Иван Александрович Смекалин

стажер-исследователь Лаборатории
социальной и когнитивной информатики,
аспирант Департамента социологии
Национального исследовательского университета
«Высшая школа экономики»,
Санкт-Петербург, Россия;
e-mail: iasmekalin@hse.ru

Роль искусственного интеллекта в обнаружении недостоверной информации: обзор новейших исследований и их значение для социальных наук

УДК: 316
DOI: 10.24412/2079-0910-2025-4-155-171

Рост популярности больших языковых моделей (БЯМ) меняет информационное поведение пользователей, включая способы поиска и оценки достоверности информации. Настоящее исследование представляет обзор предметного поля, посвященный роли искусственного интеллекта (ИИ) в обнаружении и интерпретации недостоверной информации. Целью обзора стало выявление ключевых направлений исследований на стыке ИИ и недостоверной информации, а также определение существующих пробелов в понимании влияния ИИ на когнитивные процессы пользователей. Данный обзор включает 32 статьи, опубликованные преимущественно в 2019–2024 гг. Выделены четыре направления исследований: 1) применение краудсорсинговых подходов к проверке информации и сравнение их с оценками профессиональных фактчекеров; 2) обнаружение и отслеживание распространения ложной информации в социальных сетях с ИИ-методами; 3) автоматический фактчекинг как разработка алгоритмов и моделей для автоматизированной проверки достоверности утверждений; 4) когнитивные искажения и предвзятости ИИ при восприятии недостоверной информации. ИИ в форме БЯМ все активнее выступает не только средством поиска информации пользователями, но и ее источником. С вовлечением ИИ в информационное поведение пользователей возникли новые вызовы: к уже известным когнитивным искажениям добавились предвзятости ИИ. Особенно актуальным становится вопрос о том, усиливает ли ИИ склонность пользователя соглашаться с готовыми суждениями или может развивать критическое восприятие информации, способствуя более адаптивному информационному поведению индивидов. Обзор выявил отсутствие работ, напрямую изучающих влияние ИИ на восприятие и распознавание ложной информации пользователями.

Ключевые слова: искусственный интеллект, когнитивные искажения, информационное поведение, большие языковые модели.

Благодарность

Исследование выполнено в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики».

Автор выражает благодарность заведующей Лабораторией социальной и когнитивной информатики д. филол. н. Олесе Юрьевне Кольцовой за ценные советы и рекомендации в ходе подготовки статьи.

Введение

Технологии искусственного интеллекта, в основе которых лежат большие языковые модели (БЯМ), получают все более массовое внедрение в повседневную жизнь: ими все шире пользуются не только медиапрофессионалы и профессиональные фактчекеры, для которых БЯМ становятся привычным инструментом создания и проверки медиаконтента, но и обычные пользователи, которые все чаще прибегают к ИИ для поиска и верификации информации. Это существенно меняет многие социальные процессы, такие как медиапотребление, обучение или получение медицинских консультаций, и в ближайшем будущем должно привести к изменению соответствующих социальных институтов (медиа, образования, здравоохранения и др.) При этом ввиду новизны самого ИИ эффективность его использования для поиска и проверки информации остается мало исследованной.

По мере того как ИИ-системы, способные к генерации и обработке информации, становятся более совершенными, исследователи подчеркивают, что традиционных методов проверки достоверности уже недостаточно для борьбы с дезинформацией в цифровой среде [*Cabañes*, 2020]. Кроме того, особенности пользовательского контента требуют внедрения новых практик верификации, адаптированных к вызовам, связанным с платформами социальных медиа [*Brandtzaeg et al.*, 2016].

Теоретическая рамка

Базовым теоретическим подходом для этого обзора является теория информационного поведения [*Wilson*, 1981, 1997]. Информационное поведение выступает ответом индивида на потребность в информации, которая задана социальным контекстом. Между контекстом и потребностью существует мотивация, в соответствии с которой человек решает, что ему нужна информация. На поиск информации могут влиять внутренние и внешние барьеры, например психологические особенности индивида, особенности источника информации и технологические факторы. Чат-боты на основании ИИ выступают и источником информации, и способом ее поиска. В терминах теории информационного потребления ИИ могут влиять на большую часть компонентов модели: на контекст, на внешние барьеры среды и на формы информационного поведения.

Другой теоретической основой выступает междисциплинарный подход к исследованию недостоверной информации [*Lewandowsky et al.*, 2017]. Он утверждает, что недостоверная информация — это не просто ложная информация, которую достаточно опровергнуть, но альтернативная система знаний, на распространение

которых влияет не качество информации [*Fazio et al.*, 2015], а структура социальных сетей и особенности восприятия содержания пользователями. Социальным контекстом условий для распространения недостоверной информации выступают сокращение общественного доверия, рост неравенства и политическая поляризация. Техно-когнитивный подход (*technocognition*) сочетает когнитивные науки и технологии для исследования того, как пользователи обращаются с недостоверной информацией.

В статье обсуждается два вида ложной информации: недостоверная информация (*misinformation*) и дезинформация (*disinformation*). Отличительной характеристикой дезинформации является преднамеренное введение в заблуждение с целью причинения вреда или получения выгоды [*Hameleers, 2023*].

Методы

Согласно типологии [*Grant, Booth,* 2009], предлагаемый анализ относится к обзорам предметного поля (*scoping review*), поскольку имеет цель первоначальной оценки охвата публикаций по только формирующемуся направлению исследований. Результаты включают типологизацию литературы по ключевым характеристикам, а формальная оценка качества не применяется. Методологическая рамка обзора предметного поля [*Arksey, O'malley,* 2005] состоит из пяти шагов, которые применялись в рамках этого обзора.

1. Формулировка исследовательского вопроса: какие аспекты роли искусственного интеллекта в обнаружении и восприятии ложной информации исследуются в актуальной литературе?
2. Поиск релевантной литературы: в обзор были включены статьи, которые рассматривают стык ИИ и ложной информации, включая ИИ как метод исследования и как источник информации. В частности, поиск проводился по академическим базам данных, в том числе *PubMed*, *PubMed Central*, *Scopus*, *Web of Science*, *ScienceDirect*, *arXiv*, *ACL Anthology* и *SpringerLink*, а также по материалам ведущих международных конференций, таких как: *International Joint Conference on Artificial Intelligence* (IJCAI), *Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL), *ACM International Conference on Information and Knowledge Management* (CIKM). Для поиска использовались такие ключевые слова, как «автоматический фактчекинг», «ИИ и недостоверная информация», «предвзятости ИИ».
3. Отбор релевантных исследований: критериями включения публикации в обзор являлись временные рамки между 2015 и 2025 гг., рассмотрение предмета исследования на стыке ИИ и ложной информации, наличие новизны в исследовательских методах, представительство разных академических дисциплин. Изначально было отобрано 75 статей, опубликованных с 2016 по 2025 г., с общим количеством цитирований более 20 тыс.; далее из них было отобрано 32 статьи, которые лучше соответствовали задачам исследования (не дублировали уже рассмотренные решения поставленных исследовательских проблем; фокусировались на недостоверной информации, а не на языке ненависти; писали о новых подходах к верификации информации, а не о техническом увеличении точности старых).

4. Систематизация данных: полученная информация была организована по характеристикам исследований вроде ключевых концептов, данных, методов их анализа, а также вкладу в общее исследовательское поле. Вся информация представлена в результатах обзора.
5. Сводка, обобщение и представление результатов: обобщение и синтез позволили сгруппировать исследования по тематическим направлениям и сформулировать пробел в исследованиях, на стыке искусственного интеллекта и ложной информации.

Результаты

Обзор включает 32 статьи, опубликованные преимущественно в 2019–2024 гг., сгруппированные по четырем направлениям: 1) краудсорсинг и экспертная проверка; 2) обнаружение и распространение недостоверной информации в соцсетях; 3) автоматический фактчекинг; 4) когнитивные и ИИ-предвзятости в восприятии информации.

Экспертная оценка и краудсорсинг как методы проверки информации

Вопрос «Может ли толпа объективно обнаружить недостоверную информацию?» задается в одноименной публикации [*Roitero et al.,* 2020]. Авторы исходят из того, что профессионалы, которые занимаются проверкой фактов (в том числе фактчекеры и журналисты), не справляются с объемами информации, которая нуждается в проверке. Это, как отмечается в тексте, создает необходимость в децентрализованных стратегиях, таких как краудсорсинг, позволяющих привлекать множество участников к процессу проверки достоверности утверждений. Дизайн исследования состоял в том, чтобы проверить степень соответствия между оценками достоверности профессионалов и пользователей краудсорсинговых платформ. В качестве метрик использовались точность полученных оценок, степень консенсуса между участниками и уровень корреляции между коллективными суждениями и оценками профессиональных фактчекеров. Участнику предлагался опросник с 11 утверждениями, а также вопросами о социально-демографических и ценностных установках. Дизайн эксперимента включал в себя встроенную поисковую систему для проверки информации, позволяя участникам обращаться к релевантным источникам в процессе выполнения задания. Агрегированные данные пользователей краудсорсинговых платформ демонстрируют высокую степень соответствия экспертным оценкам, причем увеличение размерности шкалы (3-балльная, 6-балльная и 100-балльная) не привело к значимому улучшению качества суждений. Когнитивные способности и политическая ориентация респондентов оказали значимое влияние на их способность распознавать ложные утверждения.

Другое исследование ставит аналогичную задачу — оценить эффективность краудсорсинга как метода проверки достоверности информации [*Saeed et al.,* 2022]. В исследовании используются данные, полученные с платформы краудсорсинговой верификации фактов *Birdwatch*. Она использует данные из «Твиттера» и следует алгоритму: обнаружение и выбор утверждений, поиск доказательств, верификация

утверждений (похожий алгоритм используется при автоматическом фактчекинге). Данные с краудсорса сравниваются с данными от профессиональных фактчекеров с платформы *ClaimReview*. Сравнение источников показало, что источники людей с краудсорсинговой платформы были более разнообразными (например, *Wikipedia*, *YouTube*, научные статьи), в то время как профессиональные фактчекеры использовали меньше источников, но они были более специализированные. К похожим результатам приходит другое исследование: в отличие от профессиональных фактчекеров фактчекеры-активисты при проверке информации опирались на более узкий набор источников и менее точно различали фактические неточности и утверждения, вводящие в заблуждения [*Tsang et al.*, 2023]. Также активисты более склонны критиковать недостоверную информацию, а не приводить достоверную.

При этом есть данные, которые подтверждают влияние политической ориентации на оценку правдивости утверждений [*La Barbera et al.,* 2020]. Для этого использовался набор данных *PolitiFact*, который содержит 12 800 политических утверждений с разметкой достоверности по 6-балльной шкале, присвоенной экспертами-фактчекерами. Разметка данных проводилась при помощи 400 исполнителей, которые оценивали по 10 утверждений, в том числе два контрольных утверждения и по три утверждения на определение политических взглядов. Оценщики с крауд-платформ склонны отдавать предпочтение материалам, соответствующим их убеждениям. Шкалы, которые используют оценщики, сами по себе выступают предметом исследования. На тех же самых данных из 12 800 утверждений *PolitiFact* показано [*Soprano et al.*, 2021], что согласованность оценок варьируется в зависимости от аспекта «правдивости», который используется (правильность, нейтральность, понятность, точность, полнота, достоверность, информативность). При этом упрощение оценки достоверности до трех признаков (непредвзятость, точность и правдивость) позволило прийти к более согласованным выводам, чем в прошлых исследованиях [*Barbera et al.*, 2024]. В качестве источника информации использовался тот же *PolitiFact*, из которого выбрали 120 утверждений, по 20 на каждого оценщика.

До сих пор говорилось о сравнении оценок профессиональных оценщиков и людей с краудсорсинговых платформ. Эксперимент «Фактчекинг фактчекеров» [*Lee et al.*, 2023] демонстрирует высокую согласованность разных фактчекеров между собой. Его данные включают корпус из 24 169 фактчекинговых статей, которые были собраны с помощью веб-скрепинга с сайтов организаций-фактчекеров. Для сопоставления утверждений из разных баз использовались TF-DF-векторизация и BERT, а точность соответствия оценок составила 0,96. Некоторая «наивность» такой оценки состоит в том, что совпадение оценок у разных фактчекеров констатируется даже тогда, когда они отметили информацию ложной по разным основаниям [*Uscinski*, 2015]. Они также могут ошибаться в одних и тех же местах, особенно если оценщики являются профессиональными и имеют схожие идеологические предвзятости.

Обнаружение и отслеживание распространения недостоверной информации

Методы, основанные на машинном обучении, вносят вклад в отслеживание распространения недостоверной информации в социальных сетях. На эту тему есть си-

стематический обзор литературы [*Wang et al.*, 2019], где отмечается, что в литературе о распространении ложной информации преобладают исследования на медицинские темы. Обзор включил 57 публикаций за период с 2012 по 2018 г., посвященных недостоверной информации в сфере здравоохранения в контексте социальных медиа. Анализ совместного цитирования выявил четыре междисциплинарных кластера, к которым относятся публикации на эту тему: социальная психология, коммуникации, медицинские и биомедицинские науки, общественное здравоохранение. В этот обзор также попали статьи, которые основаны на информации по медицинским или политическим вопросам.

Ярким примером решения задачи отслеживания распространения недостоверной информации по социальным медиа выступает исследование [*Ghenai*, *Mejova*, 2017], которое посвящено выявлению недостоверной информации о вирусе Зика в социальных сетях с использованием методов машинного обучения для мониторинга общественных нарративов в области здравоохранения. Исследование анализирует, как различные типы ложной информации влияют на скорость и охват ложных нарративов, а также как можно снизить их воздействие с помощью целевых интервенций. В статье о распространении информации в различных социальных сетях описываются разные методы, которые платформы используют для того, чтобы обнаруживать недостоверную информацию [*Cohen et al.*, 2020]. В качестве доступных методов перечисляются модели на основании искусственного интеллекта, которые в качестве исходных данных принимают не только текст, но и характеристики сетевого взаимодействия: данные об оценках и комментариях поста, как связаны комментаторы друг с другом и с автором публикации, оценка тональности комментариев.

На социально-сетевой подход опирается исследование о структуре дискурсивных сообществ в социальных сетях [*Mattei et al.,* 2022]. Сравнивая распространение информации с эпидемией, авторы вводят термин «инфодемия»: есть источники инфекции и есть ее переносчики. Исследователи описывают структуру дискурсивных сообществ как галстук-бабочку: исходящий сегмент выступает односторонним источником информации для ядра, а ядро выступает односторонним источником информации для входящего сегмента. При этом на периферии, вне ядра, качество контента значимо более низкое и включает недостоверную информацию. Неполная связность сегментов способствует распространению инфодемии. Роль кластеров при распространении ложной информации известна как эффект эхо-камер. На наборе данных из более чем одного миллиарда твитов, связанных с COVID-19, сетевое моделирование и картографирование эхо-камер было дополнено методами тематического моделирования для выявления распространенных нарративов [*Chen et al.*, 2022]. Картографирование эхо-камер показало, что пользователи, которые публиковали ложную информацию, чаще принадлежали к одному кластеру. Тематическое моделирование выявило, что эхо-камеры различаются по политическим признакам.

В 2019 г. различные модели машинного обучения также использовались для решения более ранней и смежной задачи — обнаружения оскорбительного языка [*Zampieri et al.*, 2019]. Тогда же была поставлена задача о том, чтобы на основании стилистических особенностей текста отличать пропаганду от не пропаганды [*Barrón-Cedeño et al.*, 2019]. Языковые признаки помогали моделям машинного обучения лучше классифицировать тексты. Задачей исследований, в рамках которых большие

языковые модели стали самостоятельным методом анализа, стала автоматическая проверка достоверности информации, то есть автоматический фактчекинг.

Автоматический фактчекинг

Платформы вопросов и ответов (*Community Question Answering, cQA*), такие как *Yahoo! Answers* и *Stack Overflow*, стали одними из первых источников данных для использования языковых моделей [*Srba, Bielikova*, 2016]. Языковые модели использовались для поиска экспертов, которые с большей вероятностью могли бы ответить на поставленный вопрос. Одним из первых примеров системы автоматической проверки достоверности утверждений является модель: [*Karadzhov et al.*, 2017]. В публикации автоматизированный фактчекинг (*automatic fact checking*) рассматривается как вариант решения задачи валидации информации из Интернета. В качестве принципов для решения этой задачи авторы предлагают универсальность, надежность, простоту, возможность повторного использования и высокое качество моделей машинного обучения. Авторы разработали систему фактчекинга, классифицирующую публикации как истинные или ложные на основе сходства с утверждениями из *cQA*. Для сопоставления использовалось 992 кластера публикаций и три метрики: важность каждого слова в документе (TF-IDF), векторные представления слов и тематическая близость. Применение модели для проверки фактов в форумах *cQA* показало улучшение по сравнению с базовыми моделями.

Отдельная задача в рамках автоматического фактчекинга состоит в том, чтобы обнаружить, проходила ли информация фактчекинг ранее. Исследователи [*Shaar et al.*, 2020] предложили ранжировать проверенные утверждения по степени их полезности для верификации нового утверждения. Для этого использовались пары «входное — проверенное утверждение» из двух датасетов, векторизация текста с помощью BERT и расчет косинусного сходства. Метод показал значимое улучшение качества ранжирования по сравнению с предыдущими подходами. В рамках развития автоматического фактчекинга проект *CheckThat! 2020* [*Barrón-Cedeño* et al., 2020] предложил инструменты для автоматической идентификации и верификации утверждений в социальных медиа. Данные, включая публикации по COVID-19, собирались с помощью веб-скрепинга и аннотировались экспертами. Для верификации также использовались внешние датасеты с проверенными утверждениями. В качестве методов применялись трансформеры (BERT, RoBERTa) и SVM — как по отдельности, так и в гибридных комбинациях — для извлечения доказательств и предсказания достоверности утверждений.

В обзорном докладе [*Nakov et al.,* 2021] дается определение того, какие этапы включает в себя автоматический фактчекинг: извлечение утверждений из текстовой информации, отбор тех, которые достойны проверки, и определение их достоверности. Роль ИИ состоит в том, что он помогает отбирать наиболее важные утверждения. Для верификации утверждений используются алгоритмы сопоставления текста с проверенными базами. Авторы выделяют ключевые задачи автоматического фактчекинга: мониторинг недостоверной информации, отбор утверждений для проверки и их последующая верификация.

Когнитивные искажения и предвзятости ИИ при восприятии недостоверной информации

Восприимчивость к недостоверной информации в социальных сетях варьируется в зависимости от социального портрета пользователя (например, американские респонденты старше 65 лет делились фейковыми новостями в семь раз чаще, чем респонденты в возрасте 18–29 лет) [*Guess et al.*, 2019]. В статье, посвященной инфодемии во время COVID-19, делается вывод о том, что фактчекинга недостаточно для предотвращения негативного воздействия дезинформации [*Chou et al.*, 2021], и в первую очередь — из-за когнитивных искажений, которые ускоряют распространение недостоверной информации.

В процессе фактчекинга, проводимого как с помощью автоматических средств, так и с помощью краудсорсинга, важно учитывать, что на оценщиков влияют когнитивные искажения. В обзоре когнитивных предвзятостей в фактчекинге выделяется 39 искажений, которые могут влиять на качество фактчекинга, и 11 контрмер против них [*Soprano et al.*, 2024]. Значимость статьи заключается в акценте на предвзятости автоматизации — явлении, при котором автоматизированные системы предоставляют информацию для пользователя и искажают правильные решения, принятые оценщиком. В качестве контрмер предлагается контроль поисковой системы исследователем (рекомендательные алгоритмы могут закладывать индивидуальные смещения для оценщиков), информирование оценщиков о наличии автоматизированной системы поиска информации, а также инструкции о скептическом отношении к информации, которую они могут получить из этих систем. Это ставит вопрос об ИИ не только как о средстве для проверки достоверности информации, но и как об ее источнике. Пользователи также склонны больше верить той ложной информации от ИИ, которая соответствует их представлениям о том, как правильно проверять достоверность информации [*Shin et al.*, 2024]. Это согласовывается с тем, что новости с доминирующим нарративом воспринимаются респондентами как более достоверные [*Bryanov et al.*, 2023].

В контексте недостоверной информации у искусственного интеллекта есть еще одно применение — это проверка информации при помощи чат-ботов с генеративным искусственным интеллектом. В этом инструменте большие языковые модели выступают методом не анализа данных, а генерации ответов о достоверности информации. При этом возникают риски предвзятости искусственного интеллекта — систематических ошибок, приводящих к несправедливым результатам [*Ferrara*, 2023]. Источниками предвзятости могут быть смещения в тренировочных данных, особенности алгоритма и интерпретация ответа пользователем. Типология [*Ferrara*, 2023] включает: 1) предвзятость выборки, возникающую из-за нерепрезентативных данных или систематических ошибок измерения, и 2) предвзятость алгоритма, при которой отдельные атрибуты оказываются непропорционально значимыми. В контексте машинного обучения «галлюцинацией» называют явление, при котором модель генерирует выходные данные, не соответствующие входным данным, создавая ложную или бессмысленную информацию. В академической литературе вместо термина *«галлюцинация»* все чаще используется концепт «искажения ИИ» (*AI misinformation*) — то есть ложная информация, сгенерированная ИИ [*Hatem et al.*, 2023]. При этом авторами подчеркивается, что интерпретация таких ответов во многом определяется самим пользователем.

В исследовании [*Pan et al.*, 2023] о применении искусственного интеллекта в контексте проверки достоверности информации используется следующий дизайн: унифицированные запросы с утверждениями на медицинскую тематику направляются различным генеративным моделям искусственного интеллекта; далее ответы чат-ботов оцениваются по достоверности, практической полезности и читаемости. По результатам опытов авторов статьи достоверность выдачи оказалась высокой, понятность средней, а полезность — низкой, что отчасти может объясняться особенностью сформулированного запроса. Дизайн другого исследования состоял в том, чтобы оценить, насколько успешно чат-бот на основании БЯМ может пройти стандартизированный медицинский экзамен [*Kung et al.*, 2023]. Результаты показали, что даже без дополнительной настройки модели специализированными данными ее ответы позволяют достичь проходного балла медицинского экзамена.

У экспериментов с искажениями ИИ есть другое ограничение, которое выражается в настройке точного запроса для получения более точного ответа. Это ограничение учитывается при добавлении вариативности по модели, строению запроса и сложности тематики [*Wang et al.*, 2024]. Исследователи рассматривали точность выдачи генеративного ИИ на примере *ChatGPT* в разрезе разных моделей, запросов разных грамматических структур и тематик от простой до сложной. Исследование показало, что формулировки запросов способны влиять на точность информации в выдаче ИИ, но даже при одинаковых вопросах достоверность ответов чат-бота может отличаться. Отдельным сценарием использования генеративного ИИ является получение рекомендаций, а не просто информации. В исследовании ставится вопрос о том, насколько будут похожими рекомендации чат-бота на реальные решения врачей [*Williams et al.,* 2024]. На материале более чем 10 тыс. случаев неотложной медицинской помощи чат *GPT-3.5-turbo* показал точность 24% в сравнении с решениями врачей-ординаторов. Информирование со стороны ИИ, в том числе посредством диалога с чат-ботом, может снижать уровень доверия теориям заговора: в эксперименте с 2 190 участниками участие в диалоге с чат-ботом на основании большой языковой модели вера в теории заговора снижалась на 20–21% [*Costello et al.,* 2024].

Дискуссия

Распространение недостоверной информации является одним из самых острых вызовов для цифровых платформ и общества в целом ввиду его все большей информатизации и цифровизации. Применение инструментов проверки информации означает методологический компромисс, поскольку каждый инструмент решает одни методологические проблемы и создает другие. Эти ограничения необходимо учитывать при работе с инструментами верификации, и они представляют собой самостоятельный предмет исследования. Инструменты обнаружения недостоверной информации за десять лет эволюционировали от краудсорсинга до широкого использования искусственного интеллекта для автоматизированного фактчекинга. При этом на своем пути развитие инструментов сталкивалось с барьерами: когнитивные способности и политическая ориентация оценщиков, размерность шкал для оценки, когнитивные искажения и точность данных. Использование искусственного интеллекта для проверки достоверности утверждений на индивидуальном уровне

сталкивается с новым вызовом — искажениями ИИ, когда предвзятость ИИ приводит к тому, что инструмент выступает не средством обнаружения недостоверной информации, а ее источником.

Для проверки информации можно использовать экспертов. Их оценки хорошо согласовываются между собой [*Lee et al.*, 2023], однако это совпадение может быть следствием общих когнитивных искажений, институциональных рамок [*Uscinski*, 2015] или идеологической предвзятости [*La Barbera et al.,* 2020]. Применение этого инструмента опосредовано относительно более высокими издержками по времени и привлечению экспертов. Коллективный способ проверки информации опирается не на узкоспециализированных экспертов, а на пользователей краудсорсинговых платформ. С одной стороны, у них нет смещения из-за своей специализации в теме. С другой стороны, они используют более доступные источники информации [*Saeed et al.,* 2022] и хуже различают манипулятивные утверждения [*Tsang et al.*, 2023]. В этом способе угроза политической и институциональной предвзятости становится меньше, поскольку самих оценщиков становится меньше. Однако недостаток их квалификации приводит к снижению точности определения недостоверной информации. Таким образом, выбор между экспертной и коллективной оценками состоит в методологическом компромиссе между глубиной экспертизы и широтой вовлечения участников, а также ограничениями по точности классификации и по специализации источников информации.

Автоматический фактчекинг с использованием алгоритмов машинного обучения позволяет обрабатывать большие массивы информации и отбирать утверждения, требующие проверки. Этот метод появился как продолжение коллективных методов проверки информации и использовался профессиональными исследователями. Он частично снимает ограничения масштабируемости и скорости проверки, присущие экспертам. Тем не менее точность оценки ограничена качеством тренировочных данных для модели и предвзятости самих алгоритмов [*Soprano et al.*, 2024; *Ferrara*, 2023]. Когнитивные искажения начинают играть важную роль, когда результаты ИИ интерпретируют пользователи без профессиональных навыков проверки информации.

Направления будущих исследований

В то время как изучение эффективности БЯМ для автоматического фактчекинга только начинается, исследования краудсорсинга показывают, что, несмотря на разногласия в оценках отдельных участников, грамотная агрегация данных повышает их соответствие экспертным оценкам, делая краудсорсинг эффективным инструментом борьбы с дезинформацией. Кроме того, методы коррекции когнитивных искажений, включая обратную связь от ИИ, помогают снижать доверие к ложным сведениям и теориям заговора. Вместе с тем большинство существующих работ сосредоточено либо на технической стороне задач обнаружения недостоверной информации, либо на анализе роли различных социальных и когнитивных факторов в точности оценки информации. В меньшей степени исследуется, как именно взаимодействие с ИИ (в том числе с языковыми моделями) влияет на когнитивные процессы, лежащие в основе восприятия, интерпретации и доверия к информации.

Таким образом, обзор выявляет существенный пробел — отсутствие работ, напрямую изучающих влияние ИИ на восприятие и распознавание ложной информации пользователями. Неясно, способствует ли обращение к ИИ формированию более критического отношения к информации или усиливает предвзятость в условиях, когда пользователь получает готовые утверждения от ИИ. Также недостаточно изучены механизмы, посредством которых ИИ-ответы могут менять суждения пользователей: за счет аргументации и стилистических особенностей. В этом контексте возможно более системное и междисциплинарное исследование того, как взаимодействие пользователя с ИИ влияет на способность оценивать достоверность информации. Такая способность напрямую связана с возможностями адаптации индивидов в обществе, с продуктивностью их решений и, как следствие, с функциональностью общественных систем в целом, поэтому такое исследование представляло бы существенный научный и практический интерес.

Заключение

В условиях цифровизации общества искусственный интеллект в форме больших языковых моделей все активнее включается в информационное поведение пользователей, выступая не только средством поиска информации, но и ее источником, посредником в интерпретации и оценке достоверности. С позиции техно-когнитивного подхода это означает, что ИИ становится фактором, опосредующим восприятие достоверности, аргументации и доверия к полученной информации.

История развития инструментов проверки достоверности информации началась с экспертного фактчекинга и краудсорсинговых инициатив, которые позволяли сообществу пользователей участвовать в оценке утверждений. Со временем эти подходы дополнились методами машинного обучения, которые начали использоваться для анализа больших объемов данных и выявления закономерностей в распространении недостоверной информации. Это привело к появлению систем автоматического фактчекинга, основанных на технологиях обработки естественного языка и предсказательных моделей. Однако с вовлечением ИИ в информационное поведение пользователей возникли новые вызовы: к уже известным когнитивным искажениям добавились предвзятости ИИ, связанные с алгоритмами, обучающими выборками и автоматической генерацией ответов.

Особенно актуальным становится вопрос о том, усиливает ли ИИ склонность пользователя соглашаться с готовыми суждениями или может развивать критическое восприятие информации, способствуя более адаптивному поведению индивидов в социуме. Будущие исследования могут быть направлены на роль, которую ИИ играет при оценке достоверности информации и принятии решений на ее основе, в том числе — представителями разных социальных групп.

Литература

Arksey H., O'Malley L. Scoping Studies: Towards a Methodological Framework // International Journal of Social Research Methodology. 2005. Vol. 8. No. 1. P. 19–32. DOI: 10.1080/1364557032000119616.

Barrón-Cedeño A., Elsayed T., Nakov P., Da San Martino G., Hasanain M., Suwaileh R., Ali Z.S. Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media // Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11. Springer International Publishing, 2020. Vol. 12036. P. 215–236. DOI: 10.48550/arxiv.2007.07997.

Barrón-Cedeño A., Jaradat I., Da San Martino G., Nakov P. Proppy: Organizing the News Based on Their Propagandistic Content // Information Processing & Management. 2019. Vol. 56. No. 5. P. 1849–1864. DOI: 10.1016/j.ipm.2019.03.005.

Barbera D.L., Maddalena E., Soprano M., Roitero K., Demartini G., Ceolin D., Mizzaro S. Crowdsourced Fact-Checking: Does It Actually Work? // Information Processing & Management. 2024. Vol. 61. No. 5. Article 103792. DOI: 10.1016/j.ipm.2024.103792.

Brandtzaeg P.B., Lüders M., Spangenberg J., Rath-Wiggins L., Følstad A. Emerging Journalistic Verification Practices Concerning Social Media // Journalism Practice. 2016. Vol. 10. No. 3. P. 323–342. DOI: 10.1080/10584609.2023.2172492.

Bryanov K., Kliegl R., Koltsova O., Miltsov A., Pashakhin S., Porshnev A., Sinyavskaya Y., Terpilovskii M., Vziatysheva V. What Drives Perceptions of Foreign News Coverage Credibility? A Cross-National Experiment Including Kazakhstan, Russia, and Ukraine // Political Communication. 2023. Vol. 40. No. 2. P. 115–146. DOI: 10.1080/17512786.2015.1020331.

Cabañes J. Digital Disinformation and the Imaginative Dimension of Communication // Journalism & Mass Communication Quarterly. 2020. Vol. 97. No. 2. P. 435–452. DOI: 10.1177/1077699020913799.

Chen E., Jiang J., Chang H.-C.H., Muric G., Ferrara E. Charting the Information and Misinformation Landscape to Characterize Misinfodemics on Social Media: COVID-19 Infodemiology Study at a Planetary Scale // JMIR Infodemiology. 2022. Vol. 2. No. 1. Article e32378. DOI: 10.2196/32378.

*Chou W.-Y. S., Gaysynsky A., Vanderpool R.*C. The COVID-19 Misinfodemic: Moving beyond Fact-Checking // Health Education & Behavior. 2021. Vol. 48. No. 1. P. 9–13. DOI: 10.1177/1090198120980675.

Cohen R., Moffatt K., Ghenai A., Yang A., Corwin M., Lin G., Gray L. Addressing Misinformation in Online Social Networks: Diverse Platforms and the Potential of Multiagent Trust Modeling // Information. 2020. Vol. 11. No. 11. P. 539. DOI: 10.3390/info11110539.

Costello T.H., Pennycook G., Rand D.G. Durably Reducing Conspiracy Beliefs through Dialogues with AI // Science. 2024. Vol. 385. No. 6714. Article eadq1814. DOI: 10.31234/osf.io/xcwdn.

Fazio L.K., Brashier N.M., Payne B.K., Marsh E.J. Knowledge Does Not Protect Against Illusory Truth // Journal of Experimental Psychology: General. 2015. Vol. 144. No. 5. P. 993–1002. DOI: 10.1037/xge0000098.

Ferrara E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. 2023. Vol. 6. No. 1. Article 3. DOI: 10.48550/ARXIV.2304.07683.

Ghenai A., Mejova Y. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter // Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI). 2017. P. 518–518. DOI: 10.48550/arxiv.1707.03778.

Grant M.J., Booth A. A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies // Health Information & Libraries Journal. 2009. Vol. 26. No. 2. P. 91–108. DOI: 10.1111/j.1471-1842.2009.00848.x.

Guess A., Nagler J., Tucker J. Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook // Science Advances. 2019. Vol. 5. No. 1. Article eaau4586. DOI: 10.1126/sciadv.aau4586.

Hameleers M. Disinformation as a Context-Bound Phenomenon: Toward a Conceptual Clarification Integrating Actors, Intentions and Techniques of Creation and Dissemination // Communication Theory. 2023. Vol. 33. No. 1. P. 1–10. DOI: 10.1093/ct/qtac021.

Hatem R., Simmons B., Thornton J.E. A Call to Address AI 'Hallucinations' and How Healthcare Professionals Can Mitigate Their Risks // Cureus. 2023. Vol. 15. No. 9. Article e44720. DOI: 10.7759/cureus.44720.

Karadzhov G., Nakov P., Màrquez L., Barrón-Cedeño A., Koychev I. Fully Automated Fact Checking Using External Sources // Proceedings of RANLP 2017 — Recent Advances in Natural Language Processing Meet Deep Learning. 2017. P. 344–353. DOI: 10.26615/978-954-452-049-6_046.

Kung T.H., Cheatham M., Medenilla A., Sillos C., De Leon L., Elepaño C., Tseng V. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models // PLOS Digital Health. 2023. Vol. 2. No. 2. Article e0000198. DOI: 10.1371/journal.pdig.0000198.

La Barbera D., Roitero K., Demartini G., Mizzaro S., Spina D. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias // Lecture Notes in Computer Science. 2020. Vol. 12036. P. 207–214. DOI: 10.1007/978-3-030-45442-5_26.

Lee S., Xiong A., Seo H., Lee D. "Fact-Checking" Fact Checkers: A Data-Driven Approach // Harvard Kennedy School Misinformation Review. 2023. DOI: 10.37016/mr-2020-126.

Lewandowsky S., Ecker U.K.H., Cook J. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era // Journal of Applied Research in Memory and Cognition. 2017. Vol. 6. No. 4. P. 353–369. DOI: 10.1016/j.jarmac.2017.07.008.

Mattei M., Pratelli M., Caldarelli G., Petrocchi M., Saracco F. Bow-Tie Structures of Twitter Discursive Communities // Scientific Reports. 2022. Vol. 12. No. 1. Article 12944. DOI: 10.1038/s41598-022-16603-7.

Nakov P., Corney D., Hasanain M., Alam F., Elsayed T., Barrón-Cedeño A., Da San Martino G. Automated Fact-Checking for Assisting Human Fact-Checkers // Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. 2021. P. 4551–4557. DOI: 10.24963/ijcai.2021/619.

Pan A., Musheyev D., Bockelman D., Loeb S., Kabarriti A.E. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries about Cancer // JAMA Oncology. 2023. Vol. 9. No. 10. P. 1437–1440. DOI: 10.1001/jamaoncol.2023.2947.

Roitero K., Soprano M., Fan S., Spina D., Mizzaro S., Demartini G. Can the Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background // arXiv. 2020. DOI: 10.48550/arXiv.2005.06915.

Saeed M., Traub N., Nicolas M., Demartini G., Papotti P. Crowdsourced Fact-Checking at Twitter // Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022. P. 3815–3819. DOI: 10.48550/arXiv.2208.09214.

Shaar S., Babulkov N., Da San Martino G., Nakov P. That Is a Known Lie: Detecting Previously Fact-Checked Claims // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P. 3607–3618. DOI: 10.48550/arXiv.2005.06058.

Shin D., Jitkajornwanich K., Lim J.S., Spyridou A. Debiasing Misinformation: How Do People Diagnose Health Recommendations from AI? // Online Information Review. 2024. Vol. 48. No. 5. P. 1025–1044. DOI: 10.1108/OIR-04-2023-0167.

Soprano M., Roitero K., La Barbera D., Ceolin D., Spina D., Demartini G., Mizzaro S. The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale // Information Processing & Management. 2021. Vol. 58. No. 6. Article 102710. DOI: 10.1016/j.ipm.2021.102710.

Soprano M., Roitero K., La Barbera D., Ceolin D., Spina D., Demartini G., Mizzaro S. Cognitive Biases in Fact-Checking and Their Countermeasures: A Review // Information Processing & Management. 2024. Vol. 61. No. 3. Article 103672. DOI: 10.1016/j.ipm.2024.103672.

Srba I., Bielikova M. A Comprehensive Survey and Classification of Approaches for Community Question Answering // ACM Transactions on the Web. 2016. Vol. 10. No. 3. P. 1–63. DOI: 10.1145/2934687.

Tsang N.L.T., Feng M., Lee F.L.F. How Fact-Checkers Delimit Their Scope of Practices and Use Sources: Comparing Professional and Partisan Practitioners // Journalism. 2023. Vol. 24. No. 10. P. 2232–2251. DOI: 10.1177/14648849221100862.

Uscinski J.E. The Epistemology of Fact Checking (Is Still Naïve): Rejoinder to Amazeen // Critical Review. 2015. Vol. 27. No. 2. P. 243–252. DOI: 10.1080/08913811.2015.1055892.

Wang L., Chen X., Deng X., Wen H., You M., Liu W., Li J. Prompt Engineering in Consistency and Reliability with the Evidence-Based Guideline for LLMs // NPJ Digital Medicine. 2024. Vol. 7. No. 1. Article 41. DOI: 10.1038/s41746-024-01029-4.

Wang Y., McKee M., Torbica A., Stuckler D. Systematic Literature Review on the Spread of Health-Related Misinformation on Social Media // Social Science & Medicine. 2019. Vol. 240. Article 112552. DOI: 10.1016/j.socscimed.2019.112552.

Williams C.Y.K., Miao B.Y., Kornblith A.E., Butte A.J. Evaluating the Use of Large Language Models to Provide Clinical Recommendations in the Emergency Department // Nature Communications. 2024. Vol. 15. No. 1. Article 8236. DOI: 10.1038/s41467-024-52415-1.

Wilson T.D. On User Studies and Information Needs // Journal of Documentation. 1981. Vol. 37. No. 1. P. 3–15.

Wilson T.D. Information Behaviour: An Interdisciplinary Perspective // Information Processing & Management. 1997. Vol. 33. No. 4. P. 551–572. DOI: 10.1016/S0306-4573(97)00028-9.

Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R. Predicting the Type and Target of Offensive Posts in Social Media // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. P. 1415–1420. DOI: 10.18653/v1/N19-1144.

The Role of Artificial Intelligence in Detecting Misinformation: A Review of Recent Research and Its Implications for Social Sciences

Ivan A. Smekalin

National Research University "Higher School of Economics",
St. Petersburg, Russia;
e-mail: iasmekalin@hse.ru

The growing popularity of large language models (LLMs) is reshaping users' information behavior, including how they search for and assess the credibility of information. This study presents a scoping review focused on the role of artificial intelligence in the detection and interpretation of misinformation. The review aims to identify key research directions at the intersection of AI and misinformation, as well as to highlight existing gaps in understanding how AI influences users' cognitive processes. The corpus includes 32 publications, mostly published between 2019 and 2024. Four main areas of research are identified: 1) the use of crowdsourcing approaches for information verification and their comparison with professional fact-checkers; 2) detection and monitoring of misinformation spread on social media using AI-based methods; 3) automatic fact-checking as the development of models and algorithms for verifying claims; and 4) cognitive biases and algorithmic biases in users' perception of misinformation. LLMs increasingly function not only as tools for information retrieval, but also as sources of information themselves. The integration of AI into users' information behavior has introduced new challenges: in addition to existing cognitive biases, AI biases and AI misinformation

have occurred. A pressing question emerges: does interaction with AI reinforce users' tendency to accept ready-made judgments, or can it foster more critical engagement with information, leading to more adaptive behavior? The review reveals a notable gap in empirical research directly examining the impact of AI on users' perception and recognition of false information.

Keywords: artificial intelligence, cognitive biases, information behavior, large language models.

Acknowledgment

The research is completed within the framework of Fundamental Research Program of National Research University "Higher School of Economics".

The author expresses gratitude to Dr. Olesya Y. Koltsova, Head of the Laboratory for Social and Cognitive Informatics, for her valuable advice and recommendations during the preparation of this article.

References

Arksey, H., O'Malley, L. (2005). Scoping Studies: Towards a Methodological Framework, *International Journal of Social Research Methodology, 8* (1), 19–32. DOI: 10.1080/1364557032000119616.

Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Ali, Z.S. (2020). Overview of CheckThat! 2020: Automatic Identification and Verification of Claims in Social Media, *Lecture Notes in Computer Science, 12036*, 215–236. DOI: 10.48550/arxiv.2007.07997.

Barrón-Cedeño, A., Jaradat, I., Da San Martino, G., Nakov, P. (2019). Proppy: Organizing the News Based on Their Propagandistic Content, *Information Processing & Management, 56* (5), 1849–1864. DOI: 10.1016/j.ipm.2019.03.005.

Barbera, D.L., Maddalena, E., Soprano, M., Roitero, K., Demartini, G., Ceolin, D., Mizzaro, S. (2024). Crowdsourced Fact-Checking: Does It Actually Work? *Information Processing & Management, 61* (5), 103792. DOI: 10.1016/j.ipm.2024.103792ISBN: 0306-4573.

Bryanov, K., Kliegl, R., Koltsova, O., Miltsov, A., Pashakhin, S., Porshnev, A., Sinyavskaya, Y., Terpilovskii, M., Vziatysheva, V. (2023). What Drives Perceptions of Foreign News Coverage Credibility? A Cross-National Experiment Including Kazakhstan, Russia, and Ukraine, *Political Communication*, *40* (2), 115–146. DOI: 10.1080/10584609.2023.2172492.

Brandtzaeg, P.B., Lüders, M., Spangenberg, J., Rath-Wiggins, L., Følstad, A. (2016). Emerging Journalistic Verification Practices Concerning Social Media, *Journalism Practice*, *10* (3), 323–342. DOI: 10.1080/17512786.2015.1020331.

Cabañes, J. (2020). Digital Disinformation and the Imaginative Dimension of Communication, *Journalism & Mass Communication Quarterly, 97* (2), 435–452. DOI: 10.1177/1077699020913799.

Chen, E., Jiang, J., Chang, H.-C.H., Muric, G., Ferrara, E. (2022). Charting the Information and Misinformation Landscape to Characterize Misinfodemics on Social Media: COVID-19 Infodemiology Study at a Planetary Scale, *JMIR Infodemiology, 2* (1), e32378. DOI:10.2196/32378.

Chou, W.-Y.S., Gaysynsky, A., Vanderpool, R. C. (2021). The COVID-19 Misinfodemic: Moving Beyond Fact-Checking, *Health Education & Behavior, 48* (1), 9–13. DOI: 10.1177/1090198120980675.

Cohen, R., Moffatt, K., Ghenai, A., Yang, A., Corwin, M., Lin, G., Gray, L. (2020). Addressing Misinformation in Online Social Networks: Diverse Platforms and the Potential of Multiagent Trust Modeling, *Information, 11* (11), 539. DOI: 10.3390/info11110539.

Costello, T.H., Pennycook, G., Rand, D.G. (2024). Durably Reducing Conspiracy Beliefs through Dialogues with AI, *Science, 385* (6714), eadq1814. DOI: 10.31234/osf.io/xcwdn.

Fazio, L.K., Brashier, N.M., Payne, B.K., Marsh, E.J. (2015). Knowledge Does Not Protect Against Illusory Truth, *Journal of Experimental Psychology: General, 144* (5), 993–1002. DOI: 10.1037/xge0000098.

Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci, 6* (1), 3. DOI: 10.48550/ARXIV.2304.07683.

Ghenai, A., Mejova, Y. (2017, August). Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter, in *2017 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 518–518). DOI: 10.48550/arXiv.1707.03778.

Grant, M.J., Booth, A. (2009). A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies, *Health Information & Libraries Journal, 26* (2), 91–108. DOI: 10.1111/j.1471-1842.2009.00848.x.

Guess, A., Nagler, J., Tucker, J. (2019). Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook, *Science Advances, 5* (1), eaau4586. DOI: 10.1126/sciadv.aau4586.

Hameleers, M. (2023). Disinformation as a Context-Bound Phenomenon: Toward a Conceptual Clarification Integrating Actors, Intentions and Techniques of Creation and Dissemination, *Communication Theory*, *33* (1), 1–10. DOI: 10.1093/ct/qtac021.

Hatem, R., Simmons, B., Thornton, J.E. (2023). A Call to Address AI 'Hallucinations' and How Healthcare Professionals Can Mitigate Their Risks, *Cureus, 15* (9), e44720. DOI: 10.7759/cureus.44720.

Karadzhov, G., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Koychev, I. (2017, November 10). Fully Automated Fact Checking Using External Sources, in *RANLP 2017 — Recent Advances in Natural Language Processing Meet Deep Learning* (pp. 344–353). DOI: 10.26615/978-954-452-049-6_046.

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models, *PLOS Digital Health, 2* (2), e0000198. DOI: 10.1371/journal.pdig.0000198.

La Barbera, D., Roitero, K., Demartini, G., Mizzaro, S., Spina, D. (2020). Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias, in *Lecture Notes in Computer Science*, vol. 12036, 207–214. DOI: 10.1007/978-3-030-45442-5_26.

Lee, S., Xiong, A., Seo, H., Lee, D. (2023). "Fact-Checking" Fact Checkers: A Data-Driven Approach, *Harvard Kennedy School Misinformation Review*. DOI: 10.37016/mr-2020-126.

Lewandowsky, S., Ecker, U.K.H., Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era, *Journal of Applied Research in Memory and Cognition, 6* (4), 353–369. DOI: 10.1016/j.jarmac.2017.07.008.

Mattei, M., Pratelli, M., Caldarelli, G., Petrocchi, M., Saracco, F. (2022). Bow-Tie Structures of Twitter Discursive Communities, *Scientific Reports, 12* (1), 12944. DOI: 10.1038/s41598-022-16603-7.

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Da San Martino, G. (2021, August). Automated Fact-Checking for Assisting Human Fact-Checkers. in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence* (pp. 4551–4557). DOI: 10.24963/ijcai.2021/619.

Pan, A., Musheyev, D., Bockelman, D., Loeb, S., Kabarriti, A.E. (2023). Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries about Cancer, *JAMA Oncology, 9* (10), 1437–1440. DOI: 10.1001/jamaoncol.2023.2947.

Roitero, K., Soprano, M., Fan, S., Spina, D., Mizzaro, S., Demartini, G. (2020). Can the Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background, *arXiv*. DOI: 10.48550/arXiv.2005.06915

Saeed, M., Traub, N., Nicolas, M., Demartini, G., Papotti, P. (2022, October 17). Crowdsourced Fact-Checking at Twitter, in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 3815–3819). DOI: 10.48550/arXiv.2208.09214.

Shaar, S., Babulkov, N., Da San Martino, G., Nakov, P. (2020). That Is a Known Lie: Detecting Previously Fact-Checked Claims, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3607–3618). DOI: 10.48550/arXiv.2005.06058.

Shin, D., Jitkajornwanich, K., Lim, J.S., Spyridou, A. (2024). Debiasing Misinformation: How Do People Diagnose Health Recommendations from AI?, *Online Information Review, 48* (5), 1025–1044. DOI: 10.1108/OIR-04-2023-0167.

Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Demartini, G., Mizzaro, S. (2021). The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale, *Information Processing & Management, 58* (6), 102710. DOI: 10.1108/OIR-04-2023-0167.

Soprano, M., Roitero, K., La Barbera, D., Ceolin, D., Spina, D., Demartini, G., Mizzaro, S. (2024). Cognitive Biases in Fact-Checking and Their Countermeasures: A Review, *Information Processing & Management, 61* (3), 103672. DOI: 10.1016/j.ipm.2024.103672.

Srba, I., Bielikova, M. (2016). A Comprehensive Survey and Classification of Approaches for Community Question Answering, *ACM Transactions on the Web, 10* (3), 1–63. DOI: 10.1145/2934687.

Tsang, N.L.T., Feng, M., Lee, F.L.F. (2023). How Fact-Checkers Delimit Their Scope of Practices and Use Sources: Comparing Professional and Partisan Practitioners, *Journalism, 24* (10), 2232–2251. DOI: 10.1177/14648849221100862.

Uscinski, J.E. (2015). The Epistemology of Fact Checking (Is Still Naïve): Rejoinder to Amazeen, *Critical Review, 27* (2), 243–252. DOI: 10.1080/08913811.2015.1055892.

Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, J. (2024). Prompt Engineering in Consistency and Reliability with the Evidence-Based Guideline for LLMs, *NPJ Digital Medicine, 7* (1), 41.

Wang, Y., McKee, M., Torbica, A., Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-Related Misinformation on Social Media, *Social Science & Medicine*, vol. 240, 112552. DOI: 10.1016/j.socscimed.2019.112552.

Williams, C.Y.K., Miao, B.Y., Kornblith, A.E., Butte, A.J. (2024). Evaluating the Use of Large Language Models to Provide Clinical Recommendations in the Emergency Department, *Nature Communications, 15* (1), 8236. DOI: 10.1038/s41467-024-52415-1.

Wilson, T. D. (1981). On User Studies and Information Needs, *Journal of Documentation, 37* (1), 3–15.

Wilson, T.D. (1997). Information Behaviour: An Interdisciplinary Perspective, *Information Processing & Management, 33* (4), 551–572. DOI: 10.1016/S0306-4573(97)00028-9.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1415–1420). DOI: 10.18653/v1/N19-1144.