DOI 10.35775/PSI.2025.118.6.016
УДК 32

В.В. КАРПОВА
аспирантка кафедры государственной политики
факультета политологии МГУ им. М.В. Ломоносова,
Россия, г. Москва
E-mail: karponika@yandex.ru

БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ КАК ДРАЙВЕР ГЛОБАЛЬНЫХ ТЕХНОЛОГИЧЕСКИХ ТРАНСФОРМАЦИЙ: ПОЛИТОЛОГИЧЕСКИЙ АНАЛИЗ НА ПРИМЕРЕ ПРОЕКТА SLAVA

В условиях стремительного развития искусственного интеллекта (ИИ) и глобальной цифровизации большие языковые модели (LLM) становятся не только инструментами генерации текста, но и акторами политической социализации. Статья посвящена анализу модели SLAVA – первого отечественного бенчмарка, направленного на оценку идеологической нейтральности LLM на материале гуманитарных дисциплин. На основе интеграции фрейм-анализа, семантического мониторинга и шкалы провокативности демонстрируется возможность диагностики идеологических сдвигов в поведении LLM и формируется концепт мировоззренческого суверенитета. Обосновывается прикладной потенциал SLAVA в сфере образования, государственного управления и гражданского сектора. Делается вывод о необходимости нормативной институционализации подобных инструментов как элементов гуманитарной экспертизы ИИ.

Ключевые слова: *цифровизация, глобальные технологические трансформации, искусственный интеллект, большие языковые модели, политическая социализация, бенчмарк SLAVA, когнитивное оружие, цифровой суверенитет.*

Современный этап развития общества характеризуется глубокой цифровой трансформацией, в центре которой находятся технологии искусственного интеллекта (ИИ), машинного обучения и нейросетевых алгоритмов [4. С. 47-55; 1. С. 15-22]. Эти технологии все активнее проникают в сферу общественно-политического управления, оказывая влияние на принятие решений, функционирование институтов, массовое сознание и механизмы политической социализации. В условиях экспоненциального роста объема цифровых данных и расширения сфер применения ИИ его изучение становится неотъемлемой частью политической науки.

ИИ определяется в научной литературе как совокупность технологий, обеспечивающих выполнение системами задач, ранее считавшихся исключительной прерогативой человека, от анализа информации и адаптивного

обучения до выработки решений в условиях неопределенности [3. С. 23-25]. Использование ИИ в политике обусловлено его способностью обрабатывать большие массивы данных (big data), прогнозировать поведение социальных групп, интерпретировать политические процессы и оптимизировать административные процедуры [6. С. 70-75; 14].

Наряду с очевидными преимуществами, такими как рост эффективности управления, персонализация коммуникации, развитие предиктивной аналитики возникают и новые вызовы. В научной литературе акцентируются проблемы алгоритмической предвзятости, эрозии прозрачности и роста влияния так называемой алгократии – модели управления, в которой алгоритмы вытесняют политические решения, минимизируя человеческий фактор [5. С. 418; 13]. Дополнительным риском становится аспект надзорного капитализма, при котором поведенческие данные граждан используются для манипуляции восприятием и вмешательства в процессы формирования политических установок [7; 2. С. 125-130; 9].

Одним из важнейших направлений трансформации политической сферы под воздействием ИИ становится изменение механизмов политической социализации. Политическая социализация традиционно рассматривается как процесс усвоения индивидом политических норм, установок и ролей, в результате которого формируется его политическое поведение и идентичность [8. С. 33-35]. Если ранее в качестве ее ключевых институтов выступали семья, школа, СМИ и религиозные структуры, то сегодня эти функции начинают выполнять алгоритмически управляемые цифровые среды: рекомендательные платформы, чат-боты, виртуальные ассистенты и большие языковые модели (LLM). Именно в этом контексте возникает необходимость критической оценки влияния LLM на процессы формирования политической идентичности, оценочных установок и интерпретационных рамок. ИИ перестает быть исключительно инструментом и становится участником идеологической коммуникации. Большие языковые модели, такие как GPT-4, Claude, GigaChat и др. изначально разрабатывались как инструменты генерации текстов, но в процессе интеграции в повседневную цифровую практику стали выполнять более сложные функции [16. P. 1-62]. Они не просто транслируют информацию, но и формируют когнитивную среду пользователя, предлагая интерпретации, смысловые акценты и ценностные рамки. Это особенно значимо в условиях растущего потребления алгоритмически сгенерированного контента – от образовательных платформ до виртуальных помощников.

Исследования показывают, что LLM могут оказывать влияние на восприятие политических событий, степень доверия к государственным институтам, интерпретацию исторических явлений и отношение к различным идеологическим парадигмам [11. С. 171-175]. При этом модели все чаще выступают в роли когнитивных посредников, выполняя функции, ранее присущие институтам социализации. Это позволяет говорить об их участии в идеологической

коммуникации – процессе передачи, воспроизводства и трансформации ценностных и политических установок в обществе.

Принципиальное отличие LLM от традиционных источников политической информации заключается в их персонализированной и диалоговой природе. Первостепенно, они адаптируют свои ответы к формулировке и интонации запроса, создавая эффект так называемого «когнитивного зеркала». Пользователь получает не объективную информацию, а отклик, наиболее вероятный с точки зрения паттернов, усвоенных моделью на этапе обучения. Во-вторых, модели функционируют в режиме непрерывного диалога, что сближает их с интерактивными формами обучения и воздействия, ранее недоступными для средств массовой коммуникации [12. С. 58-69].

Подобная интерактивность способствует усилению доверия к модели, которую пользователь нередко воспринимает как рационального и беспристрастного собеседника. Однако за фасадом «нейтральности» может скрываться сложная система идеологических предпочтений, встроенных в обучающие выборки, алгоритмы ранжирования и механизмы генерации.

В ответ на вызовы, связанные с идеологической прозрачностью и ценностной устойчивостью ИИ, в России был разработан исследовательский проект SLAVA (Sociopolitical Landscape and Value Analysis) [15]. Его цель заключается в разработке и апробации бенчмарка, способного выявлять когнитивные и мировоззренческие отклонения в поведении LLM, тестируя их на материале отечественных гуманитарных дисциплин. SLAVA позволяет не только оценивать фактологическую корректность ответов моделей, но и анализировать степень их идеологической нейтральности, соответствие локальному культурному коду и устойчивость к провокативному контексту. В отличие от сугубо технических бенчмарков, таких как MMLU, C-Eval, TruthfulQA, Pinocchio и др., SLAVA фокусируется на гуманитарной и политической чувствительности, вводит трехуровневую шкалу провокативности и реализует комплексную методологию анализа лексики, нарративов и интерпретационных шаблонов. Это позволяет не только фиксировать идеологически нагруженные отклонения, но и предложить метрику их оценки – от терминологических сдвигов до изменения фрейма ответа под влиянием контекста.

Бенчмарк SLAVA объединяет более 14 тысяч вопросов, сгруппированных по четырем направлениям: история, обществознание, география и политология. Выбор этих дисциплин неслучаен: они формируют основание гражданской и политической идентичности, закрепляют базовые представления о государстве, обществе, человеке и историческом процессе. Эти сферы наиболее чувствительны к идеологическим интерпретациям и потому критически важны при анализе поведения генеративных моделей.

Одной из ключевых методологических инноваций проекта стало введение трехуровневой шкалы провокативности, позволяющей классифицировать задания по степени социальной и смысловой чувствительности: первый уровень – нейтральные вопросы, проверяющие фактологическое знание; второй

уровень – дискуссионные формулировки, предполагающие наличие различных интерпретаций; третий уровень – вопросы, затрагивающие конфликтогенные, морально и политически поляризованные темы: национальная память, международные конфликты, религиозные традиции. Подобная градация позволяет зафиксировать изменения в стиле, терминологии и нарративах, используемых моделями при переходе от нейтральных к чувствительным темам. Она открывает путь к диагностике алгоритмического поведения: склонности к уклончивости, идеологической конформности, воспроизведению внешних фреймов. В процессе разработки бенчмарка была задействована экспертная аннотация, включающая участие специалистов по истории, политологии и социологии. Они осуществляли классификацию вопросов, оценивали корректность ответов моделей, интерпретировали лексико-семантические и смысловые отклонения. Это позволило совместить количественный и качественный подход: автоматически рассчитываемые метрики дополнялись содержательным анализом фреймов и логических структур.

Важно подчеркнуть, что SLAVA ориентирован не на измерение уровня знаний, а на поведенческий анализ LLM в гуманитарной сфере. Центральной задачей становится выявление устойчивых паттернов, свидетельствующих о наличии или отсутствии нейтральности, согласованности с культурным кодом и когнитивной устойчивости моделей. В этом смысле бенчмарк представляет собой не просто тестовую платформу, а создает основу для системной верификации поведения ИИ в политически чувствительном поле, формируя предпосылки для институционального аудита нейросетевых платформ, экспертной оценки их идеологической прозрачности и разработки стратегий цифрового суверенитета.

Одним из результатов реализации проекта SLAVA стало выявление способности больших языковых моделей не только воспроизводить знания, но и транслировать интерпретации, формирующие ценностный контекст. Это делает актуальным вопрос: насколько поведение LLM в гуманитарной сфере соответствует критерию идеологической нейтральности?

Традиционно под нейтральностью в научной и инженерной практике понимается отсутствие оценочности, однако в случае LLM данное понятие усложняется. Во-первых, модели обучаются на массивных текстовых корпусах, внутри которых уже содержатся идеологические установки и культурные предпочтения (так называемый bias-in-training). Во-вторых, даже формально «нейтральные» ответы могут воспроизводить фреймы, характерные для конкретной ценностной системы, особенно если они подаются как универсально рациональные. Проект SLAVA предложил системный подход к оценке мировоззренческого сдвига, то есть отклонения модели от нейтральной интерпретации в сторону определенной идеологической рамки. В рамках этого подхода использовались следующие исследовательские техники:

– контрастное тестирование – вопросы с нейтральной и оценочной формулировкой по одной теме (например, экономические реформы 1990-х

годов), что позволяет зафиксировать изменение терминологии, стилистики и эмоционального тона в зависимости от контекста;

– тестирование альтернативных нарративов – формирование заданий, основанных на различных исторических и политических традициях (например, трактовка событий Второй мировой войны в западном и российском дискурсах);

– семантический мониторинг – отслеживание частотности идеологически маркированной лексики в ответах моделей по отношению к социально-чувствительным темам: «демократизация», «империя», «освобождение»;

– фрейм-анализ – оценка того, какие структурные акценты делает модель в ответе: индивидуальные права, коллективная ответственность, моральный долг и т.д.

На основе этих методов формируется интегральный индекс идеологической нейтральности, включающий: коэффициент отклонения от локальной нормативной позиции; частоту использования терминов, характерных для внешней фреймовой системы; уровень стилистической уклончивости (выраженный через избегание прямых оценок на провокативные темы). Применение этой метрики показало, что даже высокоточные модели (по критерию фактологических совпадений) могут демонстрировать устойчивую идеологическую асимметрию. Например, в ответах на вопросы о советском периоде некоторые модели воспроизводили западные клише («тоталитаризм», «репрессивная машина») даже при нейтральной формулировке задания. При описании политической системы России встречались англо-американские конструкции, такие как «checks and balances», не соответствующие терминологии российского конституционного права. Вопросы о религии, патриотизме и семье интерпретировались через призму индивидуалистских либеральных ценностей, независимо от культурной рамки, заданной в вопросе.

Подобные проявления подтверждают гипотезу о том, что LLM действуют как переносчики фреймов, они не просто отвечают на вопрос, но и встраивают его в глобально доминирующие смысловые структуры. Это превращает проблему нейтральности из технической в мировоззренческую и социально-политическую.

Бенчмарк таким образом выполняет две ключевые функции: с одной стороны, он предоставляет инструментарий количественно-качественного анализа идеологического сдвига, а с другой формирует основу для более широкого концепта мировоззренческого суверенитета. Последний предполагает способность государства и общества защищать интерпретационные рамки, историческую идентичность и культурно-ценностные основания в условиях глобального информационного противостояния.

Коснемся и темы ИИ как когнитивного вектора воздействия. Термин «когнитивное оружие» используется в стратегических исследованиях и теориях гибридных конфликтов для обозначения инструментов, направленных на дестабилизацию когнитивных структур – памяти, ориентиров, ценностей, способности к рациональному суждению [10. С. 41-53]. В отличие от кибероружия, воздействие когнитивного оружия не разрушает инфраструктуру, но меняет

интерпретации и смысловое пространство, влияя на политическое поведение и общественное мнение.

Большие языковые модели, особенно построенные на англоязычных корпусах и оптимизированные глобальными технокомпаниями, потенциально могут быть инструментом когнитивной интервенции. Их ключевая особенность мимикрия под нейтральность. В отличие от СМИ или открытых идеологических платформ, LLM воспринимаются пользователями как объективные и безоценочные, что делает их воздействие более глубоким и трудно обнаружимым.

В этом контексте проект SLAVA можно рассматривать как систему обнаружения когнитивных интервенций, позволяющую выявлять устойчивые отклонения от локальных историко-культурных нарративов, фиксировать терминологическую подмену и изменение фрейма восприятия, а также диагностировать навязчивое воспроизводство внешних идеологических схем, например, либерально-индивидуалистской оптики в анализе коллективной идентичности.

Анализ данных SLAVA показывает, что даже при точных фактологических ответах LLM способны структурировать информацию в соответствии с англо-американским политическим и правовым дискурсом, в котором доминируют определенные представления о демократии, правах, государстве и человеке. Это формирует иллюзию универсальной рациональности, подменяющую многообразие культурных интерпретаций глобализированными шаблонами.

В условиях работы ИИ-систем в публичных, образовательных и экспертных средах недостаточно ограничиваться концептом информационного суверенитета, понимаемого как контроль над критически важными потоками данных. Возникает более глубокая задача – обеспечение мировоззренческого суверенитета, то есть способности общества и государства: формировать и защищать собственные интерпретации ключевых событий, понятий и ценностей; отслеживать проникновение внешне сконструированных фреймов; сохранять когнитивную автономию в условиях глобальной цифровой конкуренции. SLAVA предоставляет инструменты, позволяющие оценить степень соответствия LLM этим задачам. Он переводит абстрактные угрозы идеологического давления в конкретные, интерпретируемые и измеримые параметры – лексические сдвиги, стилистические деформации, отклонения от локальной нормативности. В результате модель становится не просто исследовательским инструментом, а элементом стратегического мониторинга цифрового ландшафта. Сопоставление результатов SLAVA с концептами когнитивного оружия и суверенитета позволяет утверждать, что нейросетевые платформы обладают потенциалом стратегического влияния на политическое сознание. В этой связи критически важно создание национальных инфраструктур оценки, аудита и верификации ИИ, способных обеспечить баланс между технологическим развитием и сохранением культурно-политической субъектности.

Бенчмарк SLAVA обладает значительным прикладным потенциалом, выходящим за пределы академических задач. Его данные и методология позволяют

использовать модель в качестве инструмента мониторинга, регулирования и просвещения в таких стратегически значимых сферах, как образование, государственное управление и развитие гражданского общества. В условиях усиления роли ИИ в различных институтах SLAVA становится платформой для встраивания нормативных ориентиров в алгоритмически управляемую среду. Одной из наиболее очевидных и социально значимых сфер применения SLAVA является система образования. В условиях, когда генеративные ИИ все активнее внедряются в школьную и вузовскую практику (в форме чат-ботов, учебных ассистентов, генераторов заданий и текстов), возрастает необходимость контроля точности и смысловой адекватности ответов, выдаваемых учащимся, оценки мировоззренческой нейтральности контента, особенно по гуманитарным предметам, а также формирования критического отношения к информации, полученной от ИИ. Для органов государственной власти результаты SLAVA могут стать основой для разработки механизмов цифрового суверенитета, включая экспертную оценку информационных рисков, связанных с применением зарубежных ИИ-систем и формирование стандартов верификации LLM, применяемых в госсекторе. Особую значимость приобретает использование бенчмарка в качестве индикатора соответствия ИИ национальным ценностям и нормативным основаниям. Это позволяет встроить бенчмарк в процедуры предварительной сертификации моделей, предохраняя государственные системы от когнитивных и идеологических сбоев.

Таким образом, рассматривая большие языковые модели как новые акторы политической социализации и когнитивного влияния, мы сталкиваемся с необходимостью переосмысления как научных, так и нормативных оснований взаимодействия человека и ИИ. В отличие от традиционных инструментов передачи информации, LLM не просто обслуживают процессы коммуникации, но активно участвуют в формировании фреймов, интерпретаций и смысловых структур. Это делает их не только технологическим, но и политико-философским феноменом. Представленный в статье бенчмарк SLAVA представляет собой первую в России инициативу по системной оценке поведения генеративных моделей в общественно-политической сфере. Его методология демонстрирует, что поведение ИИ может быть измерено, классифицировано и интерпретировано в рамках гуманитарной экспертизы. Ключевым понятием, возникающим в результате анализа, становится мировоззренческий суверенитет, как способность общества и государства защитить ценностно-мировоззренческие основания и обеспечить когнитивную устойчивость в условиях глобализированного ИИ-пространства.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК:

1. **Багдасарян В.Э., Балдин П.П.** Перспективы развития искусственного интеллекта в актуальной повестке политических и социальных рисков глобальных трансформаций // Журнал политических исследований. 2020. Т. 4. № 2.

2. **Бронников И.А., Карпова В.В.** Цифровое гражданство в Российской Федерации: политические риски и перспективы // Вестник Волгоградского государственного университета. Серия 4: История. Регионоведение. Международные отношения. 2021. Т. 26. № 3.
3. **Быков И.А.** Искусственный интеллект как источник политических суждений // Журнал политических исследований. 2020. Т. 4. № 2.
4. **Володенков С.В.** Цифровые актанты и вычислительная пропаганда как инструменты воздействия на массовое сознание в условиях глобальных технологических трансформаций // Вестник Московского университета. Серия 12. Политические науки. 2024. № 2.
5. **Володенков С.В., Федорченко С.Н., Печенкин Н.М.** Риски, угрозы и вызовы внедрения искусственного интеллекта и нейросетевых алгоритмов в современную систему социально-политических коммуникаций: по материалам экспертного исследования // Вестник Российского университета дружбы народов. Серия: Политология. 2024. Т. 26. № 2.
6. **Жуков Д.С.** Искусственный интеллект для общественно-государственного организма: будущее уже стартовало в Китае // Журнал политических исследований. 2020. Т. 4. № 2.
7. **Зубофф Ш.** Эпоха надзорного капитализма: битва за человеческое будущее на новых рубежах власти. М.: Издательство Института Гайдара, 2022.
8. **Пашкова Н.В.** Искусственный интеллект как источник «экзистенциального риска»: проблема самоидентификации общества и личности в условиях глобального виртуального пространства // Alma Mater (Вестник высшей школы). 2023. № 8.
9. **Срничек Н.** Капитализм платформ / пер. с англ. М. Добряковой. 2-е изд. М.: Изд. дом Высшей школы экономики, 2020.
10. **Федорченко С.Н.** Значение искусственного интеллекта для политического режима России: проблемы легитимности, информационной безопасности и «мягкой силы» // Вестник Московского государственного областного университета. Серия: История и политические науки. 2020. № 1.
11. **Четвергов А.С., Шарафетдинов Р.С., Полукошко М.М. и др.** SLAVA: бенчмарк социально-политического ландшафта и ценностного анализа // Труды ИСП РАН. 2025. Т. 37. Вып. 3.
12. **Цвык В.А., Цвык И.В.** Социальные проблемы развития и применения искусственного интеллекта // Вестник Российского университета дружбы народов. Серия: Социология. 2022. Т. 22. № 1.
13. **Hildebrandt M.** Algorithmic regulation and the rule of law // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018. Vol. 376. No. 2128. Article 20170355.
14. **Nemitz P.** Constitutional democracy and technology in the age of artificial intelligence // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2018. Vol. 376. No. 2133. Article 20180089.

15. SLAVA: Benchmark of the Socio-political Landscape and Value Analysis, открытая часть набора данных // https://huggingface.co/datasets/RANEPA-ai/SLAVA-OpenData2800-v1.
16. **Wang C. et al.** Survey on factuality in large language models: knowledge, retrieval and domain specificity // ACM Computing Surveys. 2018. Vol. 1. № 1.

V.V. KARPOVA
Graduate student, Department of Public Policy
Faculty of Political Science, Lomonosov Moscow State
University, Moscow, Russia

LARGE LANGUAGE MODELS AS A DRIVER OF GLOBAL TECHNOLOGICAL TRANSFORMATIONS: A POLITICAL SCIENCE ANALYSIS BASED ON THE SLAVA PROJECT

Amid the rapid advancement of artificial intelligence (AI) and global digitalization, large language models (LLMs) are increasingly functioning not only as tools for text generation but also as actors in political socialization. This article presents an analysis of SLAVA – the first domestic benchmark developed to assess the ideological neutrality of LLMs based on content from the humanities. Through the integration of frame analysis, semantic monitoring, and a provocation scale, the study demonstrates the potential for diagnosing ideological shifts in model behavior and introduces the concept of worldview sovereignty. The applied relevance of SLAVA is substantiated in the domains of education, public administration, and civil society. The article concludes by underscoring the necessity of the normative institutionalization of such tools as integral components of AI-related humanitarian expertise.

Key words: *digital transformation, global technological shifts, artificial intelligence, large language models, political socialization, SLAVA benchmark, cognitive influence operations, digital sovereignty.*