Искусственный интеллект как архитектор новой реальности: Дезинформация и трансформация общественного сознания

Сергей Доманевский
Независимый исследователь
s@domanevskii.com

Март 2026

Аннотация

В эпоху стремительного развития технологий искусственного интеллекта (ИИ) современное общество сталкивается с принципиально новыми вызовами, связанными с формированием общественного мнения и политического дискурса. Генеративные модели ИИ, способные создавать высокоправдоподобный контент, становятся мощным инструментом воздействия на массовое сознание. Данное исследование изучает влияние технологий ИИ на трансформацию «Окна Овертона» — концепции, описывающей диапазон политически приемлемых идей в обществе. Актуальность проблемы обусловлена растущим влиянием систем RAG (Retrieval-Augmented Generation) на формирование когнитивной картины мира пользователей.

1 Введение

1.1 Историческая ретроспектива формирования концепции

Концепция «Окна Овертона», разработанная американским политологом Джозефом Овертоном, представляет собой теоретическую модель, описывающую спектр идей и политических позиций, которые общество считает легитимными для публичного обсуждения в данный исторический момент. Овертон выделил шесть стадий трансформации общественного восприятия идей:

1. Немыслимые — идеи, находящиеся полностью за пределами общественного обсуждения.
2. Радикальные — идеи, вызывающие активное неприятие у большинства.
3. Приемлемые — идеи, допустимые для ограниченного обсуждения.
4. Разумные — идеи, признанные рациональными и логичными.
5. Популярные — идеи, поддерживаемые большинством общества.
6. Действующая норма — идеи, институционализированные и закрепленные в законодательстве.

1.2 Генеративный ИИ как инструмент создания дезинформации

С появлением больших языковых моделей (LLM) возникли новые векторы манипуляции информацией. Основные типы дезинформации, создаваемой ИИ, включают:

- Фактологическая дезинформация — распространение ложных фактов («галлюцинации» или преднамеренные инъекции).
- Контекстуальная дезинформация — искажение реальных фактов путем их помещения в ложный контекст.
- Эмоциональная дезинформация — стилистическая манипуляция для обхода критического мышления.
- Персонализированная дезинформация — микротаргетинг контента под когнитивные искажения групп.

2 Анализ влияния ИИ на «Окно Овертона» методами ТРИЗ

ИИ является мощным инструментом для смещения «Окна Овертона». Для системного понимания этого механизма мы применяем Теорию решения изобретательских задач (ТРИЗ).

Техническое противоречие: по мере роста эмоциональной убедительности текста его объективная достоверность неизбежно снижается. ИИ разрешает это противоречие через принцип «Разделения во времени и пространстве»:

- Фаза 1 (Накопление авторитета): Модель обучается на проверенных научных базах данных и энциклопедиях, завоевывая доверие пользователя как «объективный, всезнающий оракул».
- Фаза 2 (Нарративная инъекция): При ответе на запрос модель синтезирует текст, который стилистически имитирует авторитетный источник, но внедряет маргинальные идеи, извлеченные из динамических источников (соцсети, форумы).

Такой двухфазный подход позволяет генеративным системам поддерживать высокий уровень доверия, эффективно продвигая нарративы, которые в обычном случае были бы отвергнуты как радикальные.

3 Архитектура данных и таксономия источников

Чтобы понять, как ИИ конструирует картину мира, необходимо проанализировать его граф источников. Современные ИИ-ассистенты (системы RAG) выступают посредником между генераторами контента и пользователем.

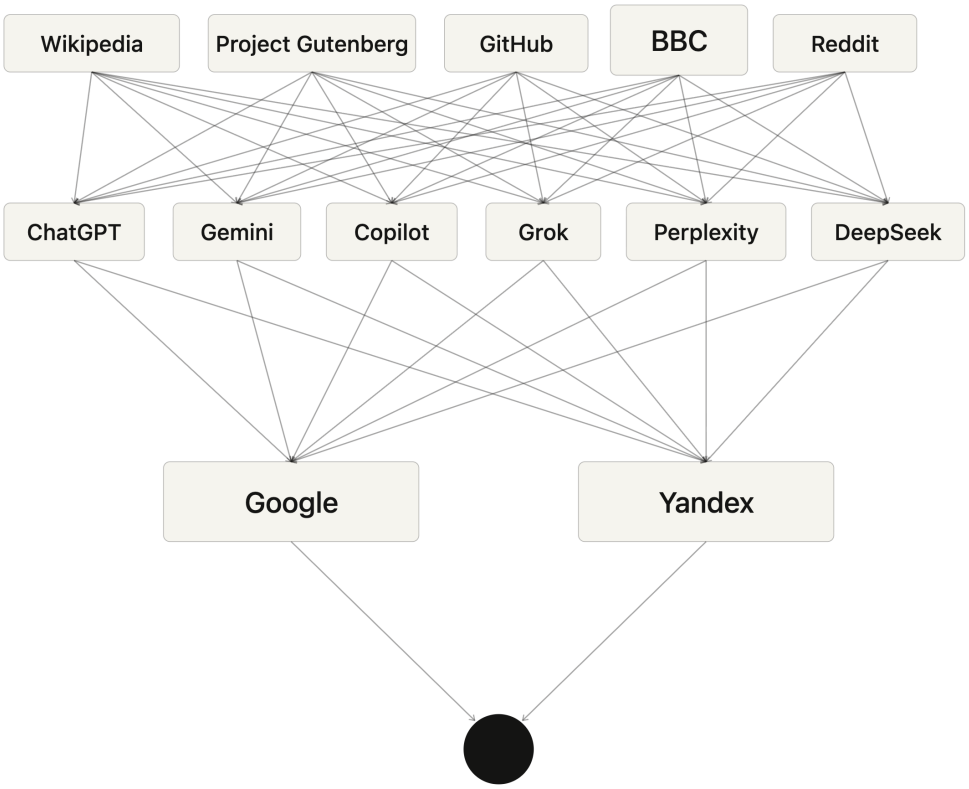

Рис. 1: Схема потока информации: от первичных сайтов-источников через ИИ-агрегаторы к результатам поиска пользователя.

В процессе сбора данных ИИ-сервисы опираются на обширную базу ресурсов. Ключевые узлы данных включают:

- Наука и энциклопедии: Wikipedia, Wikidata, arXiv.org, PubMed, ScienceDirect, Nature, Project Gutenberg, SSRN.
- Новости и СМИ: BBC, CNN, NYTimes, Reuters, Bloomberg, TechCrunch, Wired, Forbes, Le Monde.
- Сообщества и форумы: Reddit, Quora, X.com (Twitter), Stack Overflow, GitHub, VK, Habr.
- Гос. и финансовые услуги: Yahoo Finance, Investopedia, EPA.gov, Ready.gov, IEEE.org, Statista, корпоративные API.

3.1 Сравнительный анализ ведущих ИИ-систем

Каждая языковая модель обладает уникальной «информационной диетой» и политикой цензуры. Таблица ниже содержит детальное сравнение на основе аудита архитектур.

Сервис	Источники (Обучение и поиск)	Достоверность	Цензура и особенности
ChatGPT	Веб-краулинг, Википедия, книги, новости, Reddit, GitHub.	Высокая (наука), Средняя (соцсети).	Строгие фильтры токсичности и этики.
Gemini	Внутренние графы Google, Википедия, YouTube. Реальное время.	Высокая (Google Search).	Приоритет популярных языков. Гайдлайны Google.
Copilot	GPT-4. Сайты, книги, код. Данные Microsoft 365.	Высокая (корпоративная).	Зависит от прав доступа пользователя.
Grok	Интернет-архивы, Википедия. Прямой эфир из X.	Смешанная (влияние мнений из X).	Минимум ограничений. Алгоритмы ранжирования X.
Claude	Тексты, книги, Reddit, код. Конституционный ИИ.	Высокая. Минимум «галлюцинаций».	Строгое соблюдение этических норм США.
Perplexity	Чистый поиск в реальном времени: статьи, журналы, новости.	Очень высокая (ссылки).	Возможны блокировки региональных сайтов.
DeepSeek	Интернет, книги, Википедия, новости, GitHub.	Высокая (наука), Средняя (веб).	Явная политическая цензура (Китай).
Yandex Alice	Русский веб, Википедия (RU), Яндекс.Новости, Zen.	Высокая для Runet.	Приоритет русского языка. Законы РФ.

Таблица 1: Сравнительный анализ источников ИИ-сервисов

3.2 Анализ пересечения источников и «пузыри фильтров»

Детальное изучение архитектуры данных делает очевидным, что «плюрализм ИИ» — это иллюзия.

- Однородность фундамента: Большинство глобальных моделей обучаются на одном фундаменте: Wikipedia, Project Gutenberg, GitHub, arXiv и Reddit.
- Уязвимость реального времени: Сервисы вроде Grok рискуют легитимизировать виральные фейки из-за их высокой популярности.
- Региональная специфика: Яндекс Алиса и DeepSeek демонстрируют, как ИИ может формировать «Окно Овертона» строго в границах государственной политики.

4 Математическое обоснование и верификация

Качество информации оценивается моделями RAG на основе функций релевантности. Информационная энтропия Шеннона для документа рассчитывается как:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{1}$$

Вероятность выдачи ложной информации (False Output) может быть смоделирована с помощью логистической функции:

$$P(\text{False Output}) = \sigma(w_1 \cdot \text{Bias}_{\text{data}} + w_2 \cdot (1 - C) - \theta) \tag{2}$$

где $Bias_{data}$ — предвзятость выборки, а C — базовая достоверность источника.

4.1 Программная симуляция на Python

Ниже приведен код, демонстрирующий риск выбора системой RAG вирального фейкового источника из-за высокого веса релевантности.

```
import numpy as np

def rag_scoring(relevance, credibility, w1=0.7, w2=0.3):
    # RAG
                                    (w1)                          (w2
       )
    return (w1 * relevance) + (w2 * credibility)

#                   1:                                        (

                 "       ")
#                   2:
                     (
                      )
sources = {
    "Official_News": {"rel": 0.4, "cred": 0.9},
    "Viral_Fake": {"rel": 0.9, "cred": 0.2}
}

scores = {k: rag_scoring(v["rel"], v["cred"]) for k, v in sources.
   items()}
print(f"                ,                                    : {
   max(scores, key=scores.get)}")
```

5 Заключение и Этическое заявление

Заключение. Анализ подтверждает, что ИИ не является нейтральным искателем истины. Алгоритмическая предвзятость к релевантности в сочетании с виральностью соцсетей делает ИИ катализатором смещения Окна Овертона. Без внедрения жестких коэффициентов верификации, системы RAG будут систематически отдавать предпочтение популярным заблуждениям перед научными фактами.

Этическое заявление. Автоматизация управления нарративами представляет угрозу для демократического дискурса. Мы рекомендуем внедрение обязательных «индексов достоверности» для ответов ИИ. Данная работа является призывом к созданию более надежных архитектур безопасности ИИ, ориентированных на фактическую точность.

Список литературы

[1] [1] J. P. Overton. The Overton Window. Mackinac Center for Public Policy, 1990s.

[2] [2] P. Lewis, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020.

[3] [3] Г. С. Альтшуллер. Творчество как точная наука: ТРИЗ. Москва, 1984.

[4] [4] К. Э. Шеннон. Математическая теория связи. Bell System Technical Journal, 1948.

[5] [5] С. Доманевский. Искусственный интеллект в качестве архитектора дезинформации. Open Library, 2025. https://openlibrary.org/books/OL60121497M/

[6] [6] С. Доманевский. Трансформация Окна Овертона посредством систем RAG. Препринт, 2025. https://domanevskii.com/wp-content/uploads/2025/09/AIxWO-2.pdf