Иван Владимирович БРЮХОВЕЦКИЙ,
соискатель кафедры информационного права и цифровых технологий Университета имени О.Е. Кутафина (МГЮА)
wangmiczuro@yandex.ru
125993, Россия, г. Москва, ул. Садовая-Кудринская, д. 9

Правовые особенности маркировки созданного генеративным искусственным интеллектом контента в КНР

Аннотация. *Данная работа посвящена анализу проблемы маркировки создаваемых искусственным интеллектом материалов, а также изучению практических подходов к данной проблеме в КНР. В исследовании приводятся данные об особенностях работы генеративного искусственного интеллекта с информационными массивами данных. Рассматривается проблема деградации информации, накопления в ней ошибок, последствий неконтролируемого распространения синтезированного контента в сети. Изучается отмеченный исследователями вопрос коллапса знаний. Приводятся основные подходы китайских законодателей относительно регулирования соответствующей проблемы. Отмечается ряд конкретных технических механизмов, содержащихся в нормативных документах КНР, о встраивании скрытой и явной маркировки в создаваемый ИИ контент.*

Ключевые слова: *искусственный интеллект, генеративный ИИ, нормативно-правовое регулирование генеративного ИИ в КНР, законодательство КНР, маркировка контента, деградация данных, коллапс знаний*

DOI: 10.17803/2311-5998.2024.122.10.176-181

Ivan V. BRUHOVETSKY,
Applicant of the Department of Information Law and Digital Technologies of the Kutafin Moscow State Law University (MSAL)
wangmiczuro@yandex.ru
9, ul. Sadovaya-Kudrinskaya, Moscow, Russia, 125993

Legal peculiarities of watermarking of content created by generative AI in PRC

Abstract. *This work is dedicated to analyzing the problem of labeling materials created by artificial intelligence, as well as studying approaches to this issue in the People's Republic of China (PRC). The study provides data on the features of generative artificial intelligence's work with large datasets. The problem of information degradation, the accumulation of errors within it, and the consequences of the uncontrolled spread of synthesized content on the internet are examined. The researchers also explore the issue of the "knowledge collapse." The main approaches of Chinese legislators regarding the regulation of this issue are outlined. Several specific technical mechanisms, contained in the regulatory documents of the PRC, related to embedding hidden and explicit labeling in AI-generated content, are noted.*

Keywords: *artificial intelligence, generative AI, regulatory framework for generative AI in the PRC, PRC legislation, content labeling, data degradation, knowledge collapse.*

Введение

В настоящее время большую актуальность приобретает вопрос определения достоверности материалов, распространяемых в сети Интернет. Сервисы генеративного искусственного интеллекта все активнее используются для создания текстов, изображений, аудиовизуального контента. Согласно оценкам, на данный момент ИИ генерирует значительный по масштабам современного цифрового мира массив контента. Ежедневно создается более 34 млн единиц визуального контента (фотографических и видеоматериалов)[1]. При этом, по данным на август 2023 г., за год использования сервисов генеративного ИИ было сформировано около 15,5 млрд единиц различного рода визуальных материалов. С учетом высокой скорости прироста сгенерированных данных возможно предположить, что уже в ближайшее время ИИ может быть одним из основных средств создания сетевого контента.

Рост качества и правдоподобности генерируемых сервисами генеративного ИИ фотографий и видеороликов создает широкое поле для неправомерного их использования. Возрастает риск распространения нежелательной информации политического характера, фейков, порочащих честь и достоинство граждан материалов. При создании материалов отмечена задача фильтрации ошибок (например, «галлюцинации» ИИ и др.)[2]. Накопление большого объема синтетической информации в сети неминуемо поставит вопрос необходимости выявления достоверных и недостоверных данных. Часть указанных проблем с распространением соответствующей информации может быть решена за счет эффективных механизмов их отслеживания, в частности маркировки.

Практические вопросы отслеживания синтетических данных

Важность определения достоверности и оригинальности контента также связана с необходимостью качественного обучения больших языковых моделей, являющихся составной частью современного генеративного ИИ. В частности, согласно ряду исследований, некачественные массивы данных приводят к быстрой деградации генеративного ИИ и невозможности его практического использования[3]. Ука-

[1] *Valyaeva A.* AI Image Statistics for 2024: How Much Content Was Created by AI // Everypixel Journal. 2023. URL: https://journal.everypixel.com/ai-image-statistics (дата обращения: 03.10.2024).

[2] European Parliament. Generative AI and watermarking // European Parliament Research Service. 2023. URL: https://www.europarl.europa.eu/regdata/etudes/brie/2023/757583/eprs_bri(2023)757583_en.pdf (дата обращения: 03.10.2024).

[3] *Shumailov I., Shumaylov Z., Zhao Y.* AI models collapse when trained on recursively generated data // Nature. 2024. Т. 631. С. 755—759. URL: https://www.nature.com/articles/s41586-024-07566-y (дата обращения: 03.10.2024).

занная проблема может приобрести большое значение в обозримом будущем, так как языковые модели искусственного интеллекта обучаются на синтетических данных и на информации, свободно размещаемой в сети, которая в большинстве случаев не может быть верифицирована.

Более того, неконтролируемое распространение недостоверных синтезированных данных в информационном пространстве несет значительные потенциальные угрозы для общего понимания достоверности данных[4]. Отмечают явление так называемого коллапса знаний (knowledge collapse), когда использование ИИ становится настолько распространенным, что человек утрачивает возможность самостоятельного критического формирования новых идей, начинает полностью полагаться на созданный контент и не может самостоятельно дифференцировать информацию по степени достоверности. С большой степенью вероятности данный процесс может привести к накоплению ошибок и деградации массива доступной человеку информации[5].

Согласно проведенным исследованиям, необратимые изменения моделей ИИ происходят достаточного быстро[6]. В частности, значительные изменения были заметны после 10 поколений рекурсивного обучения моделей.

Таким образом, в сценарии, при котором одни модели производят обучение других моделей ИИ 10-е поколение (10-я последовательно обученная модель) будет обладать информацией, отдаленно напоминающей оригинал, что ощутимо повлияет на результат ее работы. Указанная проблема может иметь большую важность, так как в случае неконтролируемого доступа обучаемого ИИ к результатам генерации других ИИ-сервисов и к свободно распространяемым сведениям, а также при отсутствии проверки на достоверность вероятно появление и постепенное встраивание в массив данных ИИ существенных ошибок. Частным случаем вышеуказанной проблемы может быть явление «канибализации» данных генеративного ИИ, когда, не отслеживая этого, модель ИИ обучается на сгенерированных ею самой данных, размещенных в сети[7]. Это приводит к быстрому накоплению ошибок и утрате достоверности.

Указанное явление обусловливает задачу поиска, фильтрации, отбора достоверного и отвечающего задаче массива информации, а также наиболее удачных образцов сгенерированных данных. Одним из решений соответствующей проблемы может быть маркирование данных.

---

[4] *Peterson A. J.* (2024). AI and the Problem of Knowledge Collapse // URL: https://arxiv.org/abs/2404.03502 (дата обращения: 03.10.2024).

[5] AI Entropy: The Vicious Circle of AI-Generated Content // URL: https://towardsdatascience.com/ai-entropy-the-vicious-circle-of-ai-generated-content-8aad91a19d4f (дата обращения: 03.10.2024).

[6] *Shumailov R., Papernot N.* The curse of recursion: training on generated data makes models forget // URL: https://arxiv.org/pdf/2305.17493v2 (дата обращения: 03.10.2024).

[7] AI is cannibalizing itself. And creating more AI // URL: https://theweek.com/tech/ai-cannibalization-model-collapse (дата обращения: 03.10.2024).

Подходы КНР к маркированию информации

Примечателен взгляд на данную проблему законодателей Китая. В июле 2023 г. Государственная канцелярия по делам интернет-информации КНР опубликовала Временные меры по управлению сервисами генеративного искусственного интеллекта, определяющие правила публичного использования ИИ для создания текста, аудиовизуального контента на территории Китая[8]. В них также впервые отмечается необходимость маркировки информации. Отметим, что в соответствии с действующим в КНР подходом в документе излагаются обобщенные принципы работы механизма маркирования контента, но не вводятся конкретные алгоритмы осуществления данного процесса[9]. Предполагается, что они будут уточнены в нормативных документах отраслевого характера. Так, в статье 8 отмечается, что разработчик сервиса генеративного ИИ должен применять соответствующую национальному законодательству маркировку создаваемого контента. При этом данный процесс должен отвечать принципам ясности, конкретности, практичности (практической реализуемости). Ответственное лицо также должно осуществлять оценку качества соответствующей маркировки, ее точности.

В сентябре 2024 г. Государственная канцелярия по делам интернет-информации КНР разместила для публичного обсуждения проект Мер по идентификации контента, сгенерированного с помощью генеративного ИИ. Указанный документ раскрывает обозначенные во Временных мерах по управлению сервисами генеративного искусственного интеллекта практические механизмы реализации маркировки информации. Несмотря на нахождение указанного проекта на стадии рассмотрения, ожидается, что большинство его положений будут закреплены в итоговом нормативном акте.

Аналогично предыдущему документу, данный проект включает значительное число положений, напрямую или косвенно затрагивающих вопросы безопасного использования сервисов генеративного ИИ[10]. При этом указанные подходы распространяются только на публичные сервисы генеративного ИИ, к которым имеют доступ пользователи интернет-пространства. В этой связи прослеживается его роль — дополнения действующего нормативного документа.

Согласно статье 3, информация, сгенерированная ИИ, включая текст, аудио, изображения, видео и другие данные, должны помечаться как материалы, созданные ИИ. Для целей маркировки вводят явную и скрытую маркировки информации. Явный тип маркировки — заметные для пользователя элементы, которые могут включать надписи, аудио и графические элементы и т.д., встраиваемые в сам контент или в интерфейс взаимодействия с пользователем. Скрытые метки внедрены в данные и не могут быть явно выделены пользователем.

---

[8] 生成式人工智能服务管理暂行办法 // 中央网络安全和信息化委员会办公室 Государственная канцелярия по делам интернет-информации КНР. URL: https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm (дата обращения: 03.10.2024).

[9] *Трощинский П. В.* Правовая система Китая. М. : ИДВ РАН, 2016.

[10] *Брюховецкий И. В.* Об актуальных вопросах регулирования генеративного искусственного интеллекта в КНР // Информационное право. 2023. № 3 (77). С. 28—31.

Отдельно подчеркивается важность маркировки контента, который впоследствии применяется на публичных информационных ресурсах, в СМИ и т.д. и потенциально способен ввести интернет-аудиторию в заблуждение (ст. 4). Подобный вид информации должен помечаться как явный тип маркировки, т.е. изображения должны иметь заметные метки, а видео, аудио, текст, виртуальный контент должны содержать метки в начале, середине и конце.

Интересен подход по внедрению в контент скрытых идентификаторов, позволяющих отследить как источник информации, так и конкретный индекс сгенерированной информации (ст. 5). В случае необходимости использования контента без меток на предоставляющих доступ к нему операторов сервисов возлагаются повышенные требования по протоколированию действий с информацией и по хранению соответствующих служебных данных не менее полугода (ст. 9).

Дополнительно отмечается недопустимость неправомерного использования меток или предоставления услуг по маркированию информации для последующего ее использования в незаконных целях (ст. 10). Конкретные алгоритмы маркирования сгенерированного контента должны быть определены в технических регламентах и стандартах. За несоблюдение соответствующих требований по маркированию контента вводится административная ответственность.

Механизмы маркирования информации

В отношении конкретных технических способов реализации маркировки материалов, создаваемых Национальным техническим комитетом по стандартизации в области информационной безопасности КНР, в августе 2023 г. подготовлен технический регламент — Методика маркирования создаваемого генеративным ИИ контента[11]. В нем описывается конкретный алгоритм осуществления маркировки различного вида материалов. Согласно его положениям, к явной маркировке относят нанесение поверх созданного изображения видимого глазу полупрозрачного текста, содержащего сведения о том, что информация сгенерирована ИИ. Надпись должна занимать не менее 0,3 % общей площади изображения при высоте текста не менее 20 пикселей. Уровень прозрачности подбирается с учетом возможности нормального использования изображения и, например, может составлять 90 %. Дополнительно, в целях идентификации контента в его файл вносятся сведения о провайдере услуг, времени создания, идентификаторе контента.

В случае скрытой маркировки минимальными условиями являются размещение названия оператора сервиса и идентификационного номера контента. Скрытая маркировка осуществляется путем замещения или подмены области данных и наносится на изображения размером от 384 на 384 пикселей. Она должна покрывать более 50 % площади и содержать соответствующую идентифицирующую информацию. Подобная маркировка наносится на видеоматериалы длительностью от 5 секунд и на аудиоматериалы — длительностью от 10 секунд.

[11] TC260-PG-20233A. 网络安全标准实践指南 —生成式人工智能服务内容标识方法 [Текст] / National Technical Commitee 260 on Cybersecurity of Standartization Administration of China. Beijing : TC260, 2023. URL: https://www.tc260.org.cn/ (дата обращения: 03.10.2024).

Заключение

С учетом проведенного анализа возможно заключить, что проблема выделения синтетического или созданного генеративным искусственным интеллектом контента имеет важное значение для дальнейшего развития информационного пространства. Соответствующие данные требуется маркировать не только для предотвращения фактов их незаконного использования, но также для минимизации искажения достоверности существующих в сети данных. В КНР уделяется особое внимание указанному вопросу. Имеются механизмы маркирования данного вида материалов, активно развивается процесс совершенствования профильной нормативной базы. В этой связи отслеживание созданной генеративными сервисами информации в сетевом пространстве Китая может быть реализовано на практике.

БИБЛИОГРАФИЯ

1. *Амелин Р. В., Чаннов С. Е.* Эволюция права под воздействием цифровых технологий : монография. — М. : Норма, 2023. — 280 с.
2. *Брюховецкий И. В.* Об актуальных вопросах регулирования генеративного искусственного интеллекта в КНР // Информационное право. — 2023. — № 3 (77). — С. 28—31.
3. *Ефремова М. А., Русскевич Е. А.* Дипфейк (deepfake) и уголовный закон // Вестник Казанского юридического института МВД России. — 2024. — № 2 (56).
4. Механизмы и модели регулирования цифровых технологий / под ред. А. В. Минбалеева. — М. : Проспект, 2023. — 264 с.
5. *Полякова Т. А., Минбалеев А. В.*, *Кроткова Н. В.* Основные тенденции и проблемы развития науки информационного права // Государство и право. — 2022. — № 9. — С. 94—104.
6. *Трощинский П. В.* Правовая система Китая. — М. : Институт Дальнего Востока РАН, 2016. — 471 с.
7. Цифровое право : учебник / под ред. В. В. Блажеева, М. А. Егоровой. — М. : Проспект, 2020. — 640 с.