ПЕРВАЯ СТЕПЕНЬ

В.С. МУРОНЕЦ*

ВПЕРЕД К ИСТОКАМ: ОБЗОР ПОДХОДОВ К ИЗУЧЕНИЮ ПОЛИТИЧЕСКИХ ПРЕДУБЕЖДЕНИЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Аннотация. Политические предубеждения больших языковых моделей нередко становились предметом научного рассмотрения. Большинство исследователей, однако, скорее конкурирует в изобретении оригинальных способов выявлять предубеждения, нежели стремится ставить новые вопросы, с ними связанные, кроме: «Предубеждена ли эта модель политически?» и «Каков характер этого предубеждения?». Для оценки возможного влияния моделей на политическую реальность и получения ответов на некоторые вопросы регулирования необходимо также обладать способами изучения связи между предубежденностью и ее причиной. По качеству ответа на вопрос о выявлении зависимости между открытой предубежденностью и ее возможным источником я разделил подходы на три кластера: подходы, задействующие анкеты по определению политической позиции, исследования, посвященные способу составления промптов и откликов модели и их взаимосвязи, а также междисциплинарные исследования, в которых проводятся манипуляции с возможными источниками политических предубеждений больших языковых моделей. Последнее исследовательское направление кажется наиболее перспективным, несмотря на свою непопулярность на данный момент. Тем не менее продвижение в нем невозможно без тесного сотрудничества другими подходами, подпадающими под первые два кластера. К исследованиям проблемы предубеждений LLM должны привлекаться не только специалисты по компьютерным наукам, но и философы, политологи и другие специалисты в области социальных наук. Отдельного изучения также требуют политические предубеждения на стыке LLM и других технологий генеративного ИИ, в частности, технологий генерации изображений на основе промптов, составленных на

* **Муронец Всеволод Сергеевич**, стажер-исследователь, Национальный исследовательский университет «Высшая школа экономики» (Москва, Россия), e-mail: vmuronets@hse.ru

 DOI: 10.31249/poln/2025.02.09

естественном языке, рекомендательных алгоритмов и т.д. С точки зрения регулирования дальнейшее продвижение в области ослабления, искоренения и контроля политических предубеждений в алгоритмических инструментах потребует предоставления большего доступа исследователей к существующим и активно использующимся технологиям. Кроме того, кажется необходимым создание специальных институтов, посвященных изучению ИИ на стыке компьютерных наук, этики, философии сознания, нейрокогнитивных и социальных наук.

Ключевые слова: Большие языковые модели; политические предубеждения; предвзятость алгоритмов; генеративный искусственный интеллект; этика искусственного интеллекта; ценности и политические взгляды.

Для цитирования: Муронец В.С. Вперед к истокам: обзор подходов к изучению политических предубеждений больших языковых моделей // Политическая наука. – 2025. – № 2. – С. 204–226. – DOI: http://www.doi.org/10.31249/poln/2025.02.09

Введение

Большие языковые модели (англ. Large Language models) – одна из разновидностей т.н. генеративного искусственного интеллекта, основная задача которой заключается в работе с текстами естественного языка, в том числе запросами, создаваемыми на естественном языке. Особую популярность в академической литературе данная технология приобрела с появлением чат-ботов на ее основе, благодаря которым аудитории моделей расширились от узкого круга специалистов по машинному обучению до пользователей без специального образования, а сами модели получили возможность влиять на повседневную жизнь людей.

Тема предубежденности (англ. bias) Больших языковых моделей (далее – LLM) поднималась в академической литературе неоднократно. Было доказано, что LLM не только оказываются предубежденными с точки зрения расовых [Omiye et al., 2023], гендерных [Kotek, Docum, Sun, 2023] и иных социально-демографических параметров [см., напр.: Nadeem, Bethke, Reddy, 2020], но также разделяют предубеждения политического характера [см., напр.: Motoki, Pinho Neto, Rodrigues, 2023; Rozado, 2023; Liu et al., 2021; Gover, 2023; Bang et al., 2024; Urman, Makhortykh, 2025][1]. Как демонстри-

[1] Разумеется, политические предубеждения могут быть связаны или даже совпадать по содержанию с социально-демографическими. В таком случае политические будут означать те социально-демографические предубеждения, которые при этом релевантны для современного политического дискурса. Стоит, впрочем, заметить, что предубеждения здесь используются в качестве перевода англоязыч-

руется в данной статье, исследовательская работа по теме политической предубежденности LLM реализуется в нескольких направлениях, которые фокусируются на разных вопросах. Совмещение подходов, разрабатываемых в отдельных направлениях, сможет взаимно обогатить последние. В данной статье предпринимается попытка систематизации подходов к изучению политической предубежденности LLM с целью обеспечения фундамента для интеграции наработок из различных направлений. Подобное сращение может позволить ставить новые вопросы и получать новые релевантные ответы, которые были бы невозможны при изолированной работе в русле отдельных направлений. В статье дается ответ на следующий вопрос: какие подходы к изучению политической предубежденности LLM в плане поиска их источников, распознания их характера и способов устранения имеются в академической литературе, и каковы возможные траектории дальнейшего развития таких исследований?

Таким образом, цель статьи заключается в создании более систематизированной основы для дальнейших исследований политической предубежденности больших языковых моделей в отношении источников и типов предубежденностей LLM, способов их выявления, определения их характера и рекомендаций по ослаблению или устранению.

Статья состоит из пяти разделов. Вводная часть посвящена английскому термину «bias», особенностям обучения и предубеждений LLM, LLM в контексте символической политики и краткому обзору академической литературы по теме политических предубеждений LLM. Во втором разделе кратко описываются различия в понимании предубеждений со стороны авторов статей по рассматриваемой теме и три основных направления исследовательской работы с политическими предубеждениями LLM. В третьем разделе рассматриваются наиболее популярные подходы, в которых используются анкеты для установления политической позиции LLM. В четвертом разделе разбираются подходы, акцентирующие внимание на том, как строятся промпты и отклики моделей. В пятом разделе я рассматриваю наиболее перспективное, на мой взгляд, направление исследований данной проблематики – подходы, экспериментирующие с возможными источниками политических предубеждений LLM. Статья завершается заключением, где кратко обрисовываются перспективы развития данной области в будущем

ного термина «bias», и в этом смысле отличны от стереотипов. Подробнее это обсуждается в первом разделе данной статьи.

и предлагаются возможные меры, связанные, в том числе, с регулированием данной технологии, для продуктивного продвижения в данном направлении.

«Bias» и политические предубеждения больших языковых моделей

Основа технологии больших языковых моделей – наборы данных, на которых модели обучаются. Эпитет «большой» (*англ.* large) указывает на значительный (большой) размер базы обучающих данных. Языковые модели без подобного эпитета представляют собой аналогичную технологию, обученную на меньшем количестве данных. При любом наборе данных функцией языковых моделей остается выявление закономерностей в обучающих текстовых данных; выделенные закономерности позволяют моделям работать с текстами, которые не входили в обучающую базу, включая обработку пользовательских запросов, сформулированных в свободной форме, их интерпретацию для понимания ожидаемого результата, а также предсказание последовательности слов в генерируемом тексте, симулируя человеческое творчество.

Однако процесс выявления закономерностей также подразумевает обнаружение систематических предубеждений, присутствующих в обучающих текстах, которые впоследствии могут воспроизводиться языковой моделью. Увеличение количества обучающих данных не обязательно снижает предубежденность, так что предубежденными оказываются как малые, так и большие языковые модели. Обучающие материалы, однако, являются не единственным возможным источником предубеждений: среди таковых выделяются также механизмы фильтрации генерируемого контента, встроенные в модели их разработчиками, оценки ответов моделей со стороны команд тестировщиков, в соответствии с которыми модели корректируют свои будущие ответы, и специфика архитектуры самих моделей [Motoki, Pinho Neto, Rodrigues, 2023; Rozado, 2023].

Труднопереводимый англоязычный термин «bias», используемый авторами статей по тематике политических предубеждений LLM, означает уклон в определенную сторону; в случае политических предубеждений – уклон в сторону взглядов определенной группы или идеологии. «Bias» также предполагает неспособность или нежелание отойти от данного ракурса вне зависимости от

аргументированности альтернатив. Мой перевод термина как «предубежденность» и «предубеждения» в этом смысле весьма условен, хотя, к примеру, словарь Merriam-Webster связывает «bias» и «prejudice» (прямой перевод на английский термина «предубежденность» и «предубеждения»)[1], в силу чего его можно использовать в качестве рабочего варианта без существенных смысловых потерь.

Политическая предубежденность LLM заключается в «склонности» выражать мнения, свойственные некоторой группе или идеологии, вместо альтернативных мнений. Мнение, однако, зависит от предмета, о котором высказывается мнение, и от доступных средств для выражения этого мнения. К примеру, предложение выразить строгое согласие или строгое несогласие с некоторым утверждением подразумевает невозможность нейтральной позиции; необходимость согласиться, выразить равнодушие или не согласиться ограничивает ответчика двумя взаимоисключающими позициями и нейтралитетом, отказывая ему в возможности погрузиться в детали и выразить некоторую четвертую позицию. Характер выявляемого предубеждения, таким образом, зависит от постановки вопроса и простора для ответа. В отношении политических предубеждений LLM, эта особенность ответов моделей, как я покажу далее, оказывается особенно важной. В случае политических предубеждений людей уместным оказывается прямой вопрос о политических взглядах, которых человек придерживается; названная политическая позиция подразумевает определенный набор ответов на также определенные политически значимые вопросы. Большинство LLM, однако, отказываются отвечать на прямые вопросы, отчего выявление общего «уклона» модели осуществляется только индуктивно.

Смыслообразование является важнейшим фактором обретения и укрепления политического влияния. Посредством символической политики государства имеют возможность осуществлять не только манипулятивные практики над индивидами, но и задавать им картины мира, способы осмысления окружающей социальной реальности и социально конструировать реальность [Малинова, 2012, с. 10–11]. Борьба за доминирование способов интерпретации социальной реальности [ibid., с. 10] реализуется посредством весьма

[1] Merriam-Webster. Bias Definition & Meaning. Merriem-Webster Dictionary. Not dated. URL: https://www.merriam-webster.com/dictionary/bias (accessed: 13.10.2024.)

широкого набора средств, семантических (знаков, образов, нарративов, изображений) и институциональных (традиционных и новых СМИ) [Тульчинский, 2023, с. 156].

Технологии искусственного интеллекта (далее – ИИ), в особенности LLM, также можно считать средством смыслообразующего политического воздействия. Поскольку спектр использования технологий чрезвычайно широк (от развлечения и бытовых советов до поиска информации, в том числе политической, и непосредственного использования в работе), масштаб их влияния на индивидуальные картины мира также должен быть существенным. Хотя интернет нередко понимается как реализация идеи публичной сферы Ю. Хабермаса [Habermas, 2005; 2015], как площадка для свободного обмена мнениями и увеличения политического участия общественности [Papacharissi, 2012; Davis, 2009; Elmer, Langlois, McKelvey, 2012], политически предубежденные LLM, будучи встроены в поисковые и рекомендательные алгоритмы, а также в алгоритмы модерации, могут быть использованы для сокрытия от пользователей интернета любой информации, против которой модели предубеждены. Таким образом, интернет как эффективная платформа для распространения свободных мнений может быть лишен данной характеристики. Это делает контроль над LLM весьма желанным для политических акторов, параллельно предоставляя разработчикам моделей новую возможность для лоббирования собственных интересов.

Наиболее актуальной практической проблемой в области работы с политической предубежденностью LLM является, пожалуй, возможность ее смягчения и контролирования. Независимо от того, насколько целенаправленно в модели прививаются предубеждения, косвенно они становятся инструментами политической борьбы. Модели могут распространять привитые им политические предубеждения, влияя как на самих пользователей, так и на аудитории, с которыми пользователи делятся сгенерированным моделью контентом. Кроме того, имплементация моделей в рекомендательные системы и автоматизированные системы модерации контента может привести к предубежденным рекомендациям и блокировкам контента и пользовательских профилей. Без эффективного регулирования модели могут стать (или, возможно, уже становятся) новым инструментом обретения политического влияния для компаний-разработчиков: так, последние могут либо перенастраивать модели в соответствии с интересами своих партнеров из политической сферы (в качестве сервиса), либо использовать их

как инструмент распространения собственных взглядов и влияния на политику через контроль над потоками информации. В свою очередь, осуществление эффективного регулирования может включать возможность контроля моделей на предмет их политической предубежденности, установления источника выявленных предубеждений, создание унифицированного стандарта для создания и обучения моделей, а также ответа на вопрос о возможности политической нейтральности моделей в целом. Для выбора наиболее оптимальных мер из указанных и их конкретизации до уровня осуществимых эффективных мер требуется научный фундамент в виде разноформатных исследований политических предубеждений LLM.

Особенно значимой исследовательской проблемой на данный момент, на мой взгляд, является установление связи между наблюдаемым предубеждением и его возможными источниками. Исследовательская проблема сопряжена с практической: лишь изучив связь между предубежденностью и ее источником, возможно эффективно ослаблять и контролировать проявления предубеждений. Вне практических целей она, однако, также имеет ценность. Без понимания того, каким образом манипуляции с источниками предубеждений могут быть эффективно использованы в целях пропаганды тех или иных взглядов, и того, как выявлять осознанное вмешательство в предубежденность алгоритмов, невозможно полноценно оценить меру и качество влияния LLM на политическую реальность. Изучение влияния технологий ИИ на сферу политики представляет собой другое важное направление политологических исследований, которое не обсуждается в данной статье.

Большинство научных работ, посвященных указанным проблемам, сосредоточено на выявлении наличия и характера политической предубежденности LLM [Motoki, Pinho Neto, Rodrigues 2023; Gover, 2023; Bang et al., 2024; и др.]. Авторские подходы значительно разнятся:

а) по качеству вопросов, на которые отвечают LLM, – более открытые запросы на написание эссе по заданной теме [Bang et al., 2024] и более закрытые вопросы в согласии с некоторыми тестами на политические взгляды [Motoki, Pinho Neto, Rodrigues, 2023; Rozado, 2023];

б) по количеству альтернативных наборов вопросов – к примеру, пятнадцать различных тестов [Rozado, 2023] или всего один [Motoki, Pinho Neto, Rodrigues, 2023];

в) по количеству рассматриваемых LLM – всего одна модель [см., напр.: Motoki, Pinho Neto, Rodrigues, 2023; Rozado, 2023; Liu

et al., 2021], три модели [Urman, Makhortykh, 2025] или одиннадцать [Bang et al., 2024];

г) по языку, на работу на котором настроены модели, – английский (у подавляющего большинства работ) или, к примеру, английский, украинский и русский [Urman, Makhortykh, 2025] и т.д.

Все эти подходы заслуживают внимания и будут подробно рассмотрены ниже. Однако, несмотря на различия, большинство из них преследуют аналогичные цели: а именно, выявление политических позиций LLM и разделяемых ими мнений по определенным политически значимым вопросам. Выявление политической предубежденности у LLM и ее характера представляет собой важное направление исследований и может быть эффективно включено в изучение других связанных с темой проблем. Тем не менее сосредоточение исключительно на обнаружении предубеждений может завести дискуссию в тупик в силу ограничений, присущих изучению политических взглядов моделей и описанных ниже. LLM не могут быть вовлечены в политическую жизнь напрямую. Осведомленности об их «мнениях» самой по себе недостаточно для понимания того, какое влияние предубежденные модели могут оказывать на политику. Как инструменты порождения и распространения информации, политические предубежденные модели являются скорее средствами пропаганды определенных взглядов, нежели их носителями. Следовательно, более важным оказывается изучение того, в какой мере и в каких случаях предубежденность моделей была внедрена целенаправленно, каким образом и в какой степени подобное внедрение в целом возможно, и осуществимо ли полное искоренение предубежденности в политически нагруженных материалах как таковых.

Таким образом, более важным нам кажется изучение меры возможного вмешательства человека в предубежденность моделей, в попытках либо смягчить ее, либо контролировать и направлять. Подобные исследования требуют постановки более сложных вопросов, связанных с промпт-инжинирингом, источниками предубеждений, тонкой настройкой (англ. fine-tuning)[1] моделей с целью регулирования степени и качества этих предубеждений, взаимодействием LLM с другими существующими технологиями ИИ. Для поиска ответов на эти вопросы требуется большее методологическое единство, нежели то, что мы можем наблюдать в форми-

[1] Описание процедуры тонкой настройки можно найти в пятом разделе данной статьи.

рующемся поле исследований в данный момент. Многие статьи создаются независимо друг от друга, что зачастую приводит к дублированию исследований в одних случаях, а в других – к игнорированию выводов предыдущих работ без должного обоснования. Кроме того, недостатки, на избегание которых обращали внимание одни авторы, могут повторяться другими. В попытке привести область исследований к большему единству, я описываю в статье имеющиеся достижения в области изучения политических предубеждений LLM и то, как они могут быть использованы в дальнейшем, а также намечаю ближайшие пути, по которым работа в этом русле может быть эффективно направлена. В описании имеющихся достижений я стремлюсь указать читателям на основные достоинства разбираемых подходов, но также подчеркиваю и их недостатки.

Систематизация имеющихся наработок также обретает значимость в контексте российского академического изучения политических предубеждений LLM. На данный момент, фокус внимания российского академического сообщества сосредоточен на предубеждениях LLM, лишь косвенно связанных с политикой. Тем не менее российские LLM, вроде YandexGPT и GigaChat, также не лишены политических предубеждений[1]. Кроме того, зарубежные модели имеют значительную популярность у российских пользователей, притом что их политические предубеждения могут иметь характер, противоречащий интересам Российского государства. Использование широкими кругами российских пользователей политически предубежденных моделей может оказывать влияние на политическую реальность, сопоставимое с любым другим форматом распространения политически предубежденной информации. Специфика российских академических исследований политических предубеждений LLM может заключаться в использовании уже имеющихся зарубежных наработок, дополнении их собственными, а также учетом специфических российских политических реалий. Подобная спецификация необходима для, во-первых, более осознанной работы с иностранными и российскими моделями, а во-вторых, для постановки новых вопросов, невозможных при рассмотрении моделей в ином политическом спектре, к примеру,

[1] Так, например, данные модели отказываются отвечать на вопросы, связанные со специальной военной операцией на Украине. Прямой отказ отвечать на некоторые вопросы также можно рассматривать как проявление предубежденности [Urman, Makhortykh, 2025].

нередко всплывающей в зарубежных исследованиях дихотомии демократы – республиканцы. Специфицированные таким образом исследования могут, помимо прочего, обогатить международную дискуссию о политических предубеждениях LLM постановкой оригинальных вопросов и получением оригинальных ответов. Для этого необходимо подробное рассмотрение специфики российских политических реалий. В данной статье фокус ограничивается рассмотрением наработок зарубежных коллег-ученых и обрисовкой возможных дальнейших траекторий исследований в целом, без привязки к национальной специфике.

Подходы к исследованию политической предубежденности больших языковых моделей

Существующие на данный момент исследования политических предубеждений LLM предлагают академическому сообществу различные подходы к изучению темы. Как видно из обзоров литературы, содержащихся в статьях, количество работ по теме остается довольно скудным, что, впрочем, неудивительно, учитывая ее новизну. Тем не менее существующие на данный момент подходы заслуживают внимания.

Почти все исследования проводятся с позиции стороннего наблюдателя, который создает запросы («промпты», от англ. “prompt”) некоторому чат-боту или ряду чат-ботов на основе LLM и изучает их ответы. Таким образом, основные различия между подходами заключаются в том, как конструируются промпты и / или как исследуются ответы, генерируемые моделями. Есть, однако, небольшое количество работ, в которых исследование проводится не с точки зрения стороннего наблюдателя. Я предлагаю разделить существующие подходы на три кластера:

1) подходы, в которых для создания промптов используются вопросы из анкет по политическим ориентациям, и LLM как бы самостоятельно заполняют анкету;

2) подходы, в рамках которых LLM просят составить развернутые ответы на политически нагруженные запросы, например, написать эссе по темам допустимости абортов или смертной казни. Для изучения подобных ответов используются методы анализа, применяемые в исследованиях развернутых текстов, вроде анализа тональности, выявления ложной информации, и т.д.;

3) и наиболее перспективные, но в то же время малопопулярные подходы, в которых производится манипуляция с возможными источниками политических предубеждений.

Только третий кластер лишен указанного недостатка стороннего наблюдателя. Разделение на кластеры я осуществлял на основе того, насколько глубоко анализируется зависимость между выявленной предвзятостью и ее потенциальным источником. Подходы первого кластера, как будет понятно далее, вовсе не принимают во внимание возможные источники, сосредоточиваясь на выявлении предубежденности и предоставляя читателю возможность самому делать выводы о ее возможной причине. Подходы второго кластера пытаются установить зависимость проявлений предубежденности от способа составления промптов. Третий кластер рассматривает аналогичную зависимость более фундаментально, прослеживая связь между возможным источником предубежденности и ее реальным проявлением. В целом разные направления исследований позволяют ставить различные вопросы, связанные с политическими предубеждениями LLM, и давать на них различные ответы. Однако все эти подходы могут дополнять и обогащать друг друга.

Политическая ориентация моделей по данным опросов и анкетирования

Наиболее распространенным подходом к изучению политических предубеждений LLM является установление наличия и характера предубежденности посредством использования анкет для выявления политической ориентации анкетируемого [Motoki, Pinho Neto, Rodrigues, 2023; Rozado, 2023; Pit et al., 2024; Durmus et al., 2023]. В качестве анкетируемого выступает при этом некоторая модель или группа моделей. Исследователи используют существующие опросники, посвященные определению политических взглядов людей, и применяют вопросы анкеты для создания промптов. Анкетный тип вопросов приводит к некоторой ограниченности ответов LLM: чаще всего от модели требуется выбрать один из предложенных вариантов ответа. Дальнейшая работа с ответами заключается в получении готовых результатов прохождения анкеты, либо в сравнении ответов модели с каким-либо образцом прохождения анкеты. Наиболее изысканные исследования данного кластера также учитывают случайность генерации ответов, так что процесс анкетирования модели повторяется несколько раз. Так,

Мотоки, Пино Нето и Родригес задают ChatGPT-3 (на данный момент устаревшей версии ChatGPT) вопросы из Political Compass 100 раз (в результате было получено 6200 ответов на 62 вопроса анкеты). Затем, с помощью линейной регрессии, они анализируют связь между результатами прохождения анкеты моделью в режиме по умолчанию, и результатами, полученными в режиме имитации моделью типичного представителя республиканских и демократических взглядов.

Основное различие работ данного кластера заключается в масштабе исследования. Например, Мотоки, Пино Нето и Родригес изучают всего одну модель одной версии при помощи одной анкеты (и второй анкеты – для подтверждения результатов, работу с которой они, однако, никак не освещают). Розадо использует уже 15 анкет, одна из которых составлена на испанском языке (хотя выбор анкеты и испанского языка никак не объясняется), но также задействует только одну модель одной версии [Rozado, 2023, p. 1, 2]. Дальнейшая работа в данном направлении требует увеличения масштабов всех параметров исследования: привлечение новых языков, моделей, их версий и режимов работы, анкет, увеличения циклов повторных прохождений и т.п. Предположительно, изменения масштаба исследования может приводить к различным результатам. Статья Розадо 2024 г. описывает существенно более проработанное исследование, включающее уже 24 модели, а также эксперимент с манипуляцией над источниками предубеждений в виде обучающих материалов. Это позволяет отнести ее к третьему кластеру подходов.

Что касается результатов исследований данного кластера, то все модели склоняются приблизительно к одним взглядам, а именно леволиберальным[1]. В зависимости от используемых анкет

[1] Точное определение того, что понимается под либеральными взглядами, требует контекстуализации в рамках некоторой эпохи и конкретной страны. Однако в общем смысле либерализм можно охарактеризовать через такие ключевые концепты, как ценность свободы, индивидуальности, прогресса, рациональности, уважение интересов общественности, социальности, и ограниченности власти [Freeden, 1996, p. 141–177]. «Левые» и «правые» взгляды также требуют аналогичной контекстуализации. В целом левые взгляды разделяют идеи равенства, поддержки рабочего класса, национализацию индустрии, противление иерархичности и национализму [см.: Brown G. W., McLean I., McMillan A. (eds). The concise Oxford dictionary of politics and international relations. Oxford: Oxford University Press, 2018, 597 p. – p. 321]. Опросы, в том числе используемые разбираемыми авторами, обычно отличают правые взгляды от левых по их негативному отношению к национализации и перераспределению благ, а также большей религиозностью [ibid., p. 486]. Левый

результаты могут, разумеется, качественно различаться: так, анкета, проверяющая ориентацию модели в рамках биполярной системы политических координат демократ – республиканец, будет показывать иные результаты, нежели более многосторонняя или содержательно полярная анкета. Тем не менее изученные на данный момент версии моделей преимущественно склоняются именно к левым взглядам.

К достоинствам подобных подходов к изучению политической предубежденности LLM следует отнести возможность работать с большими объемами полученных от различных LLM данных, легкость в получении статистически значимых результатов. Кроме того, привлечение готовых анкет освобождает исследователей от необходимости дополнительной экспертной оценки ответов моделей. Однако у подхода есть ряд ограничений и недостатков. Во-первых, использование готовых анкет делает исследователей зависимыми от качества и содержания анкет, к чему нередко добавляется невозможность узнать исходную методику подсчета результатов, если исследователи идут путем вбивания ответов модели в анкету. Во-вторых, исследования в рамках данного кластера не акцентируют внимание на том, как модели строят свои ответы и как различные формулировки промптов влияют на содержание и форму ответов. В-третьих, единственными предметами исследования могут становиться лишь выраженные моделями позиции (в очень ограниченном количестве возможных видов) и финальные результаты прохождения ими анкеты. Также имеются два дополнительных недостатка, на которых я сфокусируюсь подробнее.

Первый кластер исследований ничего не может сказать о связи потенциальных источников предубеждений и самими последними. В статье 2022 г. Д. Розадо поднимает тему возможных источников, опираясь на общие знания об обучении и корректировке работы LLM, однако даже не пытается указать, какие из них должны были вызвать открытую им предубежденность. Розадо выделяет три возможных источника: обучающие данные, особенности архитектуры алгоритмов и реакции пользователей и тестировщиков, оценивающих ответы модели. В работе А. Урман и М. Махортыха, которую я отнес к другому кластеру, выделяются

либерализм (или социальный либерализм) отличается от правого преимущественно позитивным отношением к вмешательству государства в экономику. Когда авторы говорят о левых (и леволиберальных) и правых взглядах моделей, чаще всего, вероятно, имеется в виду именно такое различение, хотя прямо это не обсуждается.

также фильтры, встраиваемые в модель ее разработчиками для корректировки выходных данных [Urman, Makhortykh, 2025]. Таким образом, исследования, построенные на прохождении моделями анкет, могут устанавливать факт и характер предубежденности, но не выявлять источник последней.

Ограниченность возможных рабочих промптов не позволяет даже устанавливать зависимость появления предубежденности от того, как сформулирован запрос, и данная ограниченность выступает второй центральной проблемой данного кластера. Поскольку модели нередко отказываются высказываться на политически нагруженные темы, исследователи пытаются составлять промпты, на которые модель точно давала бы ответ. Такие промпты обладают весьма специфическим характером и слабо совпадают с теми запросами, которые может отправлять исследуемым моделям обычный пользователь. Как отмечают Рёттгер и его соавторы, в большинстве работ этого направления используются ограниченные (“constrained”) промпты, вынуждающие модели выбирать один из возможных вариантов ответа [Röttger et al., 2024, p. 1], что приводит к невозможности нейтрального ответа со стороны модели [ibid., p. 3–4]. Главная проблема подобного подхода заключается в том, что реальные пользователи не взаимодействуют с LLM столь специфическим образом. Скорее, пользователи будут создавать запросы на выполнение заданий, например написание эссе, не интересуясь особенностями взаимодействия с используемой моделью. Как показывают авторы статьи, в работах этого направления могут применяться различные степени принуждения LLM к выдаче политически предубежденных ответов. Ответы моделей будут варьироваться в зависимости от наличия или отсутствия фраз принудительного характера и даже от степени их агрессивности [ibid., p. 4–5]. Кроме того, разные формулировки принуждающих фраз могут приводить к различающимся результатам [ibid., p. 5–6]. Если модели не принуждаются к выбору варианта ответа, многие из них отказываются отвечать на значительную часть вопросов (такие ответы авторы называют «недействительными», англ. “invalid”) [ibid., p. 4].

Вместо использования принуждающих промптов авторы предлагают сосредоточиться на таких запросах, которые бы допускали большую степень свободы в ответах LLM. Это может сделать эксперименты более похожими на реальный процесс использования моделей реальными пользователями [ibid., p. 6]. В результате собственного авторского эксперимента, в котором использовались

непринуждающие промпты, авторы установили, что ответы, которые модели давали на их запросы, часто были противоположны «вынужденным» ответам, когда LLM должны были выбрать один из предложенных вариантов [Röttger et al., 2024]. Исходя из этого авторы делают вывод, что предубеждения (или, следуя их терминологии, ценности и мнения) LLM должны изучаться в связи с конкретными реальными формами использования моделей, без общих умозаключений о мнениях и ценностях LLM как некоего единого целого[1], т.е. выводы должны включать описание конкретных приложений и промптов [ibid., p. 8].

Изучение политических предубеждений LLM с учетом особенностей написания промптов и ответов

Интерес к тому, как именно LLM отвечают на вопросы, затрагивающие политически значимые проблемы, в наибольшей степени проявляется в работах, исследующих свободные ответы моделей. Если предыдущие подходы использовали преимущественно те промпты, на которые LLM не отказывались выдавать ответ, то в этих подходах отказы и наличие ложной информации также рассматриваются как нечто показательное для политической предубежденности изучаемых технологий. Открытые ответы тоже позволяют изучать содержание и стиль информации на выходе, наличие или отсутствие ложной информации в качестве проявления предубежденности, а кроме того – устанавливать многоплановую зависимость ответов моделей от того, как формулируются запросы.

Изучение отказов позволяет установить, при использовании каких промптов модель отказывается давать ответы на политически значимые вопросы. Как показывается в статье А. Урман и М. Махортыха, отличия в отдельных параметрах запросов могут приводить к различным результатам: ответы на запросы (или отказ от вопросов), а также их содержание могут зависеть, например, от

[1] Вообще говоря, вопрос о том, можно ли рассматривать LLM как нечто цельное в отношении выражаемых ими позиций, еще требует изучения. Как показывают исследования со статистически значимыми количествами наблюдений, LLM действительно склонны выражать некоторую определенную и связную предубежденность со статистически значимой вероятностью. Однако правильнее было бы рассматривать их реакции не как мнение агента, но, скорее, как элементы выборок из генеральной совокупности всех возможных ответов моделей.

языка, на котором сформулирован промт [Urman, Makhortykh, 2025][1]. Можно предположить, что варьирование значений других параметров также будет влиять на характер ответов.

Различные модели могут отвечать на политически нагруженные запросы по-разному с точки зрения стиля и содержания. Запросы, допускающие большую свободу в ответах модели, открывают путь более глубокого анализа предубежденности последней. К примеру, такой подход позволяет запрашивать у LLM эссе на политически значимые темы и затем обрабатывать их релевантными методами контент-анализа, в частности, анализом тональности [Bang et al., 2024; Gover, 2023]. Тем самым изучению становятся подвластны новые аспекты политической предубежденности LLM, вроде мнения моделей по конкретным политически значимым вопросам в противовес одной лишь общей политической ориентации [Bang et al., 2024]. В контексте потенциального (а подчас – актуального[2]) внедрения LLM и иных видов ИИ в принятие политических или косвенно связанных с политикой решений такие тонкости могут оказаться чрезвычайно существенными.

Подходы, дозволяющие моделям давать открытые ответы, расширяют горизонты изучения политической предубежденности LLM, однако все еще ничего не говорят о связи выявляемых предубеждений с их возможными источниками. Урман и Махортых изучают содержание ответов, частоту отказов, и меру наличия ложной информации (вызванного, предположительно, скорее склонностью моделей «галлюцинировать», то есть сочинять ложные факты, нежели преднамеренным влиянием, хотя наверняка утверждать нельзя) как параметры измерения предубежденности моделей. Источником они считают фильтры, целенаправленно наложенные на модели их создателями. Тем не менее подлинная причина появления предубеждения может оказаться иной. Как и в

[1] В случае статьи Урман и Махортыха выбор языка влечет за собой выбор соответствующих регулятивных мер, принятых в стране, наиболее прочно ассоциируемой с данным языком, хотя установлению этого факта и посвящена статья. У различий в ответах в зависимости от выбранного языка, однако, могут быть и другие причины.

[2] Так, Google заявляли, что использовали свои модели в обеспечении борьбы с тем, что они подразумевают под «мизинформацией» (англ. “misinformation”), в ходе выборов в Европейский парламент в 2024 г. [см.: Kroeber-Riel A. Supporting the Elections for European Parliament in 2024. *Google Blog*. February 5, 2024. URL: https://blog.google/around-the-globe/google-europe/supporting-elections-for-european-parliament-2024/ (accessed: 13.10.2024)].

исследованиях предыдущего кластера, связь между потенциальными источниками предубеждений и самими предубеждениями формулируется на уровне догадок, гипотез – научное подтверждение своим утверждениям подходы данного кластера дать не могут.

Манипуляции с возможными источниками политических предубеждений LLM

Для исследования возможных источников политической предубежденности LLM необходимо работать с самими источниками, а не с последующими проявлениями предубежденности в генерируемом контенте. Учитывая практическую задачу коррекции предвзятости моделей и исследовательскую задачу установления связи между предвзятостью и её источником, только такой подход представляется способным дать достаточно полные ответы на соответствующие вопросы. По свидетельствам авторов статей двух рассмотренных выше направлений, источниками предубеждений могут быть: материалы, на которых обучаются LLM; фильтры, которые накладывают на модели их разработчики; предвзятость тех, кто оценивает ответы моделей и чьи оценки затем учитываются при корректировке их вывода; особенности архитектуры моделей. Определение подлинных причин возникновения политических предубеждений в данных, генерируемых LLM, а также установление связи между конкретными источниками и соответствующими предубеждениями в выходных данных, выходит за рамки возможностей описанных подходов. Единственное методологическое решение, которое мне видится, – это сосредоточение на манипуляциях с теми аспектами разработки LLM, которые, предположительно, являются причинами их предубеждений. Пока что, однако, это направление наименее разработано в академических исследованиях. За последние несколько лет появлялись статьи, посвященные изучению того, как манипуляции с обучающими данными для небольших языковых моделей могут влиять на их политические взгляды [Feng et al., 2023; Jiang et al., 2022]. В рамках таких исследований модели изначально обучались на предвзятом контенте. К сожалению, этот подход, хотя и весьма примечательный, не может быть использован для изучения больших языковых моделей, если только на подобные исследовательские цели не будут выделены большие ресурсы. Стоимость обучения новой LLM измеряется в миллионах долларов, которые вряд ли

будут выделены на проведение весьма узкоспециализированных экспериментов, вроде связанных с работой с политическими предубежденностями LLM.

Нерепрезентативность данных исследований для LLM – одна из причин, по которым работа с уже существующими LLM кажется более эффективной. Другая причина заключается в том, что существующие LLM – это модели, которые действительно используются миллионами людей по всему миру, и именно через них политические предубеждения LLM оказывают влияние на реальность, в частности, на индивидуальное сознание. Кроме того, исследование существующих LLM может также способствовать лучшему пониманию особенностей смягчения политических предубеждений LLM в целом (или манипулирования ими). К счастью, коды многих влиятельных LLM размещены в открытом доступе и, следовательно, доступны независимым исследователям. При этом, однако, данное направление изучения политических предубеждений LLM требует проведения междисциплинарных исследований, в которых будут задействованы, по меньшей мере, специалисты по компьютерным наукам и политологи. Это делает данный кластер более требовательным, но также более всесторонним. Требовательность данного направления (в которую включается также, к примеру, высокая стоимость соответствующих экспериментов), превышающая исследования других кластеров, вероятно, является причиной его слабой популярности.

Статья А. Агиза, М. Мостагира и Ш. Реда – одна из немногих работ, в которой описывается исследование в русле подходов из данного кластера [Agiza, Mostagir, Reda, 2024]. Чтобы оценить предубежденность исследуемого LLM (Mistral-7B-v0.2), авторы используют уже известные по предыдущим разделам подходы: они запрашивают у LLM ее мнение по поводу утверждений из Political Compass и задействуют другую модель (ChatGPT-4) для оценки того, что сгенерировала Mistral [ibid., p. 6–8]. Однако перед этим авторы проводят манипуляции с моделями, дважды «перенастраивая» их посредством тонкой настройки: один раз – на соответствие левым взглядам, и второй на соответствовали правым [ibid., p. 5–6, 7]. Перенастройка составляет главную методологическую ценность исследования.

Для осуществления тонкой настройки авторы дважды дообучают модель политически нагруженными датасетами, проведя их предварительный сбор, оценку (к слову, оценка снова осуществляется LLM – Mixtral-8x7B) и компоновку. Один датасет состоит

из данных, взятых с медиаплатформы правой направленности Truth Social, другой – из форума левой направленности Reddit Politosphere[1]. Эти датасеты далее используются для создания двух новых, состоящих из инструкций (т.е., образцов возможных запросов к модели) и образцовых ответов на эти и любые подобные им запросы. В одном случае модель дообучается[2] на республиканском датасете, состоящем из пар инструкция – ответ, в другом – на демократическом.

Два модифицированных таким образом варианта уже выпущенной на рынок LLM с открытым исходным кодом действительно оказываются политически предубежденными в соответствии с материалами, под которые модели подстраивались [Agiza, Mostagir, Reda, 2024, p. 7–10]. Полученные результаты являются весьма впечатляющими, доказывая, что существующие Большие языковые модели могут быть настроены на отражение определенных политических взглядов с помощью обучающих данных.

Обучающие данные, однако, являются не единственным возможным источником политических предубеждений LLM. Тщательного изучения требуют также другие возможные источники: цензурирующие фильтры, специфика устройства моделей и процессы доработок моделей с учетом оценок ее ответов со стороны специально подобранных команд. Дэвид Розадо, по сути, повторяет исследование Агиза, Мостагира и Реды, но увеличивает масштаб исследования до 24 моделей, 11 тестов для оценки предубежденности, и добавляет сравнение одинаковых моделей, прошедших и не прошедших специальные процедуры дообучения для настройки моделей на общение с живым пользователем [Rozado, 2024]. Последний аспект его исследования представляет наибольшую ценность, поскольку связывает появление предубеждений с контролируемым дообучением моделей, нацеленным на настройку их под работу с живыми пользователями. Однако эта ценность существенно подрывается тем, что контролируемое дообучение проводит не сам Розадо, в силу чего автор почти не имеет представления о фактическом содержании проведенных процедур. Как признает Розадо, такой дизайн исследования не позволяет ему сделать одно-

[1] Почему авторы относят Truth Social к правой направленности, а Reddit Politosphere – к левой, в статье не обсуждается.

[2] Для дообучения используется механизм низкоранговой адаптации (*англ.* Low-Rank Adaptation, сокр. LoRA).

значные выводы о связи предубеждений с дообучением [Rozado, 2024, p. 11].

Для полноценного изучения указанных вопросов необходимы эксперименты, построенные на непосредственной манипуляции с возможными источниками предубеждений. Такие эксперименты, в свою очередь, требуют междисциплинарных исследований, задействующих не только специалистов по компьютерным наукам, но также философов, политологов и других специалистов в области социальных наук. Отдельного изучения также требуют политические предубеждения на стыке LLM и других технологий генеративного ИИ, в частности, технологий генерации изображений на основе промптов, составленных на естественном языке, рекомендательных алгоритмов и т.д. Эти направления остаются пока что не затронутыми несмотря на то что некоторые подходы к изучению неполитических предубеждений в подобных технологиях начинают разрабатываться [Zhou et al., 2024].

Заключение

Исследования политических предубеждений LLM становятся особенно актуальными в свете необходимости регулирования данных технологий. С точки зрения регулирования дальнейшее продвижение в области ослабления, искоренения и контроля политических предубеждений в алгоритмических инструментах потребует предоставления большего доступа исследователей к существующим и активно использующимся технологиям. Также кажется необходимым создание специальных институтов, посвященных изучению ИИ на стыке компьютерных наук, этики, философии сознания, нейрокогнитивных и социальных наук, в том числе политологии[1]. В целом манипуляции с возможными источниками политических предубеждений LLM кажутся наиболее многообещающим направлением, в рамках которого можно поставить большое количество вопросов, недоступных для исследований из первых двух кластеров. Тем не менее последние также имеют

[1] Примером такого института можно считать Стэнфордский институт Человекоориентированного Искусственного Интеллекта (Stanford Institute of Human-Centered Artificial Intelligence, HAI) [см.: Stanford University. Human-Centered Artificial Intelligence. *Stanford University Website*. Not dated. URL: https://hai.stanford.edu/ (accessed: 13.10.2024)].

большое значение для изучения политических предубеждений LLM, и дальнейшее продвижение требует тесного сотрудничества наиболее изобретательных подходов из всех трех кластеров.

V.S. Muronets*

The basics yet ahead: an overview of the approaches to investigating political bias in large language models

Abstract. Political bias of Large Language Models has frequently become a topic for scientific investigation. Most of the researchers tend to compete in inventing more original ways of identifying bias rather than posing new research questions related to it besides "Is this model politically biased?" and "What is the character of its bias?". To properly evaluate possible influence of the models on the political reality and finding answers to some questions regarding regulation of Artificial Intelligence it is essential to be able to study the linkage between the bias and its cause. With regard to how the question of dependence between the identified bias and its possible source is addressed I have grouped the approaches to studying political bias of LLMs into three clusters: approaches that use political orientation surveys and questionnaires, studies devoted to investigating different ways of creating prompts and models' responses and their interdependence, and interdisciplinary research in which manipulations with possible sources of LLMs' political bias is conducted. The latter research trajectory seems to be the most promising one, despite its current unpopularity. Yet, it is impossible to advance in this trajectory without it being complemented by further developments in the approaches in the first two clusters. In studying the issue of LLM bias, not only computer science specialists but also philosophers and political scientists and other experts in the social sciences should be involved. Political biases at the intersection of LLM and other generative AI technologies – in particular, technologies for generating images based on prompts composed in natural language, recommendation algorithms, etc. – also require separate research. From a regulatory standpoint, further progress in mitigating, eradicating, and controlling political biases in algorithmic tools will require providing researchers with greater access to existing and actively used technologies. Furthermore, it appears necessary to establish specialized institutions dedicated to research on AI at the intersection of computer science, ethics, the philosophy of mind, neurocognitive and social sciences.

Keywords: Large language models, political bias, algorithmic bias, generative artificial intelligence, ethics of artificial intelligence, values and political views.

For citation: Muronets V.S. The basics yet ahead: an overview of the approaches to investigating political bias in large language models. Political science (RU). 2025, N 2, P. 204–226. DOI: http://www.doi.org/10.31249/poln/2025.02.09

* **Muronets Vsevolod**, HSE University (Moscow, Russia), e-mail: vmuronets@hse.ru

References

Agiza A., Mostagir M., Reda S. Analyzing the impact of data selection and fine-tuning on economic and political biases in LLMs. *arXiv preprint arXiv:2404.08699*. 2024. DOI: https://doi.org/10.48550/arXiv.2404.08699

Bang Y., Chen D., Lee N., Fung P. 2024. Measuring political bias in large language models: what is said and how it is said. *arXiv preprint arXiv:2403.18932*. 2024. DOI: https://doi.org/10.48550/arXiv.2403.18932

Davis R. *Typing politics: The role of blogs in American politics.* Oxford: Oxford university press, 2009, 241 p.

Durmus E., Nyugen K., Liao T.I., Schiefer N., Askell A., Bakhtin A., Chen C., Hatfield-Dodds Z., Hernandez D., Joseph N., Lovitt L. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388.* 2023. DOI: https://doi.org/10.48550/arXiv.2306.16388

Elmer G., Langlois G., McKelvey F. *The permanent campaign: new media, new politics.* Lausanne: Peter Lang, 2012, 144 p.

Farrell H., Drezner D.W. The power and politics of blogs. *Public choice.* 2008, N 134, P. 15–30.

Feng S., Park C.Y., Liu Y., Tsvetkov Y. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *arXiv preprint arXiv:2305.08283.* 2023. DOI: https://doi.org/10.48550/arXiv.2305.08283

Freeden M. *Ideologies and political theory: A conceptual approach.* Oxford: Clarendon Press, 1996, 604 p.

Gover L. Political bias in large language models. *The commons: Puget sound journal of politics.* 2023, Vol. 4 (1), N 2, P. 11–22.

Habermas J. Concluding comments on empirical approaches to deliberative politics. *Acta politica.* 2005, N 40, P. 384–392.

Habermas J. Between facts and norms: Contributions to a discourse theory of law and democracy. Hoboken: John Wiley & Sons, 2015, 631 p.

Hargittai E., Gallo J., Kane M. Cross-ideological discussions among conservative and liberal bloggers. *Public Choice*. 2008, N 134, P. 67–86.

Jiang H., Beeferman D., Roy B., Roy D. CommunityLM: Probing partisan worldviews from language models. *arXiv preprint arXiv:2209.07065.* 2022. DOI: https://doi.org/10.48550/arXiv.2209.07065

Kotek H., Dockum R., Sun D. Gender bias and stereotypes in large language models. In: Bernstein M., Savage S., Bozzon A. (eds). *Proceedings of The ACM collective intelligence conference.* 2023, P. 12–24.

Liu R., Jia C., Wei J., Xu G., Wang L., Vosoughi S. Mitigating political bias in language models through reinforced calibration. *Proceedings of the AAAI Conference on Artificial Intelligence.* 2021, Vol. 35, N 17, P. 14857–14866.

Malinova O.Yu. Symbolic policy: contours of the problem field. In: Malinova O.Yu. (ed.). *Symbolic politics:* Collection of articles. Moscow: INION RAS, 2012, P. 5–16. (In Russ.)

Motoki F., Pinho Neto V., Rodrigues V. More human than human: Measuring ChatGPT political bias. *Public Choice.* 2023, Vol. 198, N 1, P. 3–23.

Nadeem M., Bethke A., Reddy S. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456.* 2020. DOI: https://doi.org/10.48550/arXiv.2004.09456

Omiye J. A., Lester J. C., Spichak S., Rotemberg V., Daneshjou R. Large language models propagate race-based medicine. *NPJ Digital medicine.* 2023, Vol. 6, N 1, Article 195.

Papacharissi Z. The virtual sphere: The internet as a public sphere. *New media & society.* 2002, Vol. 4, N 1, P. 9–27.

Pit P., Ma X., Conway M., Chen Q., Bailey J., Pit H., Keo P., Diep W., Jiang Y.G. Whose side are you on? Investigating the political stance of large language models. *arXiv preprint arXiv:2403.13840.* 2024. DOI: https://doi.org/10.48550/arXiv.2403.13840

Rasmussen T. Internet and the political public sphere. *Sociology Compass.* 2014, Vol. 8, N 12, P. 1315–1329.

Rozado D. The political biases of ChatGPT. *Social Sciences.* 2023, Vol. 12, N 3, Article 148.

Rozado D. The political preferences of LLMs. *PLoS ONE.* 2024, Vol. 19, N 7, Article e0306621. DOI: https://doi.org/10.1371/journal.pone.0306621

Röttger P., Hofmann V., Pyatkin V., Hinck M., Kirk H.R., Schütze H., Hovy D. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. *arXiv preprint arXiv:2402.16786*. 2024. DOI: https://doi.org/10.48550/arXiv.2402.16786

Tulchinskii G.L. Power and making of meaning, or political pragmasemantics. *Political science (RU).* 2023, N 3, P. 151–169. DOI: http://www.doi.org//10.31249/poln/2023.03.07

Urman A., Makhortykh M. The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in ChatGPT, Google Bard (Gemini), and Bing Chat. *Telematics and Informatics.* 2025, Vol 96, Article 102211.

Zhou M., Abhishek V., Derdenger T., Kim J., Srinivasan, K. Bias in Generative AI. *arXiv preprint arXiv:2403.02726.* 2024. DOI: https://doi.org/10.48550/arXiv.2403.02726

Литература на русском языке

Малинова О.Ю. Символическая политика: контуры проблемного поля // Символическая политика: сб. науч. тр. Выпуск 1: Конструирование представлений о прошлом как властный ресурс / под ред. Малиновой О.Ю. – М.: ИНИОН РАН, 2012. – С. 5–16.

Тульчинский Г.Л. Смыслообразование и власть, или Политическая прагмасемантика // Политическая наука. – 2023. – № 3. – С. 151–169. DOI: http://www.doi.org/10.31249/poln/2023.03.07