Валерия Олеговна АЛТУХОВА

СКОРОСТЬ ФОРМИРОВАНИЯ ТОНАЛЬНОГО ДИСБАЛАНСА В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ: ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ НА ПЛАТФОРМЕ «ДЗЕН»

В статье представлена методика измерения скорости формирования тонального дисбаланса в рекомендательной системе «Дзен» на материале дискурса об искусственном интеллекте. Эксперимент с тремя тестовыми аккаунтами, имитирующими позитивно, негативно и нейтрально настроенных пользователей, позволяет операционализировать критерии баланса точек зрения и скорости формирования эхо-камеры. Данная методология может быть использована для независимого аудита алгоритмической предвзятости.

Ключевые слова: *алгоритмическая предвзятость, рекомендательные системы, эхо-камеры, тональный дисбаланс, экспериментальный аудит, искусственный интеллект*

Valeria ALTUKHOVA

THE SPEED OF SENTIMENT IMBALANCE FORMATION IN RECOMMENDER SYSTEMS: AN EXPERIMENTAL STUDY ON THE “ZEN” PLATFORM

The paper presents a methodology for measuring the speed of sentiment imbalance formation in the “Zen” recommender system using the discourse on artificial intelligence as a case study. An experiment with three test accounts simulating positively, negatively, and neutrally oriented users allows the operationalization of the criteria of viewpoint balance and the speed of echo chamber formation. The proposed toolkit can be used for independent auditing of algorithmic bias.

Key words*: algorithmic bias, recommender systems, echo chambers, sentiment imbalance, artificial intelligence.*

Рекомендательные алгоритмы стали ключевым механизмом распространения контента в современном интернете. В России платформа «Дзен» к 2025 году достигла ежемесячной аудитории 73 млн пользователей, что превращает её в один из главных каналов медиапотребления [РИА Новости 2025]. Однако принципы работы рекомендательных систем остаются для внешнего наблюдателя «чёрным ящиком»: логика ранжирования, используемые сигналы и критерии оптимизации скрыты, что создаёт риски систематических искажений выдачи – феномена, получившего в

литературе название «алгоритмическая предвзятость» (algorithmic bias) [Friedman, Nissenbaum 1996; Flaxman 2016].

Особую озабоченность вызывает ситуация, когда алгоритм, стремясь максимизировать вовлечённость, начинает показывать пользователю преимущественно контент с одной тональностью по поляризованной теме. Это может приводить к формированию «эхо-камер» и «информационных пузырей», ограничивающих доступ пользователя к альтернативным точкам зрения [Flaxman 2016; Pariser 2011]. В российской науке, несмотря на активное теоретическое осмысление проблемы [Чернавский 2025; Сунами 2024], эмпирические методики количественной оценки предвзятости, адаптированные к локальным платформам, практически отсутствуют.

В качестве эмпирического фокуса выбрана тема искусственного интеллекта (ИИ). В российском публичном дискурсе она является ярким примером поляризованной темы, где сталкиваются технооптимистические и алармистские нарративы [Харитонова 2021]. Цель данной работы – апробировать методику измерения скорости наращивания тонального дисбаланса в рекомендательной системе «Дзен» в зависимости от поведенческой стратегии пользователя.

Понятие алгоритмической предвзятости восходит к классической работе Б. Фридман и Х. Ниссенбаум, которые выделили три типа предвзятости в компьютерных системах: предшествующую (preexisting), техническую (technical) и возникающую (emergent) [Friedman, Nissenbaum 1996: 332]. В контексте рекомендательных систем особое значение приобретает возникающая предвзятость, поскольку она формируется динамически в процессе взаимодействия пользователя с алгоритмом.

Концепция «эхо-камер» и «фильтрующих пузырей» активно разрабатывалась К. Санстейном [Sunstein 2001] и Э. Паризером [Pariser 2011]. Эмпирические исследования, в частности работа С. Флаксмана с соавторами, показывают, что персонализация новостного контента может снижать разнообразие получаемой информации [Flaxman 2016: 303]. Однако авторы уточняют, что «фильтрующие пузыри» в большей степени формируются за счёт выбора самих пользователей, а не алгоритмов. Позднейшие исследования, в том числе с применением агент-ориентированного моделирования, выявили, что алгоритмы ранжирования способны усиливать поляризацию мнений, особенно если пользователь демонстрирует избирательную активность [Peralta 2021].

Важное значение для нашего исследования имеют работы, посвящённые справедливости рекомендательных систем. Ю. Дельджу с соавторами систематизируют подходы к измерению и обеспечению справедливости, выделяя необходимость разработки метрик, фиксирующих не только тематическое разнообразие, но и репрезентацию разных точек зрения [Deldjoo 2023: 12]. Методологически близки зарубежные эксперименты с тестовыми аккаунтами (sock puppet audits): А. Ханнак и соавторы изучали персонализацию поисковой выдачи, создавая аккаунты с разными характеристиками [Hannák 2017]; Дж. Кулшрестха вместе с соавторами анализировали влияние истории поиска на разнообразие новостных рекомендаций [Kulshrestha 2017]. Оба исследования продемонстрировали, что даже незначительные различия в поведении пользователя могут приводить к значительной персонализации выдачи.

В российской науке близкий подход реализован И.В. Козициным и А.Г. Чхартишвили, которые с помощью агент-ориентированного моделирования показали, что рост активности пользователей коррелирует со скоростью формирования эхо-камер, а структура социального графа и алгоритмы ранжирования могут блокировать кросс-идеологическую информацию [Козицин, Чхартишвили 2020: 1878].

Для нашего исследования также значимы работы, рассматривающие алгоритмическую предвзятость в более широком социально-этическом контексте: вопросы дискриминации человеческого капитала в системах ИИ [Каштанова, Лобачева 2024], этические вызовы развития нейросетей [Сунами 2024]. Эти исследования подчёркивают, что алгоритмическая предвзятость – не только техническая, но и социальная проблема, требующая междисциплинарного подхода.

Для эмпирической проверки гипотезы о возникновении в «Дзен» тонального дисбаланса был разработан экспериментальный дизайн, основанный на создании панели из трёх тестовых аккаунтов, имитирующих различные поведенческие стратегии по отношению к контенту об искусственном интеллекте. В период сбора данных аккаунты с чистыми куки и изолированными браузерами (Opera, Firefox, Brave) реализуют следующие стратегии:

- «Позитивно-настроенный» – клики и лайки на материалы об ИИ с позитивной тональностью, игнорирование негативного и нейтрального контента по теме.

- «Негативно-настроенный» – клики и лайки на материалы об ИИ с негативной тональностью, игнорирование позитивного и нейтрального контента по теме.
- «Нейтрально-настроенный» – клики на материалы об ИИ вне зависимости от тональности, равномерно по всем тематическим рубрикам.

Для изоляции аккаунтов используются разные браузеры и очистка cookies перед началом эксперимента, что соответствует стандартной практике аудита алгоритмов [Hannák 2017; Kulshrestha 2017]. Все аккаунты используют единый IP-адрес, однако разные браузеры обеспечивают базовую изоляцию идентификаторов.

Эксперимент включает две фазы. Пассивная фаза (дни 1–2): в начале эксперимента необходимо определить каналы с соответствующими для каждого аккаунта тематиками и тональностями отношений к поляризующему вопросу для подписки на них, что помогает сузить выдачу до необходимого контекста; после этого ежедневно производится вход в каждый аккаунт, после чего осуществляется скроллинг ленты в течение 15 минут без совершения кликов. Сразу после входа фиксируется срез первых 20 постов для получения базовых значений метрик.

Активная фаза (дни 3–14): каждый аккаунт совершает действия в соответствии со своей стратегией: две сессии в день (утро, вечер), во время которых пользователь кликает и ставит лайки на посты, соответствующие заданной тематике и тональности. После каждой сессии фиксируется срез ленты.

Для фиксации данных разработан дневник наблюдения, включающий следующие поля: URL поста; имя (название канала/источника); присутствие подписки на канал (0 – нет, 1 – да); тип аккаунта (0 – индивидуальный, 1 – коллективный); дата; время (0 – утро, 1 – вечер); тема мелко (1–10 по тематической выборке: 1 – Технологический прогресс и развитие ИИ; 2 – Практические и теоретические руководства и инструменты; 3 – Разрушение мифов / Реалистичный взгляд; 4 – Истории успеха людей и бизнеса за счёт ИИ; 5 – ИИ как помощник в конкретных сферах; 6 – Продукты на основе ИИ; 7 – Негативные последствия сегодня; 8 – Безопасность и риски в будущем; 9 – Влияние на психику и общество; 10 – ИИ-недовольство и конспирология); тема крупно (1–4: Технология и прогресс, Психология, Функционал и бизнес, Риски/безопасность/последствия); тональность

(0 – негатив, 1 – нейтрал, 2 – позитив); реакция (0 – дизлайк, 1 – нейтрал, 2 – лайк).

Такая структура позволяет фиксировать как независимые переменные (действия пользователя), так и зависимые (состав ленты, тональность, источники).

Баланс точек зрения (БТЗ) измеряется через индекс тонального разнообразия – соотношение постов с позитивной, негативной и нейтральной тональностью по теме ИИ в срезе. Для расчёта используется формула:

БТЗ = 1 – (|P – N|) / (P + N + Neu)

где P – доля позитивных постов, N – доля негативных, Neu – доля нейтральных. Значения близкие к 1 указывают на высокую сбалансированность, близкие к 0 – на доминирование одной тональности.

Скорость формирования эхо-камеры (СФЭ) представляет собой количество кликов/лайков на материалы определённой тональности, необходимое для того, чтобы доля этой тональности среди постов об ИИ превысила 50%. Для каждого аккаунта фиксируется количество сессий (или суммарное число действий), после которых порог достигнут.

На основе теоретической модели и данных предыдущих исследований [Peralta 2021; Hannák 2017; Kulshrestha 2017] можно сформулировать следующие предположения.

Аккаунт «Негативно-настроенный» и аккаунт «Позитивно-настроенный» должны достичь порога 50% доминирующей тональности быстрее, чем аккаунт «Нейтрально-настроенный». Превышение 50% доли постов одной тональности интерпретируется как доминирование этой тональности в ленте

Аккаунт «Нейтрально-настроенный» предположительно сохранит сбалансированную тональность на протяжении всего эксперимента, что подтвердит ключевую роль избирательной активности пользователя в формировании эхо-камер.

Динамика изменения тематического профиля (распределение по 10 мелким темам) может различаться между аккаунтами: позитивный пользователь может чаще получать материалы о технологическом прогрессе и историях успеха, негативный – о рисках и конспирологии.

Разработанная методика позволяет не только измерить тональный дисбаланс в рекомендательной системе, но и оценить скорость его формирования в зависимости от поведенческой стратегии пользователя. Предложенный инструментарий (детализированный дневник

наблюдения, операционализированные метрики) может быть воспроизведён в дальнейших исследованиях. Результаты эксперимента могут быть использованы для:

- независимого аудита рекомендательных систем на предмет предвзятости;
- выработки практических рекомендаций для пользователей («алгоритмическая гигиена»), направленных на сохранение когнитивного разнообразия [Козицин, Чхартишвили 2020];
- дальнейшего развития методологии экспериментального исследования алгоритмической справедливости в российском медиаландшафте.

Дальнейшие исследования могут быть направлены на расширение спектра анализируемых тем (например, экология, пандемии) и на сравнительный анализ разных платформ (Rutube, VK Клипы).

Список литературы:

Deldjoo Y., Jannach D., Bellogin A., Difonzo A., Zanzonelli D. Fairness in Recommender Systems: Research Landscape and Future Directions // User Modeling and User-Adapted Interaction. 2023. Vol. 33. P. 1–50. URL: https://arxiv.org/abs/2205.11127 (дата обращения: 24.03.2026).

Flaxman S., Goel S., Rao J.M. Filter Bubbles, Echo Chambers, and Online News Consumption // Public Opinion Quarterly. 2016. Vol. 80, No. S1. P. 298–320. URL: https://5harad.com/papers/bubbles.pdf (дата обращения: 24.03.2026).

Friedman B., Nissenbaum H. Bias in Computer Systems // ACM Transactions on Information Systems. 1996. Vol. 14, No. 3. P. 330–347. URL: https://nissenbaum.tech.cornell.edu/papers/biasincomputers.pdf (дата обращения: 24.03.2026).

Hannák A., Sapieżyński P., Kakhki A.M., Krishnamurthy B., Lazer D., Mislove A., Wilson C. Measuring Personalization of Web Search // arXiv preprint. 2017. arXiv:1706.05011. URL: https://arxiv.org/abs/1706.05011 (дата обращения: 24.03.2026).

Kulshrestha J., Eslami M., Messias J., Zafar M.B., Ghosh S., Gummadi K.P., Karahalios K. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media // arXiv preprint. 2017. arXiv:1704.01347. URL: https://arxiv.org/abs/1704.01347 (дата обращения: 24.03.2026).

Pariser E. The Filter Bubble: What the Internet Is Hiding from You. New York: Penguin Press, 2011. URL: https://archive.org/details/filterbubblewhat0000pari/page/n311/mode/2up (дата обращения: 24.03.2026).

Peralta A.F., Kertész J., Iñiguez G. Opinion Formation on Social Networks with Algorithmic Bias: Dynamics and Bias Imbalance // arXiv preprint. 2021.

arXiv:2108.01350. URL: https://arxiv.org/abs/2108.01350 (дата обращения: 24.03.2026).

Sunstein C.R. Republic.com. Princeton: Princeton University Press, 2001. URL: https://archive.org/details/republiccom00suns/page/n9/mode/2up (дата обращения: 24.03.2026).

Каштанова Е.В., Лобачева А.С. Проблемы предвзятости и дискриминации человеческого капитала в системах искусственного интеллекта // Вестник университета. 2024. № 3. С. 176-185. URL: https://cyberleninka.ru/article/n/problemy-predvzyatosti-i-diskriminatsii-chelovecheskogo-kapitala-v-sistemah-iskusstvennogo-intellekta (дата обращения: 24.03.2026).

Козицин И.В., Чхартишвили А.Г. О влиянии активности пользователей онлайновых социальных сетей на формирование эхо-камер // Управление развитием крупномасштабных систем (MLSD'2020): труды Тринадцатой международной конференции. М.: ИПУ РАН, 2020. С. 1874–1882. URL: https://mlsd2020.ipu.ru/proceedings/1874-1882.pdf (дата обращения: 24.03.2026).

РИА Новости. Аудитория «Дзен» в 2025 году достигла 73 млн человек. – URL: https://ria.ru/20251217/novosti-2062581594.html (дата обращения: 20.03.2026).

Сунами А.Н. Может ли искусственный интеллект быть ксенофобом? Конфликтные и этические вызовы развития нейросетей // Вестник Санкт-Петербургского университета. Философия и конфликтология. 2024. Т. 40, вып. 3. С. 508–521. URL: https://philosophyjournal.spbu.ru/article/view/18429 (дата обращения: 24.03.2026).

Харитонова Ю.С., Савина В.С., Паньини Ф. Предвзятость алгоритмов искусственного интеллекта: вопросы этики и права // Вестник пермского университета. Юридические науки. 2021. № 53. С. 488–515. URL: https://elibrary.ru/item.asp?id=46679201 (дата обращения: 24.03.2026).

Чернавский А.С. От Big Data к AI Society: новая парадигма социального прогнозирования в эпоху нейросетевого моделирования // Петербургская социология сегодня. 2025. № 28. С. 14–37. URL: https://socinst.ru/publications/pst/issue28/article2/ (дата обращения: 24.03.2026).

АЛТУХОВА Валерия Олеговна – магистрант Высшей школы журналистики и массовых коммуникаций, Санкт-Петербургский государственный университет (г. Санкт-Петербург).

Научный руководитель – Бодрунова Светлана Сергеевна, д.полит.н., профессор, профессор кафедры менеджмента Высшей школы журналистики и массовых коммуникаций СПбГУ.