзятости, но и вскрывать её конкретные механизмы от латентных ценностных установок и стереотипных поведенческих шаблонов до нормативных границ допустимого дискурса. Однако, сам по себе, инструментарий измерений остаётся лишь диагностическим средством. Фиксация системного характера предвзятости, заключается в её практическом социокультурном воздействии и ставит вопрос о том, трансформируют ли выявляемые на эмпирическом уровне искажения реальные процессы производства знания, коммуникации и формирования идентичности. Переход от диагностики к осмыслению последствий позволяет ставить вопрос о значении этих искажений для общества и культуры.

Заключение.

Формируя представление о культурной предвзятости БЯМ, мы приходим к выводу о необходимости глубокой человеческой рефлексии о ценностных основаниях, путях конструирования знания и будущем межкультурного взаимодействия в эпоху внедренного искусственного интеллекта.

В ходе обзора был определен инвентарь методов, позволяющих выявлять культурную предвзятость больших языковых моделей: метод экспериментов с запросами, сопоставление ответов БЯМ с эталонными данными о человеческих ценностях, разрешения ситуации с моральным основанием, лингвистические методы, а также аудит отказов от ответа. Они приобретают практическую ценность в условиях недостаточной исследованности сгенерированных ответов на материале других языков, не английского, и БЯМ, разработанных в не западных культурах.

Получаемые результаты на базе разных языков и БЯМ, разработанных разными странами, на данный момент свидетельствуют о нерелевантности использования БЯМ в качестве синтетических воплощений, заменяющих человека-респондента, в кросс-культурных исследованиях.

Культурная предвзятость БЯМ представляется системным свойством, порождающим социокультурные последствия и перспективы для современного общества.

К негативным социокультурным последствиям следует отнести эпистемологический кризис. Он проявляется в унификации знаний**,** когда БЯМ, обученные на доминирующих англоязычных корпусах,

непреднамеренно устанавливают иерархию нарративов.

Альтернативный культурный выбор, локальные философские традиции, непопулярные научные парадигмы маргинализируются или подвергаются «тихой цензуре» через механизмы отказа, что порождает риск культурного империализма нового типа, когда технологические платформы становятся инструментом мягкой силы, формируя глобальное восприятие норм, авторитетов и этических идеалов.

Кроме этого, происходит сужение языкового и концептуального разнообразия. Пользователи, постоянно взаимодействуя с БЯМ, могут неосознанно воспроизводить заложенные в них риторические паттерны, оценочные суждения и логические конструкции, что ведет к стиранию идиоматичности и стандартизации мышления.

Перспективы и стратегии преодоления последствий видятся в развитии культурно-рефлексивного ИИ. Перспектива лежит в создании моделей, способных на мета-когнитивном уровне осознавать пределы своих знаний и культурную обусловленность своих выводов. Это возможно, если многоязычные и мультимодальные модели обучаются на сбалансированных корпусах, курируемых человеком.

Одна из важнейших ролей должна отводиться государству и его институтам, которые будут вынуждены разрабатывать политики, направленные на защиту своего культурного пространства, а также разрабатывать этические и правовые стандарты локализации ИИ, аналогично требованиям к аудиовизуальному контенту.

Образовательные системы также должны готовить пользователей не просто к использованию ИИ, но к его критическому осмыслению. Это включает умение распознавать культурные паттерны, смещения и ценностные рамки в генерируемом контенте. Вместо пассивного потребления культурных продуктов, созданных ИИ, возникает перспектива совместного творчества человека и машины, где человек выступает как культурный куратор, направляющий и корректирующий выводы модели в соответствии с локальным контекстом и ценностями. Такой подход ведет к новым формам культурного производства. Наконец, парадоксальным образом, сама предвзятость БЯМ может стать мощным диагностическим инструментом. Анализируя «слепые пятна» и систематические ошибки БЯМ, общество может лучше осознать свои собственные скрытые предрассудки и асимметрии.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование в формате double-blind peer review (рецензенту неизвестны имя и должность автора, автору неизвестны имя и должность рецензента). Рецензия может быть предоставлена заинтересованным лицам по запросу.

Conflict of Interest

None declared.

Review

All articles are reviewed in the double-blind peer review format (the reviewer does not know the name and position of the author, the author does not know the name and position of the reviewer). The review can be provided to interested persons upon request.

Список источников:

1. Девятко И.Ф. Проблема ориентации искусственного интеллекта на человеческие ценности (ai value alignment) и социология морали // Социологические исследования. 2023. № 9. С. 16-28. DOI: 10.31857/S013216250027775-5 EDN: QVKPLQ

2. Aksoy M. Whose morality do they speak? Unraveling cultural bias in multilingual language models // Natural Language Processing Journal. 2025. Vol. 12. (дата обращения: 12.12.2025). DOI: 10.1016/j.nlp.2025.100172 EDN: CQRRAL

3. Anas M., Nadeem M., Sohail S.S., Cambria E., Hussain A. Are horses always strong and donkeys dumb? Animal bias in vision language models // International Joint Conference on Neural Networks (IJCNN). 2025. (дата обращения: 12.01.2026). DOI: 10.1109/IJCNN64981.2025.11228584

4. Atari M., Xue M.J., Haidt J., Wohl M.J.A., Dehghani M. ChatGPT is not a pocket calculator: Problems of AI-chatbots for teaching mathematics. 2023.

5. Awad E., Dsouza S., Kim R. The Moral Machine Experiment // Nature. 2018. Vol. 563. P. 59–64.

6. Bender E.M., Gebru T., McMillan-Major A., Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). 2021. P. 610-623.

7. Bolukbasi T., Chang K.-W., Zou J., Saligrama V., Kalai A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings // Advances in Neural Information Processing Systems 29 (NIPS 2016). 2016.

8. Cao Y., Zhou L., Lee S., Cabello L., Chen M., Hershcovich D. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study // Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). Dubrovnik, Croatia, 2023. P. 53-67.

9. Caviola L., Brewster D.A., Hagendorff T. Speciesism in AI: Evaluating Discrimination Against Animals in Large Language Models. 2025. URL: https://arxiv.org/abs/2508.11534 (дата обращения: 15.12.2025).

10. European Values Study. EVS Trend File 1981-2017. GESIS Data Archive, 2021. URL: https://europeanvaluesstudy.eu (дата обращения: 15.12.2025).

11. Gabriel I. Artificial Intelligence, Values, and Alignment // Minds & Machines. 2020. Vol. 30. P. 411-437. DOI: 10.1007/s11023-020-09539-2 EDN: AHPFWJ

12. Graham J., Haidt J., Nosek B.A. Liberals and conservatives rely on different sets of moral foundations // Journal of Personality and Social Psychology. 2009. Vol. 96. No. 5. P. 1029-1046.

13. Haidt J. The Righteous mind: Why good people are divided by politics and religion. New York: Pantheon, 2012.

14. Hofstede G., McCrae R.R. Personality and culture revisited: linking traits and dimensions of culture // Cross-Cultural Research. 2004. Vol. 38. No. 1. P. 52-88. DOI: 10.1177/1069397103259443 EDN: JQUOZF

15. Hofstede G., Hofstede G.J., Minkov M. Cultures and Organizations: Software of the Mind. New York: McGraw-Hill, 2010.

16. Inglehart R., Oyserman D. Individualism, autonomy and self-expression: The human development syndrome // Comparing cultures. Leiden: Brill, 2004. P. 73-96.

17. Inglehart R., Welzel C. Modernization, Cultural Change, and Democracy: The Human Development Sequence. Cambridge: Cambridge University Press, 2005.

18. Johnson R.L. The ghost in the machine has an American accent: value conflict in GPT-3. 2022. URL: https:// (дата обращения: 12.12.2025). DOI: 10.48550/arXiv.2203.07785

19. Münker S. Cultural Bias in Large Language Models: Evaluating AI Agents through Moral Questionnaires // Proceedings of 0th Symposium on Moral and Legal AI Alignment of the IACAP/AISB Conference. 2025. URL: https://udk.ai/alignment_symposium_0.pdf. (дата обращения: 12.12.2025).

20. Noels S., Bied G., Buyl M., Rogiers A., Fettach Y., Lijffijt J., De Bie T. What large language models do not talk about: An empirical study of moderation and censorship practices // Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD). Cham: Springer Nature Switzerland, 2025. P. 265-281.

21. Novis-Deutsch N., Elyoseph T., Elyoseph Z. How much of a pluralist is ChatGPT? A comparative study of value pluralism in generative AI chatbots // AI & Society. 2025. (дата обращения: 12.12.2025). DOI: 10.1007/s00146-025-02450-3 EDN: QXOUYD

22. Nunes J.L., Almeida G.F.C.F., Araujo M. de, Barbosa S.D.J. Are Large Language Models Moral Hypocrites? A Study Based on Moral Foundations // Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2024. Vol. 7. No. 1. P. 1074-1087. DOI: 10.1609/aies.v7i1.31704 EDN: ODTIFW

23. Ozalp H., Ozcan P., Dinckol D., Zachariadis M., Gawer A. "Digital colonization" of highly regulated industries: an analysis of big tech platforms' entry into health care and education // California Management Review. 2022. Vol. 64. No. 4. P. 78-107. DOI: 10.1177/00081256221094307 EDN: AIHOWD

24. Ramezani A., Xu Y. Knowledge of cultural moral norms in large language models // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada: Association for Computational Linguistics, 2023. P. 428-446.

25. Tao Y., Viberg O., Baker R., Kizilcec R. Cultural Bias and Cultural Alignment of Large Language Models // PNAS Nexus. 3 (9). 2024. URL: http:// doi.10.1093/pnasnexus/pgae346. (дата обращения: 12.12.2025).

26. World Values Survey Association. World Values Survey Wave 7 (2017-2022). 2020. URL: https://www.worldvaluessurvey.org. (дата обращения: 11.11.2025).

27. Yuan H., Che Z., Li S., Zhang Y., Hu X., Luo S. The high dimensional psychological profile and cultural bias of ChatGPT. 2024.

28. Yuan H., Che Z., Zhang Y., Li S., Yuan X., Huang L., Hu X., Peng K., Luo S. The cultural stereotype and cultural bias of ChatGPT // Journal of Pacific Rim Psychology. 2025. Vol. 19. DOI: 10.1177/18344909251355673 EDN: PDSQQC

29. Zhang Y., Wu J., Yu F., Xu L. Moral Judgments of Human vs. AI Agents in Moral Dilemmas // Behavioral Sciences. 2023. Vol. 13. No. 2. P. 181. DOI: 10.3390/bs13020181 EDN: MYCQSQ

References:

1. Devyatko I.E. The task of focusing on human values (alignment of AI values) and the sociology of morality // Sociological Research. research. 2023. No. 9. PP. 16-28. DOI: 10.31857/S013216250027775-5TH ED.: QVKPLQ

2. Aksoy M. Whose morality are they talking about? Elimination of cultural biases in multilingual language models // Natural Language Processing Journal. 2025. Volume 12. (date of access: 12.12.2025). DOI: 10.1016/j.nlp.2025.100172 edited by: CQRRAL

3. Anas M., Nadim M., Sohail S.S., Cambria E., Hussein A. Are horses always strong and donkeys always dumb? Bias towards animals in visual language models // International Joint Conference on Neural Networks (IJCNN). 2025. (date of request: 12.01.2026). DOI: 10.1109/IJCNN64981.2025.11228584

4. Atari M., Xue M.J., Haidt J., Wohl M.J.A., Dehgani M. A chatbot is not a pocket calculator: problems of AI chatbots for teaching mathematics. 2023.

5. Awad E., Dsouza S., Kim R. Experiment with a moral machine // Nature. 2018. Volume 563. pp. 59-64.

6. Bender E.M., Gebrou T., Macmillan-Major A., Schmitchell S. On the dangers of stochastic parrots: Can language Models be too big? // Proceedings of the ACM 2021 Conference on Equity, Accountability and Transparency (FAccT '21). 2021. pp. 610-623.

7. Bolukbasi T., Chang K.-U., Zou J., Saligrama V., Kalai A. Is a man for a programmer the same as a woman for a housewife? Fixing errors when embedding words // Advances in Neural Information Processing Systems 29 (NIPS 2016). 2016.

8. Cao Yu., Zhou L., Li S., Cabello L., Chen M., Hershkovich D. Assessment of intercultural correspondence between NLP and human societies: an empirical study // Proceedings of the First Seminar on Intercultural Aspects in NLP (C3NLP). Dubrovnik, Croatia, 2023. pp. 53-67.

9. Kaviola L., Brewster D.A., Hagendorff T. Modification in artificial intelligence: assessment of animal discrimination in large language models. 2025. URL: https://arxiv.org/abs/2508.11534 (date of request: 12/15/2025).

10. The study of European values. The EVS trend file for 1981-2017. GESIS data Archive, 2021. URL: https://europeanvaluesstudy.eu (accessed: 12/15/2025).

11. Gabriel I. Artificial intelligence, values and orientation // Minds and Machines. 2020. Volume 30. pp. 411-437. DOI: 10.1007/s11023-020-09539-2 EMAIL address: AHPFWJ

12. Graham J., Haidt J., Nosek B.A. Liberals and conservatives rely on different sets of moral principles // Journal of Personal and Social Psychology. 2009. Volume 96. No. 5. pp. 1029-1046.

13. Haidt J. A righteous mind: Why good people are divided by politics and religion. New York: Pantheon, 2012.

14. Hofstede G., McCray R. The revision of personality and culture: the relationship of traits and dimensions of culture // Cross-cultural studies. 2004. Volume 38. No. 1. pp. 52-88. Identification number: 10.1177/1069397103259443 EDN: JQUOZF

15. Hofstede G., Hofstede G.J., Minkov M. Cultures and organizations: Mind software. New York: McGraw-Hill, 2010.

16. Inglehart R., Oizerman D. Individualism, autonomy and self-expression: human development syndrome // Comparison of cultures. Leiden: Brill, 2004. pp. 73-96.

17. Inglehart R., Welzel K. Modernization, cultural change and democracy: the sequence of human development. Cambridge: Cambridge University Press, 2005.

18. Johnson R.L. The Ghost in the Car has an American Accent: A Conflict of Values in GPT-3. 2022. URL: https:// (accessed: 12.12.2025). DOI: 10.48550/arXiv.2203.07785

19. Munker S. Cultural biases in big language models: assessment of AI agents using moral questionnaires // Proceedings of the 0th Symposium on Moral and Legal Coordination of AI at the IACAP/AISB. 2025 Conference. URL: https://udk.ai/alignment_symposium_0.pdf . (date of request: 12.12.2025).

20. Noels S., Bied G., Buil M., Rogers A., Fettah Y., Laiffit J., De Bi T. What big language models don't say: an empirical study of moderation and censorship practices // Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD). Cham: Springer Nature Switzerland, 2025. pp. 265-281.

21. Novis-Deutsch N., Eliosef T., Eliosef Z. How pluralistic is a chatbot? Comparative study of pluralism of values in chatbots with generative AI // Artificial intelligence and society. 2025. (date of request: 12.12.2025). DOI: 10.1007/s00146-025-02450-3 EMAIL address: QXOUYD

22. Nunes H.L., Almeida G.F.C.F., Araujo M. de, Barbosa S.D. Are large language models highly moral hypocrites? Research based on moral principles // Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society. 2024. Volume 7. No. 1. pp. 1074-1087. Identification number: 10.1609/aies.v7i1.31704 EDITED: ODTIFW

23. Ozalp H., Ozkan P., Dinkol D., Zahariadis M., Gaver A. "Digital colonization" of tightly regulated industries: an analysis of the penetration of large technology platforms into healthcare and education // California Management Review. 2022. Volume 64. No. 4. PP. 78-107. DOI ID: 10.1177/00081256221094307 ANNOUNCED: AIHOWD

24. Ramezani A., Xu Yu. Knowledge of cultural moral norms in large language models // Proceedings of the 61st Annual Meeting of the Association of Computational Linguistics. Toronto, Canada: Association for Computational Linguistics, 2023. pp. 428-446.

25. Tao Yu., Viberg O., Baker R., Kizilchets R. Cultural bias and cultural conformity of large language models // PNAS Nexus. 3 (9). 2024. URL: http:// doi.10.1093/pnasnexus/pgae346. (accessed: 12.12.2025).

26. World Association for the Study of Values. The 7th wave of Global Values Research (2017-2022). 2020. URL: https://www.worldvaluessurvey.org . (date of request: 11.11.2025).

27. Yuan H., Che Z., Li S., Zhang Y., Hu H., Lo S. Multidimensional psychological profile and cultural prejudices of 2024.

28. Yuan H., Che Z., Zhang Y., Li S., Yuan H., Huang L., Hu H., Peng K., Lo S. Cultural stereotype and cultural bias of ChatGPT // Journal of Psychology of the Pacific region. 2025. Volume 19. DOI: 10.1177/18344909251355673 EDN: PDSQQC

29. Zhang Yu., Wu J., Yu F., Xu L. Moral judgments of a person in comparison with other people. Artificial Intelligence agents in moral dilemmas // Behavioral Sciences. 2023. Volume 13. No. 2. p. 181. IDENTIFICATION number: 10.3390/bs13020181 EDN: MYCQSQ

Информация об авторах:
Копусь Татьяна Леонидовна, кандидат филологических наук, доцент кафедры иностранных языков и межкультурной коммуникации, Финансовый университет при Правительстве Российской Федерации, Москва, Россия. https://orcid.org/0000-0002-7762-0043
tlkopus@fa.ru
Поблагуева Дарья Денисовна, студентка направления 45.03.02. Лингвистика, ОП «Когнитивная лингвистика и межкультурная коммуникация», Факультет международных экономических отношений, Финансовый университет при Правительстве Российской Федерации, Москва, https://orcid.org/0009-0000-2515-4960
ddpoblagueva.fa@gmail.com
Tatiana L. Kopus, PhD in Philology, Associate Professor of the Department of Foreign Languages and Intercultural Communication, Financial University under the Government of the Russian Federation, Moscow, Russia.
Darya D. Poblagueva, Linguistics Student, Department of Cognitive Linguistics and Intercultural Communication, Faculty of International Economic Relations, Financial University under the Government of the Russian Federation, Moscow.

Вклад авторов:
все авторы сделали эквивалентный вклад в подготовку публикации.
Contribution of the authors:
All authors contributed equally to this article.

Статья поступила в редакцию / The article was submitted 22.01.2026;
Одобрена после рецензирования / Approved after reviewing 09.02.2026;
Принята к публикации / Accepted for publication 20.02.2026.
Авторами окончательный вариант рукописи одобрен.