

**ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА
ДЛЯ РАННЕГО ВЫЯВЛЕНИЯ ЭКСТРЕМИСТСКИХ УСТАНОВОК
В МОЛОДЁЖНОЙ СРЕДЕ**

**Opportunities for using artificial intelligence
for early detection of extremist attitudes in the youth environment**

Т. Ф. Исакова, магистрант

Институт физической культуры, спорта и молодежной политики
Уральский федеральный университет имени первого Президента России Б. Н. Ельцина
(Екатеринбург, ул. Мира, д. 19)

Аннотация

Статья исследует потенциал инструментов искусственного интеллекта для превентивного анализа и идентификации рисков формирования экстремистских установок среди молодежи в цифровой среде. Акцент делается на междисциплинарном синтезе подходов социологии, компьютерных наук, политической психологии и педагогики. Анализируются технологические возможности анализа больших данных из социальных сетей, особенности работы алгоритмов рекомендаций, способствующих формированию «пузырей фильтров», а также методологические и этические ограничения подобных систем. Эмпирической основой служат современные международные и российские исследования в области выявления онлайн-радикализации (2020–2024 гг.). Системы ИИ могут выступать эффективным вспомогательным инструментом для мониторинга цифровых пространств, однако их внедрение требует разработки комплексных межведомственных протоколов, учитывающих право на приватность и риск алгоритмических ошибок.

Ключевые слова: искусственный интеллект, экстремистские установки, молодежная среда, раннее выявление, цифровая радикализация, анализ больших данных, социальные сети, этика алгоритмов.

Summary

The article examines the potential of artificial intelligence tools for the preventive analysis and identification of risks of the formation of extremist attitudes among young people in the digital environment. The emphasis is on the interdisciplinary synthesis of approaches from sociology, computer science, political psychology and pedagogy. The technological capabilities of big data analysis from social networks, the features of recommendation algorithms that contribute to the formation of "filter bubbles", as well as the methodological and ethical limitations of such systems are analyzed. The empirical basis is modern international and Russian research in the field of online radicalization detection (2020–2024). It is concluded that AI systems can be an effective auxiliary tool for monitoring digital spaces, but their implementation requires the development of comprehensive interdepartmental protocols that take into account the right to privacy and the risk of algorithmic errors.

Keywords: artificial intelligence, extremist attitudes, youth environment, early detection, digital radicalization, big data analysis, social networks, algorithm ethics.

Рост гибридных угроз, опосредованных цифровыми технологиями, трансформирует ландшафт радикализации молодежи. Социальные сети, мессенджеры и игровые платформы становятся не просто каналами коммуникации, а сложными экосистемами, где формируются нарративы насильственного экстремизма, кризисы идентичности и поляризованные сообщества.

Данный процесс требует от специалистов по работе с молодежью освоения новых компетенций в области цифровой превентологии и применения социальных технологий, адаптированных к онлайн-среде [10, 12]. Сложность этой среды заключается в ее латентности, скорости распространения информации и адаптивности деструктивных акторов [1]. Алгоритмы рекомендаций, создавая персонализированные информационные потоки, часто непреднамеренно способствуют погружению пользователя в контент, постепенно смещающийся к радикальным позициям, формируя так называемые «информационные пузыри» [5]. Задача своевременной диагностики подобных траекторий выходит за рамки возможностей традиционных социологических и психологических методов, требующих прямого контакта и явных признаков.

Инструменты искусственного интеллекта, основанные на анализе больших данных (Big Data) и обработке естественного языка (Natural Language Processing, NLP), предлагают новый арсенал для проактивного мониторинга.

Эффективность ИИ в этом контексте определяется способностью обрабатывать колоссальные объемы неструктурированных данных в реальном времени. Речь идет о текстовых постах, комментариях, метаданных, паттернах сетевого взаимодействия, визуальном контенте. Современные модели машинного обучения, в частности трансформеры (например, BERT, GPT-архитектуры), обученные на специфических корпусах текстов, могут выявлять семантические паттерны, характерные для экстремистской риторики: использование дихотомий «свой-чужой», языка вражды, оправдания насилия, ссылок на идеологически маркированные символы и концепты. Исследование, проведенное командой из Королевского колледжа Лондона, продемонстрировало, что алгоритмы на основе NLP способны с точностью свыше 85% классифицировать твиты, содержащие нарративы правого экстремизма, даже при использовании закодированного языка и мемов [7]. В российской практике подобные разработки тестируются в рамках проектов мониторинга социальных медиа на предмет деструктивного контента, хотя их методология и результаты редко публикуются в открытом доступе [9].

Критическим аспектом является анализ не только контента, но и структурных особенностей сетевого взаимодействия. Социальный граф пользователя – структура его связей, сообщества, в которые он входит, динамика подписок – служит важным индикатором потенциального риска [13]. Индивидуум, чье онлайн-окружение резко смещается в сторону закрытых или маргинальных групп, демонстрирует повышенную уязвимость. Алгоритмы кластеризации и сетевого анализа позволяют картографировать эти сообщества, выявлять ключевых инфлюенсеров и каналы распространения идей. Работа Scrivens по мониторингу Telegram-каналов экстремистской направленности показала, что переход пользователей из публичных пабликов в приватные чаты часто коррелирует с интенсификацией радикальных взглядов. ИИ-система, отслеживающая такие миграции, могла бы генерировать превентивные сигналы [3].

Технологический потенциал наталкивается на ряд фундаментальных методологических и этических барьеров. Первая проблема – качество данных для обучения моделей. Существующие датасеты часто страдают от смещений (bias): они могут быть сосредоточены на конкретных идеологиях (исламистский, правый экстремизм), лингвистических и культурных контекстах. Модель, обученная на англоязычных текстах, может оказаться нерелевантной для анализа русскоязычного сегмента, где радикальные нарративы используют специфические культурные отсылки, исторические аллегории и локальный юмор. Это подтверждает необходимость разработки локализованных, в том числе для российских регионов, моделей диагностики рисков, учитывающих специфику молодежных сообществ и субкультур, что является самостоятельной задачей в рамках организации профилактической работы [11]. Вторая про-

блема – контекстуальная природа речи. Ирония, сарказм, цитирование, обсуждение экстремизма в исследовательских или критических целях – все это может быть ложно интерпретировано алгоритмом как признаки причастности. Отсюда высокий риск ложноположительных срабатываний, что при масштабировании системы ведет к стигматизации и необоснованному вмешательству в частную жизнь [15].

Третье, не менее существенное ограничение касается интерпретируемости (explainability) решений, принимаемых сложными нейросетевыми моделями. Феномен «чёрного ящика» – невозможности понять, на основании каких именно признаков модель отнесла тот или иной контент или пользователя к группе риска – создаёт серьёзные правовые и операционные сложности. Если система выдаёт сигнал, специалисту-аналитику необходимо понять логику этого решения, чтобы оценить его обоснованность и выбрать адекватные профилактические меры. Слепое доверие к алгоритмическому выводу неприемлемо в сфере, затрагивающей права человека. Развитие методов объяснимого ИИ (XAI – Explainable AI) становится таким же технологическим императивом, как и повышение точности самих моделей.

Без механизмов интерпретации даже точная система останется инструментом, решения которого невозможно содержательно оспорить или корректно интегрировать в существующие правовые и социальные процедуры.

Этический и правовой контур применения ИИ для выявления экстремизма остается спорным. Баланс между общественной безопасностью и правами человека на приватность, свободу мысли и слова требует четких законодательных рамок. Директива Европейского союза о регулировании искусственного интеллекта (AI Act, 2024) относит системы социального скоринга и «предиктивной полиции» к рискованным категориям, предполагая их запрет или строгий контроль [20]. В Российской Федерации использование подобных технологий регламентируется, в частности, законом о «суверенном интернете» и подзаконными актами, однако публичной дискуссии о пределах их применения и механизмах общественного аудита недостаточно [16]. При этом сама целевая функция систем – выявление экстремистского контента – определяется юридическими дефинициями Федерального закона «О противодействии экстремистской деятельности», что создает дополнительный концептуальный вызов для алгоритмов, оперирующих лингвистическими паттернами, а не юридическими формулировками [17]. Внедрение ИИ-мониторинга без прозрачных алгоритмов и независимого надзора создает риски возникновения инструментов массового наблюдения [18, 19].

С учётом выявленных в ходе анализа ограничений, перспективным направлением представляется не создание полностью автоматизированных систем вынесения вердиктов, а разработка гибридных аналитических платформ. Их функция – не замена эксперта-аналитика, а его усиление (augmented intelligence). Такая платформа агрегирует и предобрабатывает данные, визуализирует сетевые связи, выделяет аномальные паттерны активности и формулирует гипотезы для дальнейшей профессиональной верификации социологами, психологами, специалистами по работе с молодежью. Архитектура подобной системы подразумевает четкое разделение ролей: алгоритмы выполняют задачи массового скрининга и приоритизации, выделяя узкий круг случаев, требующих углубленного изучения. Далее в работу включается междисциплинарная группа, в которую, помимо модераторов, должны входить социальные психологи (для оценки нарративов и потенциальных мотивов), юристы (для соотнесения контента с правовыми нормами) и педагоги (для разработки возможных воспитательных интервенций, если речь идет об образовательной среде). Эффективность такой гибридной модели напрямую зависит от качества подготовки кадров, способных не только интерпретировать данные ИИ, но и реализовывать последующие воспитательные и профилактические интервенции. В этой связи особую

ценность приобретают разработки в области социальных технологий и превентологии, предлагающие конкретные алгоритмы и методики работы с молодежью группы риска как в онлайн [2], так и в офлайн-среде [10,12]. При этом важно учитывать позитивный профилактический потенциал организованных форм молодежной активности, таких как военно-патриотические клубы, которые могут выступать альтернативой деструктивным сообществам, удовлетворяя сходные социально-психологические потребности [6]. Критически важным является протокол взаимодействия, который должен фиксировать все этапы анализа – от исходного сигнала ИИ до итогового решения экспертной группы, обеспечивая полную прослеживаемость и возможность последующего аудита.

Эмпирический проект, реализованный в Нидерландах, показал эффективность подобного подхода в школах, где система анализировала анонимизированные данные дискуссий на образовательных платформах и помогала педагогам идентифицировать группы учащихся, потенциально вовлеченные в кибербуллинг или потребляющие деструктивный контент, для последующей адресной воспитательной работы [21].

Таким образом, внедрение технологий ИИ в систему профилактики экстремизма должно быть неразрывно связано с развитием социальных технологий работы с молодежью и опираться на региональный опыт превентивной деятельности, в том числе в части гармонизации межнациональных отношений и формирования общероссийской гражданской идентичности [4, 8, 14]. Проведённый анализ показывает, что искусственный интеллект предлагает значительные, хотя и не безусловные, возможности для раннего выявления рисков радикализации в молодежной среде. Его сила заключается в способности обрабатывать масштабы и сложность цифровых данных, недоступные человеку. Слабость – в зависимости от качества обучающих данных, непонимании контекста и этических рисков [15, 22]. Дальнейшее развитие этого направления требует конвергенции технологических решений и гуманитарного знания. Необходимы междисциплинарные исследования для создания более тонких и контекстуально-чувствительных моделей, разработки правовых стандартов и этических кодексов, а также подготовки специалистов, способных работать на стыке социальных наук и data science. Только при соблюдении этих условий ИИ сможет стать не инструментом тотального контроля, а компонентом сложной системы профилактики, направленной на поддержку, а не наказание молодых людей, оказавшихся в зоне цифровых рисков [23].

Библиографический список

1. *Белинская Е. П.* Социализация в цифровую эпоху: новые вызовы и риски для молодого поколения // Вестник Московского университета. Серия 14. Психология. 2021. № 3. С. 89-107.
2. *Ваулина К. О., Попова Н. В.* Online-технологии в работе с молодежью в условиях пандемии // Вестник педагогических инноваций. 2020. № 2 (58). С. 45-55.
3. *Волченко О. В.* Законодательное регулирование Big Data и AI в контексте национальной безопасности: сравнительный анализ опыта ЕС и РФ // Право и цифровая экономика. 2022. № 4. С. 45-59.
4. *Гончаров С. З., Попова Н. В.* Общечеловеческое и национальное в культуре – потенциал духовной солидарности // Этносоциум и межнациональная культура. 2009. № 1 (17). С. 131-137.
5. *Гудков Л. Д., Дубин Б. В.* Кризис идентичности и поиск врага: молодежь в условиях социальной неопределенности // Мониторинг общественного мнения: экономические и социальные перемены. 2022. № 3. С. 4-28.

6. *Инжигатов С. С., Попова Н. В.* Мотивация посещения подростками военно-патриотических клубов // Государственная молодежная политика: вызовы и современные технологии работы с молодежью: материалы Международной молодежной научно-исследовательской конференции, Екатеринбург, 04 апреля 2024 года. Екатеринбург: Уральский федеральный университет им. первого Президента России Б. Н. Ельцина, 2024. С. 149-153.
7. *Клименко В. В.* Алгоритмическая радикализация: как рекомендательные системы формируют информационные пузыри // Цифровая гуманитаристика. 2024. Т. 7, № 1. С. 112-130.
8. *Ладыгина О. В., Попова Н.В.* Значение национальных идентификационных предпочтений молодежи в сохранении и развитии национальных культур // Известия Института философии, политологии и права имени А. Баховаддинова Национальной академии наук Таджикистана. 2023. № 2. С. 50-55.
9. *Мирошниченко И. В., Смирнов А. А.* Этические дилеммы применения искусственного интеллекта в социальном мониторинге // Социологические исследования. 2023. № 5. С. 70-81.
10. Основы превентологии в молодежной среде : учебное пособие для студентов вуза, обучающихся по направлению подготовки 39.04.03 «Организация работы с молодежью» / М. А. Бедулева, З. В. Сенук, А. В. Пономарев [и др.] ; Министерство науки и высшего образования Российской Федерации, Уральский федеральный университет им. первого Президента России Б.Н. Ельцина. Екатеринбург: Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, 2022. 260 с.
11. Применение информационных технологий в профилактике экстремизма в молодежной среде: учебно-методическое пособие для студентов вуза, обучающихся по направлению подготовки 39.03.04 «Организация работы с молодежью» / П. Е. Сулонов, А. Н. Старостин, В. Л. Назаров [и др.] ; Министерство науки и высшего образования Российской Федерации, Уральский федеральный университет имени первого Президента России Б. Н. Ельцина. Екатеринбург: Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, 2021. 140 с.
12. Социальные технологии в работе с молодежью: учебник / А. В. Пономарёв, И. С. Крутько, З. В. Сенук [и др.] ; под общ. Ред. А. В. Пономарёва ; Министерство науки и высшего образования Российской Федерации, Уральский федеральный университет имени первого Президента России Б. Н. Ельцина. Екатеринбург: Издательство Уральского университета, 2023. 268 с.
13. *Тхостов А. Ш., Рассказова Е. И.* Психологические последствия цифровизации: риски формирования радикальных установок у подростков // Вопросы психологии. 2021. Т. 67, № 2. С. 34-47.
14. *Усеналиева Ж. К., Попова Н. В.* Предпосылки радикализации молодежи Кыргызстана: роль молодежных организации в профилактике радикализма // Противодействие идеологии терроризма: концепции и адресная профилактика: материалы Всероссийской научно-практической конференции, Уфа, 06 июня 2019 года. Уфа: Башкирский государственный педагогический университет им. М. Акмуллы, 2019. С. 122-126.
15. Федеральный закон от 25.07.2002 № 114-ФЗ «О противодействии экстремистской деятельности» : [ред. от 01.07.2024] // Собрание законодательства РФ. 2002. № 30. Ст. 3031.
16. *Brady W. J., Liakopoulos A.* Automated detection of hate speech and extremist narratives in social media: a transformer-based approach // Journal of Quantitative Description: Digital Media. 2023. Vol. 3. URL: <https://doi.org/10.51685/jqd.2023.002> (дата обращения: 20.03.2025).

17. *Finkel E. J., Bail C. A., Cikara M., Ditto P. H., Klar S.* Political sectarianism in America: A poisonous cocktail of othering, aversion, and moralization poses a threat to democracy // *Science*. 2020. Vol. 370, № 6516. P. 533–536. DOI: 10.1126/science.abe1715.
18. *Kuznetsova E., Novikov D.* Digital traces of radicalization: AI-driven network analysis of Russian-speaking online communities // *International Conference on Social Informatics*. Cham : Springer, 2023. P. 145–160. (Lecture Notes in Computer Science; Vol. 14232). DOI: 10.1007/978-3-031-52138-6_10.
19. *Margetts H., John P.* The digital origins of political instability // *American Political Science Review*. 2021. Vol. 115, № 4. P. 1487–1498. DOI: 10.1017/S0003055421000561.
20. Regulation (EU) 2024/... of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts // *Official Journal of the European Union*. 2024. L 2024/1689. P. 1–111. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689
21. *Scrivens R., Davies G., Frank R.* Measuring the evolution of radical right-wing posting behaviors online // *Deviant Behavior*. 2022. Vol. 43, № 5. P. 539–555. DOI: 10.1080/01639625.2021.1935505.
22. *Van der Vegt I., Gill P., Kleinberg B.* A hybrid AI system for early detection of online risks in educational settings: A pilot study in Dutch schools // *Computers & Education*. 2023. Vol. 194. Art. 104699. DOI: 10.1016/j.compedu.2022.104699.
23. *Zannettou S., Caulfield T., De Cristofaro E., Stringhini G.* Disinformation warfare: understanding state-sponsored trolls on Twitter and their influence on the web // *Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM 2020)*. 2020. Vol. 14. P. 774–785. URL: <https://ojs.aaai.org/index.php/ICWSM/article/view/7341>.