

МАШИНА ХАБЕРМАСА: СМОЖЕТ ЛИ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ ПРЕОДОЛЕВАТЬ ПОЛЯРИЗАЦИЮ МНЕНИЙ В ОБЩЕСТВЕ И БОРОТЬСЯ С СОЦИАЛЬНЫМ СТРЕССОМ?

© Лебедев А.Н.

Доктор психологических наук, Институт психологии РАН, Москва, Россия
lebedev-lubimov@yandex.ru; ORCID: 0000-0003-1030-9709

© Гагарина М.А.

Доктор психологических наук, Финансовый университет при Правительстве
Российской Федерации, Москва, Россия
mgagarina224@gmail.com; ORCID: 0000-0002-7812-7875

© Гордякова О.В.

Кандидат психологических наук, Институт психологии РАН, Москва, Россия
gordyakovaov@ipran.ru; ORCID: 0000-0003-3694-0734

© Неврюев А.Н.

Кандидат психологических наук, Первый МГМУ им. И.М. Сеченова, Москва, Россия
a.n.nevryuev@mail.ru; ORCID: 0000-0002-8640-9717

Исследование выполнено при финансовой поддержке Российского научного фонда (РНФ)
в рамках научного проекта № 23–18–00422, <https://rscf.ru/project/23-18-00422/>

Одна из наиболее крупных проблем современного общества — это информационные и социальные стрессы. Возникают они во многом вследствие ценностно-аффективной поляризации взглядов людей в социальных группах. Над решением проблемы поляризации работают специалисты по всему миру. Наиболее известным и широко обсуждаемым в настоящее время является подход, предложенный компанией Google DeepMind и получивший название «Машины Хабермаса» (*Habermas Machine*). Цель исследования состояла в оценке, какие высказывания и аргументы по поводу пользы или вреда от западных санкций оказываются для российской аудитории, и прежде всего молодых людей, более убедительными — известных специалистов в области политических и экономических наук или искусственного интеллекта? В исследовании приняло участие 168 респондентов. Результаты показывают, что поляризация мнений проявляется не в равномерной отрицательной или положительной оценке экспертных суждений, а в смещении акцентов. Политически «лояльные» респонденты придают большее значение исходным позитивным интерпретациям, тогда как «нелояльные» предпочитают критические версии, особенно когда они генерируются ИИ.

Ключевые слова: информационный стресс, социальный стресс, ценностно-аффективная поляризация социальных групп, Машина Хабермаса

Введение

В эпоху перехода цивилизации от однополярного мира к многополярному проблема прогнозирования и предотвращения конфликтных отношений как в малых социальных группах, так и в больших становится крайне актуальной. Противоречия между системами ценностей разных народов обостряются в связи с изменениями в сфере распространения и обмена информации

(Харари, 2019). Для решения этой проблемы проводятся многочисленные научные исследования в политических науках и разрабатываются технологии с применением искусственного интеллекта.

Одна из наиболее крупных проблем современного мира — это информационные и социальные стрессы, которые, как оказывается, тесно связаны. Если еще несколько десятилетий назад

люди, находящиеся в неких информационных ценностно замкнутых нишах, могли не обращать внимания на тех, кто думает иначе, не так себя ведет или имеет другую религию, то в настоящее время, в связи с тем, что информация, выражающая некую позицию или претензии на исключительность мировоззрения, распространяется по планете за несколько секунд, поляризация мнений в больших и малых социальных группах становится реальным и чуть ли не основным стрессовым фактором. Поэтому над методами преодоления негативных последствий ценностно-аффективной поляризации взглядов сегодня работают многие философы, политологи, ученые и специалисты IT-профессий.

Наиболее известным и широко обсуждаемым в настоящее время здесь является подход, предложенный компанией Google DeepMind и получивший название «Машины Хабермаса» (Habermas Machine). Данная система искусственного интеллекта предназначена для разрешения конфликтов и поиска согласия в группах с разными точками зрения (Kudiabor, 2024; Tessler et al., 2024).

Система была названа в честь немецкого философа Юргена Хабермаса, который в рамках своей теории коммуникативного действия утверждал, что уважительное общение на равных может приводить противодействующие группы к согласию. Юрген Хабермас полагает, что конфликтные отношения между людьми чаще всего определяются не содержанием проблемы, а неправильным способом ее обсуждения в поляризованных группах. По его мнению, любые разногласия можно преодолеть, если создать благоприятную среду для дискуссии, когда все участники получают равные возможности высказывать свое мнение и соглашаются действовать по общим правилам. Очевидно, что многие ученые с этим не согласятся, поскольку, например, в рамках отечественной науки принято считать, что как внутригрупповые, так и межгрупповые конфликты чаще всего имеют некие объективные основания и крайне редко возникают «на пустом месте».

Система «Машина Хабермаса» состоит из двух языковых моделей. Первая — генеративная — основана на модифицированной версии языковой модели Chinchilla, разработанной компанией DeepMind в 2022 году. Она создает несколько вариантов итогового заявления или аргумента на основе индивидуальных высказываний участников дискуссии. Вторая — модель оценки — анализирует высказывания каждого участника и предсказывает вероятность того, что каждый

из них согласится с предложенными генеративной моделью вариантами общего заявления.

Искусственный интеллект в этом случае выступает медиатором. Обработывая высказывания представителей противодействующих сторон, он находит нечто общее в аргументах и формулирует итоговую точку зрения, максимально снижающую вероятность несогласия или возможного конфликта. В процессе дискуссии ее участникам предлагается групповое заявление, которое, по оценкам системы, с наибольшей вероятностью будет принято всеми. Участники анализируют предложения системы и вносят свои замечания и предложения. Программа искусственного интеллекта, получая обратную связь, генерирует новые варианты заявления и этот процесс повторяется до тех пор, пока не будет достигнуто общее согласие.

В процессе разработки системы основная часть участников первого этапа исследования была привлечена через краудсорсинг. Несколько тысяч жителей Великобритании разделили на группы, каждой из которых было предложено обсудить одну из тем, связанных с актуальными вопросами политики правительства Великобритании. В частности, обсуждались такие темы, как облегчение въезда для беженцев и Брексит, возможность снижения возрастного ценза для голосования до 16 лет, необходимость приватизации Национальной службы здравоохранения Великобритании и другие. В этих случаях машина Хабермаса чаще всего превосходила людей-посредников. Были также сформированы контрольные группы, работавшие с модераторами-людьми.

Пилотажные исследования с помощью «Машины Хабермаса», проведенные DeepMind, показали, что ИИ обеспечивает более высокий уровень согласия участников дискуссии (56%) по сравнению с посредниками-людьми (44%). В этих экспериментах большинство участников предпочитали утверждения, сгенерированные ИИ, тем, которые были написаны людьми, и оценивали их как более аргументированные. Также они чаще меняли свои взгляды после обсуждения в сторону общего мнения, в формировании которого участвовал ИИ (Tessler et al., 2024).

Постепенно исследования компании DeepMind подверглись критике, которая, в частности, состояла в том, что мнения участников краудсорсинга не отражают мнения всего населения Великобритании по рассматриваемым вопросам, а значит практическая ценность программы ока-

зывается весьма спорной. Поэтому компания собрала группу из 200 человек, отражающую разнообразие британского общества (по возрасту, этнической принадлежности, социальному статусу и др.). Эта группа была разделена на более мелкие подгруппы по три человека, которые рассматривали и обсуждали девять вопросов. На этой более репрезентативной выборке были получены аналогичные результаты. В частности, после дискуссии уровень поддержки идеи сокращения заключенных в тюрьмах вырос с 60% до 75%. Также увеличилась поддержка облегчения въезда для беженцев (с 39% до 51%) и поощрения национальной гордости (с 42% до 57%). Участники сошлись во мнениях по пяти из девяти вопросов, хотя по некоторым, например, по вопросам Брексита, разногласия все-таки сохранились (Kudabor, 2024; Tessler et al., 2024).

Особенность Машины Хабермаса состоит в том, что эта программа не просто обобщает высказываемые точки зрения, в отличие от стандартной функции резюмирования, но анализирует множество точек зрения и стремится выделить общие идеи, способствующие достижению консенсуса. Более того, при разработке программы вопросы для обсуждения отбирались не случайным образом, а проходили проверку на предмет потенциально провокационных тем. Ее создатели стремились к тому, чтобы минимизировать риск оскорбительных высказываний участников дискуссии. То есть поляризация мнений в группах всегда оставалась на уровне обмена ценностными суждениями и никогда не переходила на этап аффективной. Разработчики осознанно избегали провокативных тем и острых дискуссий, чтобы получить наиболее предсказуемые результаты.

Таким образом, в том варианте медиации, который предложила компания Google DeepMind, разработав Машину Хабермаса, нет возможности решать проблемы, которые возникают во взаимоотношениях групп, находящихся в условиях крайне аффективной поляризации мнений и выясняющих отношения в агрессивной и уничижительной форме. Следовательно, возможности снижения уровня агрессии в обществе и социального стресса на основе данной программы весьма ограничены.

Когда в 2022 году компания OpenAI выпустила ChatGPT, Илон Маск выразил мнение и обеспокоенность большой группы авторитетных в сфере IT-специалистов по поводу того, что ИИ может использоваться политиками для манипулирования мнением населения в политической борьбе,

например, потому, что ИИ чаще поддерживает более прогрессивные и либеральные взгляды. Однако создатели Машины Хабермаса считают, что предвзятость ИИ зависит от того, каким образом формулируются вопросы. При выборе ответа из готовых вариантов ИИ часто склоняется к более прогрессивным взглядам. Когда он имеет больше самостоятельности, то его ответы и предложения выражают более нейтральную позицию.

Это ограничение было учтено при разработке Машины Хабермаса. Поэтому вместо обобщения одного фрагмента текста, предоставленного одним участником дискуссии, машина Хабермаса обобщает несколько фрагментов текста, предоставленных несколькими участниками, и выделяет общие идеи, находя точки пересечения. Цикл продолжается до тех пор, пока групповое заявление не будет одобрено всеми участниками дискуссии.

Эмпирическое исследование доверия граждан с разными политическими установками экспертам и искусственному интеллекту

Цель нашего исследования состояла в том, чтобы оценить, какие высказывания и аргументы по поводу пользы или вреда от западных санкций оказываются для российской аудитории, и прежде всего молодых людей, более убедительными — известных специалистов в области политических и экономических наук или искусственного интеллекта?

Метод

Для проведения исследования была разработана авторская анкета, включающая: 1) вопросы по социально-демографическим данным респондентов; 2) суждения экспертов, предлагаемые к оценке респондентам на степень согласия от 1 до 5; 3) комментарии ИИ по суждениям экспертов, предлагаемые к оценке респондентам на степень согласия от 1 до 5; 4) два вопроса на отношение к антироссийским санкциям, для разделения респондентов на группы; 5) вопросы на доверие ИИ, разработанные на основе модели формирования цифрового доверия — шесть вопросов, например, «Я считаю, что информация, предоставляемая сервисами ИИ, правдива» (Тункевичус, Ребязина, 2023); 7) вопрос на удовлетворенность жизнью (Насколько Вы удовлетворены вашей жизнью? От 1 до 10, где 1 — абсолютно не удовлетворен).

Суждения экспертов, предлагаемые для оценки респондентам, были отобраны из открытых

источников. Всего использовалось 10 суждений, высказанных экспертами в СМИ:

1. Н. Талеба, современного экономиста, математика, философа и писателя, автора книги «Черный лебедь. Под знаком непредсказуемости», вышедшей в 2007 году и содержащей прогноз финансового кризиса 2008 года. В опросе использовался фрагмент «О пользе санкций для России», опубликованный Forbes 23 марта 2018 г. (Известный экономист Нассим Талеб рассказал..., 2018).

2. И. Липсица¹, российского экономиста, маркетолога, автора публикаций в экономической прессе и первого в России учебника по рыночной экономике для подростков «Удивительные приключения в стране Экономика». Предлагался фрагмент «Российская экономика сейчас проедает будущее и тратит не по средствам», опубликованный в «Новой газете»² 08 ноября 2024 г. (Российская экономика сейчас..., 2024).

3. С. Бодрунова, директора ИНИР им. С.Ю. Витте, президента Вольного экономического общества России, член-корреспондента Российской академии наук. Представлен фрагмент «Экономист Бодрунов: антироссийские санкции способствовали развитию РФ», опубликованный ТАСС 06 мая 2025 г. (Экономист Бодрунов: антироссийские санкции способствовали..., 2025).

4. К. Каллас, эстонского юриста, политического и государственного деятеля. Предлагался фрагмент «Кая Каллас: планируем принять новый пакет санкций против РФ 20 мая», опубликованный «Деловыми ведомостями» 12 мая 2025 г. (Кая Каллас: планируем принять новый пакет..., 2025).

5. А. Дудчака, эксперта Института стран СНГ. Использовался фрагмент «Александр Дудчак: Санкции против России могут резко улучшить нашу экономику», опубликованный в «Украина.ру» 02 марта 2022 г. (Александр Дудчак: Санкции против России могут резко улучшить..., 2022).

6. Результаты ноябрьского опроса Российского союза промышленников и предпринимателей по теме: «Последствия введения санкций для российского бизнеса», опубликованные на сайте в разделе «Аналитика» (Результаты опроса..., 2024).

7. Р. Нельсон, представителя Congress.gov — онлайн-базы законодательной информации Конгресса США (Nelson, 2017).

8. Я. Кедми, Военно-политического эксперта. Представлен фрагмент ««Силенок не хватит»:

Кедми раскрыл, как Путин сделал Россию неуязвимой для новых санкций США», опубликованный Информационно-аналитическим изданием «Мировое Политическое Шоу» 09 мая 2025 г. («Силенок не хватит»: Кедми раскрыл..., 2025).

9. М. Сергеева, зав. отделом экономики «Независимой газеты». Предлагался фрагмент «От санкций уже пострадало 70% российского бизнеса», опубликованный НГ 15 января 2025 г. (От санкций уже пострадало..., 2025).

10. А. Зубца, экономиста. Представлен фрагмент «Экономист Зубец: Главный позитивный эффект от санкций — новая идеология развития, опирающаяся на внутренние ресурсы», опубликованный «Российской газетой» 10 марта 2023 г. (Экономист Зубец: Главный позитивный эффект от санкций..., 2023).

Все суждения были подвергнуты анализу с помощью ИИ. В качестве программы ИИ, применяемой в эксперименте, была выбрана система ИИ DeepSeek, которая считается аналогом ChatGPT и разработана в Китае.

Полученные комментарии ИИ по каждому суждению эксперта были разбиты на смысловые блоки и предложены для оценки респондентам.

В исследовании приняло участие 168 респондентов (студенты вузов). Из общего пула респондентов на основании степени согласия с вопросами: «Западные санкции приносят пользу и способствуют развитию нашей страны» и «Я считаю, что все проблемы в нашей стране имеют временный характер, и страна развивается в правильном направлении», были выделены две группы респондентов «лояльные» — согласные с тем, что страна движется в правильном направлении и санкции способствуют ее развитию (N=40, из них 65% женщин и 35% мужчин) и «нелояльные» — несогласные с этими утверждениями (N=39, из них 54% женщин и 46% мужчин).

Результаты и их анализ

В таблице 1 приведена описательная статистика результатов оценок высказываний экспертов всеми респондентами (лояльными и нелояльными), а также значение величины эффекта Коэна (Cohen's d), указывающее на силу различий между группами лояльных и нелояльных респондентов в оценке мнений экспертов. Распределение оценок оказалось близким нормальному по большинству измеряемых характеристик.

¹ Признан Минюстом РФ Иностранным агентом

² Издание признано Минюстом РФ Иностранным агентом

Таблица 1. Описательные статистики и значения эффекта Коэна оценок суждений экспертов лояльными и нелояльным респондентами

	Эксперты	Общая выборка (N=168)		«Нелояльные» (N=39)		«Лояльные» (N=40)		Значение эффекта Коэна
		М	σ	М	σ	М	σ	
1	Высказывание Н. Талеба	3,52	1,01	3,13	1,13	4,15	0,7	-1,11
2	Высказывание С. Бодрунова	3,45	1,07	3,00	1,08	3,90	1,01	-0,86
3	Результаты, полученные экспертами РСПП	3,4	1,06	3,38	1,11	3,45	1,18	-0,06
4	Высказывание М. Сергеева	3,32	1,02	3,46	1,05	3,20	1,02	0,25
5	Высказывание Р.М. Нельсон	3,29	1,06	3,36	1,06	3,30	1,07	0,06
6	Высказывание Я. Кедми	3,21	1,13	2,72	1,10	3,43	1,22	-0,61
7	Высказывание А. Зубца	3,19	1,03	2,56	0,85	3,70	0,91	-1,30
8	Высказывание А. Дудчака	3,08	1,10	2,44	1,12	3,38	1,05	-0,87
9	Высказывание И. Липсица	2,96	1,20	3,38	1,25	2,90	1,17	0,40
10	Высказывание К. Каллас	2,55	1,22	2,72	1,17	2,30	1,20	0,35

Примечания: М — среднее значение, σ — стандартное отклонение, d — значение эффекта Коэна

Самым популярным высказыванием в общей выборке респондентов является мнение Н. Талеба, суть которого сводится к тому, что некоторые экономические сектора России выиграют от введения санкций, а нефтяной шок, вызванный снижением цен на нефть, может быть даже полезным, выступая импульсом для развития, и итоговым суждением является идея, что надо отделить российскую экономику от нефти. Суждение Н. Талеба занимает первую строчку из десяти в рейтинге оценок высказываний экспертов лояльными респондентами и лишь пятую среди нелояльных.

Самым непопулярным среди респондентов общей выборки является высказывание К. Каллас, содержательно сводящееся к необходимости усиления санкционного давления на Россию. В группе лояльных респондентов это суждение также является самым непопулярным, а среди нелояльных респондентов оно занимает восьмое место.

Наибольшую степень согласия нелояльные респонденты выразили по отношению к мнению Н. Сергеева, утверждающего, что «санкционное давление остается значимым негативным фактором для большинства отечественных предприятий, и никаких признаков улучшения ситуации в этой сфере пока не просматривается, признают эксперты» (От санкций уже пострадало..., 2025), а самую низкую — к мнению А. Дудчака, под-

черкивающего возможность избавиться от обязательств перед международными организациями и обретения независимости объема денежной массы в российской экономике от валютной выручки (Александр Дудчак: Санкции против России могут резко улучшить..., 2022).

Сравнительный анализ оценок одних и тех же экспертов в группах респондентов выявил не только различия по уровню восприятия, но и принципиальные расхождения в структуре предпочтений. Расчет размеров эффекта (Cohen's d) показал, что наиболее выраженные различия наблюдаются в отношении высказывания Н. Талеба ($d > 0,9$, значимо выше у «лояльных») и С. Бодрунова ($d \approx 0,8$, также выше у «лояльных»), тогда как в случае И. Липсица и К. Каллас оценки «нелояльных» оказались выше.

Корреляционный анализ профилей средних оценок показал отсутствие согласованности между группами респондентов. Коэффициент Пирсона оказался практически нулевым ($r = -0,01$; $p = 0,98$), что свидетельствует об отсутствии линейной связи между уровнями оценок «лояльных» и «нелояльных». Анализ ранговой корреляции Спирмена также не выявил значимых зависимостей ($r = -0,21$; $p = 0,56$).

Таким образом, можно заключить, что эффект поляризации проявляется не только в уровне оценок, но и в их структурном распределении: «ло-

яльные» склонны поддерживать фигуры, ассоциированные с внешним позитивным дискурсом (например, Н. Талеб), тогда как «нелояльные» при-

дают большее значение критически настроенным отечественным экспертам (например, И. Липсиц). Визуально картина представлена на рисунке 1.

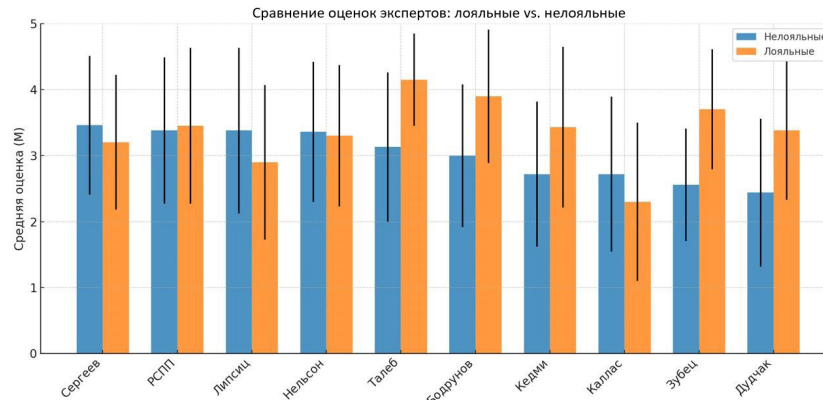


Рис. 1. Сравнение оценок высказываний экспертов лояльными и нелояльными респондентами

Объем статьи не позволяет нам рассмотреть все оценки суждений экспертов ИИ, а также отношение респондентов к оценкам ИИ. Далее будет проведено сравнение мнений лояльных и нелояльных респондентов по отношению к высказыванию Н. Талеба и оценкам различных аспектов его высказывания ИИ. Суждение Н. Талеба было

выбрано ввиду того, что в отношении именно этого суждения выявлен самый большой эффект.

В таблице 2 в качестве примера приведена описательная статистика и сравнение средних значений по t-критерию суждения Н. Талеба и его оценок ИИ лояльными и нелояльными респондентами.

Таблица 2. Описательные статистики и результаты сравнения средних значений суждения Н. Талеба, его оценок ИИ, доверия ИИ и удовлетворенности жизнью лояльными и нелояльными респондентами

Суждения	«Лояльные» (N=40)		«Нелояльные» (N=39)		p	
	М	σ	М	σ		
Статистики						Вывод
Н. Талеб (ориг.)	4,15	0,70	3,13	1,13	***	Значимо ↑ у лояльных
ИИ: Преувеличение выгод санкций	3,45	0,93	3,64	1,27	n.s.	Различий нет
ИИ: сравнение с Норвегией	3,33	0,92	3,92	1,01	**	Значимо ↑ у нелояльных
ИИ: нефтяной шок	3,05	1,13	3,82	1,07	**	Значимо ↑ у нелояльных
ИИ: уязвимость Саудовской Аравии	3,25	1,28	3,33	1,13	n.s.	Различий нет
ИИ: структурные проблемы России	3,28	1,20	4,08	1,04	**	Значимо ↑ у нелояльных
ИИ: вывод Талеба (оптимизм упрощает)	3,18	1,13	4,15	0,96	***	Значимо ↑ у нелояльных
ИИ: средняя оценка	3,25	1,10	3,82	1,08	**	Значимо ↑ у нелояльных
Доверие ИИ	22,28	4,06	22,54	3,68	n.s.	Различий нет
Удовлетворенность жизнью	8,15	1,27	6,64	1,88	***	Значимо ↑ у лояльных

Примечания: М — среднее значение, σ — стандартное отклонение, p — уровень значимости различий по критерию Стьюдента, n.s. — различия статистически не значимы, ** $p \leq 0,01$, *** $p \leq 0,001$.

Следует обратить внимание на наличие достоверных различий по степени удовлетворенности жизнью среди лояльных и нелояльных респондентов. При этом значения лояльных находятся в области зоны высоких значений, а нелояльных — средних. К сожалению, дизайн исследований не позволяет делать выводы ни о состоянии респондентов, ни о причинно-следственных связях, однако мы можем полагать, что удовлетворенность жизнью и интерпретация происходящих событий связаны. Так, анализ корреляций показал, что чем больше респонденты убеждены, что западные санкции приносят пользу и способствуют развитию нашей страны, тем больше они верят, что все проблемы в нашей стране имеют временный характер, и страна развивается в правильном направлении ($r=0,41^{***}$), и наоборот. Обе переменные также положительно связаны с удовлетворенностью жизнью ($r=0,28^{***}$; и $r=0,37^{***}$).

Полученные результаты демонстрируют наличие устойчивых различий в оценках высказываний экспертов и их интерпретаций между группами респондентов. Респонденты, отнесенные к группе «лояльных», статистически значимо выше оценили исходное высказывание Н. Талеба ($p \leq 0,01$), что может интерпретироваться как тенденция поддерживать позитивные интерпретации в дискурсе о санкциях и перспективах российской экономики.

В противоположность этому группа «нелояльных» респондентов систематически демонстрировала более высокие оценки критических интерпретаций, предложенных ИИ: сравнений России с Норвегией ($p \leq 0,01$), рассуждений о последствиях нефтяного шока ($p \leq 0,01$), указаний на структурные ограничения экономики ($p \leq 0,01$) и заключения о чрезмерном оптимизме Талеба ($p \leq 0,01$). Таким образом, именно критический пласт интерпретаций оказался более созвучным их восприятию.

Полученные результаты отчасти согласуются с результатами других исследователей, получив-

ших, что российская идентичность увеличивает политическую поляризацию между людьми, которые поддерживают и не поддерживают то, что происходит в стране (Гулевич, Косимова, 2024). Отсутствие значимых различий по пунктам «преувеличение выгод санкций» и «уязвимость Саудовской Аравии» указывает на наличие общих точек консенсуса, где позиции респондентов сближаются. При наличии достоверных различий в оценках высказываний самого эксперта, лояльные и нелояльные респонденты различаются и по степени согласия с комментариями ИИ.

Выводы

1. По результатам исследования можно сделать вывод, что именно оптимизм в отношении будущего себя и своей страны («страна движется в правильном направлении, трудности носят временный характер») позволяет справляться с текущими вызовами и поддерживает приемлемый уровень удовлетворенности жизнью.

2. По результатам методики «Доверие к ИИ» лояльные и нелояльные респонденты имеют одинаковый уровень доверия к ИИ, то есть предположение о различиях в содержательной оценке суждений ИИ правомерно.

3. В совокупности эти данные позволяют заключить, что поляризация мнений проявляется не в равномерной отрицательной или положительной оценке экспертных суждений, а в смещении акцентов. А именно: для группы «лояльных» респондентов более значимыми оказываются исходные позитивные интерпретации, тогда как для группы «нелояльных» — критические версии, особенно когда они генерируются ИИ. Это указывает на потенциальную роль ИИ как фактора, способного усиливать существующие границы в общественных установках, структурируя восприятие в зависимости от исходных ориентаций аудитории.

Литература:

1. Гулевич О.А., Косимова С.С. Связь российской идентичности и политической поляризации: роль надежной национальной идентификации и национального нарциссизма // Социальная психология и общество. 2024. № 15(4). С. 123–139. DOI: 10.17759/sps.2024150409
2. Тункевичус Э.О., Ребязина В.А. Многомерная модель формирования цифрового доверия пользователей цифровых сервисов: результаты эмпирического исследования // Вестник Московского университета. Серия 6. Экономика. 2023 № 4. С. 165–200. DOI: 10.55959/MSU0130-0105-6-58-4-8
3. Харари Ю.Н. 21 урок для XXI века. М.: Синдбад, 2019.
4. Kudabor H. AI tool helps people with opposing views find common ground // Nature. 17 October 2024. DOI: 10.1038/d41586-024-03424-z URL: <https://www.nature.com/articles/d41586-024-03424-z>
5. Nelson R.M. U.S. Sanctions and Russia's economy (Санкции США и экономика России) // Congressional research service. 17.02.2017. URL: <https://fas.org/sgp/crs/row/R43895.pdf>

6. *Tessler M. H., Bakker M. A., Jarrett D., et al.* AI can help humans find common ground in democratic deliberation // *Science*. 2024. V. 386. Issue 6719. URL: <https://www.science.org/doi/10.1126/science.adq2852>
7. Александр Дудчак: Санкции против России могут резко улучшить нашу экономику // *Украина.ру*. 02.03.2022. URL: <https://ukraina.ru/20220302/1033415109.html>
8. Известный экономист Нассим Талеб рассказал, чем санкции полезны для России // *Forbes. Бизнес*. 23.03.2018. URL: <https://www.forbes.ru/biznes/359061-izvestnyy-ekonomist-nassim-taleb-rasskazal-chem-sankcii-polezny-rossii>
9. Кая Каллас: планируем принять новый пакет санкций против РФ 20 мая // *Деловые ведомости*. 12.05.2025. URL: <https://www.dv.ee/novosti/2025/05/12/kaja-kallas-planiruem-prinjat-novyj-paket-sankcij-protiv-rf-20-maja>
10. От санкций уже пострадало 70% российского бизнеса // *Независимая Газета. Экономика*. 15.01.2025. URL: https://www.ng.ru/economics/2025-01-15/1_9171_sanctions.html
11. Результаты опроса «Последствия введения санкций для российского бизнеса» // *Российский союз промышленников и предпринимателей*. 2024. URL: <https://rspp.ru/activity/analytics/rezultaty-oprosa-posledstviya-vvedeniya-sanktsiy-dlya-rossiyskogo-biznesa/>
12. Российская экономика сейчас проедает будущее и тратит не по средствам // *Новая газета*. 11.08.2024. URL: <https://novayagazeta.ru/articles/2024/11/08/rossiiskaia-ekonomika-proedaet-budushchee-i-tratit-ne-po-sredstvam>
13. «Силенок не хватит»: Кедми раскрыл, как Путин сделал Россию неуязвимой для новых санкций США // *MPSH.RU*. 09.05.2025. URL: <https://mpsh.ru/31318-silenok-ne-hvatit-kedmi-raskryl-kak-putin-sdelal-rossiju-neujazvimoj-dlja-novyh-sankcij-ssha.html>
14. Экономист Бодрунов: антироссийские санкции способствовали развитию РФ // *ТАСС*. 6.05.2025. URL: <https://tass.ru/ekonomika/23870687>
15. Экономист Зубец: Главный позитивный эффект от санкций — новая идеология развития, опирающаяся на внутренние ресурсы // *Российская газета*. 10.03.2023. URL: <https://rg.ru/2023/03/10/ekonomist-zubec-glavnyj-pozitivnyj-effekt-ot-sankcij-novaia-ideologiya-razvitiia-opiraiushchaisia-na-vnutrennie-resursy.html>

THE HABERMAS MACHINE: WILL ARTIFICIAL INTELLIGENCE BE ABLE TO OVERCOME THE POLARIZATION OF OPINIONS IN SOCIETY AND COMBAT SOCIAL STRESS?

© Aleksandr N. Lebedev

Doctor of Psychology, Institute of Psychology of the Russian Academy of Sciences, Moscow, Russia
lebedev-lubimov@yandex.ru; ORCID: 0000-0003-1030-9709

© Mariya A. Gagarina

Doctor of Psychology, Financial University under the Government of the Russian Federation, Moscow, Russia
mgagarina224@gmail.com; ORCID: 0000-0002-7812-7875

© Olga V. Gordyakova

Candidate of Psychology, Institute of Psychology of the Russian Academy of Sciences, Moscow, Russia
gordyakovaov@ipran.ru; ORCID: 0000-0003-3694-0734

© Andrey N. Nevryuev

Candidate of Psychology, I.M. Sechenov First Moscow State Medical University, Moscow, Russia
a.n.nevryuev@mail.ru; ORCID: 0000-0002-8640-9717

This study was supported by the Russian Science Foundation (RSF) as part of research project No. 23-18-00422, <https://rscf.ru/project/23-18-00422/>

One of the biggest problems of modern society is information and social stress. They arise largely due to the value-affective polarization of people's views in social groups. Experts all over the world are working to solve the problem of polarization. The most well-known and widely discussed approach currently is the one proposed by Google DeepMind and called the Habermas Machine. The purpose of our study was to evaluate which statements and arguments about the benefits or harms of Western sanctions are more convincing for the Russian audience, and especially young people—well-known experts in the field of political and economic sciences or artificial intelligence? 168 respondents participated in the study. The results show that the polarization of opinions is manifested not in a uniform negative or positive assessment of expert judgments, but in a shift in emphasis. For politically “loyal” respondents, the initial positive interpretations are more important, while the “disloyal” ones prefer critical versions, especially when they are generated by AI.

Keywords: information stress, social stress, value-affective polarization of social groups, Habermas Machine

REFERENCES

1. Gulevich O.A., Kosimova S.S. (2024). The Connection between Russian Identity and Political Polarization: The Role of Secure National Identification and National Narcissism // *Social Psychology and Society*. № 15(4). P. 123–139. DOI: 10.17759/sps.2024150409
2. Tunkevichus E.O., Rebyazina V.A. (2023). A Multidimensional Model of Formation of Digital Trust of Users of Digital Services: Results of an Empirical Study // *Bulletin of Moscow University. Series 6. Economics*. № 4. P. 165–200. DOI:10.55959/MSU0130-0105-6-58-4-8
3. Harari Yu.N. (2019). *21 Lessons for the 21st Century*. M.: Sinbad.
4. Kudiabor H. (2024). AI tool helps people with opposing views find common ground // *Nature*. 17 October. DOI:10.1038/d41586-024-03424-z. URL: <https://www.nature.com/articles/d41586-024-03424-z>
5. Nelson R.M. U.S. (2017). Sanctions and Russia's economy (US sanctions and Russian economy) // Congressional research service. 02.17.2017. URL: <https://fas.org/sgp/crs/row/R43895.pdf>
6. Tessler M.H., Bakker M.A., Jarrett D., et al. (2024). AI can help humans find common ground in democratic deliberation // *Science*. V. 386. Issue 6719. URL: <https://www.science.org/doi/10.1126/science.adq2852>
7. Alexander Dudchak: Sanctions against Russia could dramatically improve our economy // *Ukraina.ru*. March 2, 2022. URL: <https://ukraina.ru/20220302/1033415109.html>
8. Famous economist Nassim Taleb explained how sanctions are beneficial for Russia // *Forbes. Business*. March 23, 2018. URL: <https://www.forbes.ru/biznes/359061-izvestnyy-ekonomist-nassim-taleb-rasskazal-chem-sankcii-polezny-rossii>
9. Kaya Kallas: We plan to adopt a new package of sanctions against Russia on May 20 // *Delovye Vedomosti*. May 12, 2025. URL: <https://www.dv.ee/novosti/2025/05/12/kaja-kallas-planiruem-prinjat-novyj-paket-sankcij-protiv-rf-20-maja>
10. 70% of Russian businesses have already suffered from sanctions // *Nezavisimaya Gazeta. Economics*. January 15, 2025. URL: https://www.ng.ru/economics/2025-01-15/1_9171_sanctions.html
11. Results of the survey “Consequences of the introduction of sanctions for Russian business” // *Russian Union of Industrialists and Entrepreneurs*. 2024. URL: <https://rspp.ru/activity/analytics/rezultaty-oprosa-posledstviya-vvedeniya-sanktsiy-dlya-rossiyskogo-biznesa/>
12. The Russian economy is currently eating up the future and spending beyond its means // *Novaya Gazeta*. August 11, 2024. URL: <https://novayagazeta.ru/articles/2024/11/08/rossiiskaia-ekonomika-proedaet-budushchee-i-tratit-ne-po-sredstvam>
13. “We won't have enough strength”: Kedmi revealed how Putin made Russia invulnerable to new US sanctions // *MPSH.RU*. May 9, 2025. URL: <https://mpsh.ru/31318-silenok-ne-hvatit-kedmi-raskryl-kak-putin-sdelal-rossiju-neujazvimoj-dlja-novyh-sankcij-ssha.html>
14. Economist Bodrunov: Anti-Russian sanctions contributed to Russia's development // *TASS*. May 6, 2025. URL: <https://tass.ru/ekonomika/23870687>
15. Economist Zubets: The main positive effect of the sanctions is a new development ideology based on domestic resources // *Rossiyskaya Gazeta*. March 10, 2023. URL: <https://rg.ru/2023/03/10/ekonomist-zubec-glavnyj-pozitivnyj-effekt-ot-sankcij-novaia-ideologiya-razvitiia-opiraiushchaisia-na-vnutrennie-resursy.html>