Как распознать контент, созданный искусственным интеллектом или измененный в цифровом формате?

Доступные и простые в использовании модели ИИ могут помочь людям учиться и создавать контент, но при этом, они также увеличивают риски, **связанные с дезинформацией**, создающей угрозы открытому обществу и демократическому дискурсу. Важно, чтобы наши общие информационные пространства не были перегружены дезинформацией, в том числе, созданной ИИ и измененной в цифровом формате.

Частью решения являются новые технологии, такие как программы обнаружения и установления происхождения контента. **Но технологические решения далеки от совершенства, поэтому независимые фактчекеры ставят перед собой задачу предоставить общественности общий набор способов проверки достоверности фактов.**

Вот краткий обзор того, что делают профессиональные независимые специалисты по выявлению и проверке дезинформации, в том числе генерируемой искусственным ИИ.

Контента созданный ИИ становится все больше

В современном мире дезинформация, генерируемая ИИ, составляет небольшой процент утверждений, проверенных профессиональными независимыми экспертами. Фактчекеры гораздо чаще сталкиваются в своей работе с контентом, измененным в цифровом формате.

Согласно внутреннему опросу членов EFCSN, **большинство специалистов по проверке фактов согласны с тем, что актуальность контента, созданного с помощью ИИ и измененного в цифровом виде, в будущем будет только возрастать.** Как показывают недавние примеры в контексте европейских выборов, этот прогноз верен.

Примечание: Под цифровым изменением понимается любой тип контента, который был значительно изменен с целью манипулирования общественным мнением.

Сгенерированный искусственным интеллектом относится к любой форме контента, создаваемого системой искусственного интеллекта.

Информация, предоставленная **AI@EUElections**, которая финансируется Meta и управляется совместно с членами **Европейской сети стандартов для проверки фактов** и внешними экспертами.

Технологии быстро развиваются, но мы не можем полагаться только на них.

Рецензенты сходятся во мнении: **технических инструментов обнаружения ИИ недостаточно для выявления дезинформации, созданной с его использованием.**

Прежде чем использовать их, эксперты рекомендуют ознакомиться с детекторами ИИ. Изучая статистику и понимая, как работают модели, специалисты по проверке фактов могут определить сильные и слабые стороны инструмента, а также вероятность успеха при выполнении задачи. **При этом, несомненно, такого рода инструменты могут быть полезной отправной точкой.**

Программы по выявлению происхождению контента такие, как спецификации C2PA, могут помочь проверить происхождение и историю контента, но даже в этом случае проверка не является полностью надежной.

Как дезинформация, создаваемая ИИ, влияет на людей?

- *Каждый раз, когда вы реагируете, если можно так выразиться, своим сердцем, это подавляет ваше отражение – Кристин Дюгойн**

Психология: Операции влияния часто проводятся с целью использования психологических предубеждений.

Знание себя и своей аудитории может помочь вам бороться с дезинформацией.

Цели: Почему злоумышленники могут полагаться на искусственный интеллект для создания или распространения дезинформации? Какое влияние они хотят оказать в реальном мире?

- ➔ Расширить свое влияние на другие страны или сообщества?
- ➔ Избежать обнаружения фактчекерами, создав несколько вариантов похожих утверждений?
- ➔ Повлиять на мнения и убеждения, завоевав доверие с помощью сети недостоверных аккаунтов?

**Кристин Дюгойн исследователь информационного воздействия в В университете Сорбонне.*

Проверка требует многогранного подхода и знания деталей.

Итак, если средства обнаружения не работают, что тогда? Важно понимать как контекст утверждения, так и его содержание. Профессиональные фактчекеры — это эксперты, обладающие необходимыми исследовательскими навыками. Вот несколько советов:

*«Инструменты обнаружения никогда не будут работать со 100% точностью — я и не ожидаю, что они когда-нибудь будут работать» - Хэнк Ван Эсс***

Исследуем источники: можете ли вы подтвердить их личности? О чем они говорят и чем делятся? Кто взаимодействует с их контентом? Какое влияние этот контент может оказать на читателей?

Обеспечьте доверие: проверяйте информацию с помощью надежных источников, например, экспертов с практическим опытом в этой области.

Наряду с традиционными и документальными расследованиями используйте **методы криминалистики.** Некоторые из них включают в себя: очистку данных, геолокацию, биометрическое распознавание, анализ шаблонов и многое другое.

Учитесь и адаптируйтесь: создатели дезинформации и дезинформации, генерируемых искусственным интеллектом, постоянно адаптируются. Измените свой подход в соответствии с меняющейся средой.

Поделитесь своей работой

Наряду с проверенными утверждениями эксперты рекомендуют предоставлять прозрачный анализ и ссылки на источники. Это может помочь читателю лучше понять содержание исследования и его детали. В некоторых случаях исследования важнее, чем контент, написанный ИИ.

***Хэнк Ван Эсс эксперт в области OSINT и методов проверки фактов*

Краткое руководство: что можно и чего нельзя делать

Ниже приведены признаки, которые могут указывать на то, что часть контента создана искусственно или изменена в цифровом виде. Наряду с другими советами, упомянутыми в этом путеводителе (контекст, методы расследования и средства обнаружения), они помогут вам понять правду, стоящую за тем, что вы видите.

Текст

- Часто (но не всегда) **грамматика у него лучше**, чем у людей.
- • В большей степени используется **слишком формальный или структурированный язык**, особенно в контексте социальных сетей.
- **Злоупотребление глаголами или прилагательными.**
- Отсутствие человеческих эмоций, юмора, сарказма и идиоматических выражений.
- • Может **не хватать конкретных деталей** (имен, дат, мест) или оригинальных идей.
- Самое важное: верны ли факты, изложенные в тексте?

Видео

- Не используйте детектор изображений, созданный искусственным интеллектом, для видеозаписей.
- Обратите внимание на **выражение лица и движения**, например моргание глаз, а также на то, соответствуют ли движения человека звуку.
- Могут быть **отмечены резкие переходы или разрезы.**

Аудио

- **Сравните подозрительный звук с подлинным образцом,** используя инструменты, которые могут обнаружить различия в речи и характере дыхания, интонации...
- При использовании детекторов избегайте использования аудиосэмплов низкого качества со статическим или фоновым шумом.
- Могут присутствовать **неестественные или механические речевые модели** с паузами и отсутствием естественного дыхания.

Изображений

- Ищите **неестественные детали:** слишком идеальную кожу, размытый фон, неестественную красоту или свет, а также странности, такие как дополнительные пальцы.
- **Ищите водяные знаки** часто используемые генераторами изображений.
- Обратите внимание на то, что изображено: имеет ли это смысл? Насколько оно уместно?
- При использовании детекторов **выбирайте изображение с высоким разрешением или исходную загруженную версию изображения вместо изображения, которым поделились несколько раз.**