CIVIC RESILIENCE INITIATIVE

Disinformation Tool-kit

# Прорыв в области искусственного интеллекта: Руководство по борьбе с дезинформацией

Спонсором этой публикации является компания Google

# Подробнее о нас:

www.cri.lt

CIVIC RESILIENCE INITIATIVE

@CivicResilience

@CRI

@civicresilienceinitiative

Это руководство разработано в сотрудничестве с одними из самых известных организаций в области противодействия дезинформации: Baltic Security Foundation (Латвия) и Центром информации о безопасности и обороне (kaitsen.ee) (Эстония).

Спонсором этой публикации является компания Google.

# Прорыв в области искусственного интеллекта: Руководство по борьбе с дезинформацией

За последние десять лет распространение цифровых медиа продолжает расти. Каждый день мы подвергаемся воздействию все большего количества информационных потоков, поступающих из самых разных каналов коммуникации: социальных сетей, блогов, веб-сайтов, традиционных СМИ или других электронных изданий.

Такое разнообразие информационных потоков позволяет легко выбрать источник, который лучше всего отражает наши интересы, политические или социальные взгляды. По мере роста вовлеченности в социальные сети и времени, проведенного в них, все больше людей выбирают их в качестве основного источника информации, зачастую не осознавая угроз, которые в них кроются. Но быстрое и удобное распространение информации в социальных сетях создает идеальные условия также и для быстрого распространения дезинформации.

Современная популярность искусственного интеллекта (ИИ) и инструментов, которые на него опираются, предоставляет широкие возможности для распространения дезинформации и другой вредоносной информации. Видео- и аудиоподделки с использованием ИИ могут нанести значительный ущерб обществу, распространяя фальшивые новости и подрывая доверие граждан к СМИ. Однако вместе с огромным потенциалом для

СМИ. Однако вместе с огромным потенциалом для вредоносных провокаций появляется и равный потенциал для противодействия им. Когда ИИ будет полностью освоена, изобретательность этой технологии можно будет использовать и в борьбе с дезинформацией. Страны Балтии – постоянная мишень для ложной информации и злонамеренных действий, которые активно стремятся подорвать доверие граждан.

Такие усилия подпитывают недоверие к местным властям, нашим партнерам по НАТО и Европейскому союзу и постоянно пытаются демотивировать граждан к активному участию в социальной жизни. Хотя профессионалы проделывают огромную работу по развенчанию ложных историй, зачастую ущерб уже нанесен, когда дезинформация уже распространена. Чтобы просветить общественность в бесконечной борьбе с дезинформацией, в 2019 году в Литве создали основали Civic Resilience Initiative (CRI). В сотрудничестве с экспертами в области дезинформации и СМИ эта организация разработала практический обзор «Прорыв в области ИИ: Руководство по борьбе с дезинформацией». Это пособие по выявлению дезинформации предлагает простое руководство по методам, необходимым для проверки информации. Мы надеемся, что эта публикация поможет вам легко выработать привычку проверять подлинность источников новостей, которые вы читаете, а также фотографий и видео, которые их сопровождают.

**Команда CRI поставила перед собой цель стать главным катализатором укрепления цифровой устойчивости общества в странах Балтии.**


Команда CRI

**Это практическое руководство по дезинформации даст вам основные инструменты, которые помогут:**

- проверить, является ли информация в Интернете реальной или фальшивой;
- понять, как работает Интернет вещей (IoT), и распознать изображения, которые были обработаны искусственным интеллектом;
- выявлять троллей и поддельные аккаунты в социальных сетях;
- распознавать поддельные изображения и видео в Интернете;
- принимать меры при обнаружении дезинформации;
- предупреждать других о ложных сообщениях, распространяемых в социальных сетях.

Данное пособие призвано способствовать повышению устойчивости к цифровым манипуляциям, осведомленности о безопасности и необходимости проявлять бдительность в информационном пространстве. **Цель этой публикации – предоставить информацию, которая поможет школьникам, студентам и широкой общественности повысить свою цифровую грамотность и устойчивость к ложной информации.**

Это руководство разработано в сотрудничестве с одними из самых известных организаций в области противодействия дезинформации: Baltic Security Foundation (Латвия) и Центром информации о безопасности и обороне (kaitsen.ee) (Эстония).

*Спонсором этой публикации является компания Google.*

# Существует множество форм фальшивых новостей:

## Дезинформация

– информация, **которая является ложной и намеренно создается с целью навредить человеку,** социальной группе, организации или стране.

## Неправдивая информация

– информация, **которая является ложной, но не создается с намерением причинить вред.**

## Вредоносная информация

– информация, **основанная на реальности, используемая для нанесения вреда человеку,** организации или стране.

Все эти виды опасны тем, что распространяются широко и быстро, потому что становятся вирусными, когда многие люди и даже организации не задумываясь репостят эти истории, поскольку они кажутся интересными и сенсационными.

# Идентификация.

# Как проверить новости или посты?

**В случаях, когда тема скандальная или она звучит невероятно, стоит уделить минуту и провести простую проверку, чтобы убедиться, что информация верна и заслуживает доверия.**

## Как это сделать?

Вот пять простых шагов, которые помогут вам проверить информацию:

| | |
|---|---|
| **1. Оцените источник** | Изучите веб-сайт или аккаунт в социальной сети. Подумайте, кто может стоять за распространением новости и какова цель ее создания. |
| **2. Читайте не только заголовок** | Заголовки статей могут быть скандальными, чтобы привлечь клики и способствовать распространению информации. Если вникнуть в суть истории, может оказаться, что утверждения в заголовке не соответствуют действительности. |
| **3. Проверьте автора** | Существует ли автор на самом деле? Является ли автор благонадежным человеком? |
| **4. Подтверждают ли другие источники эту историю?** | Часто в фальшивых новостях нет ссылок, по которым можно проверить факты. Если в статье есть ссылки на источники, то перейдите по ним. Может выясниться, что исходное сообщение было приукрашено или его смысл был искажен. |
| **5. Проверьте дату** | Повторная публикация старых новостей не означает, что они все еще актуальны. |

# Дополнительные советы:

## Выбирайте безопасную и надежную информацию.

Для получения новостей используйте крупные порталы, такие как ERR, DELFI, TV3 и подобные. На небольшие порталы легче повлиять, чтобы они купили определенный контент, поскольку у них меньше человеческих и финансовых ресурсов, меньше журналистов, которые могут проверить достоверность информации, и они могут быть более уязвимы к кибератакам и взломам.

**Тщательно фильтруйте информацию, поступающую с иностранных порталов или из групп социальных сетей:**

• У ваших источников должна быть достаточная аудитория. В противном случае посмотрите, кто еще опубликовал ту же информацию, и снова отфильтруйте этих людей.

• Если вы собираете информацию на платформах X (ранее известной как Twitter), TikTok или YouTube, обратите внимание на комментарии, которые должны быть включены. Более надежные записи будут иметь более одного комментария, с большими промежутками между публикациями, и не будут связаны шаблонами стиля написания. В противном случае комментарии могут быть сделаны троллями или ботами.

• Источники не должны иметь ссылок на пророссийские правительственные порталы (Sputnik, Первый Канал, «Россия 24» и другие). Хотя в новостях, публикуемых этими порталами, может быть доля правды, в условиях информационной войны не стоит рисковать, веря пророссийским источникам новостей, если они не подтверждены другими надежными порталами в странах Балтии и наших зарубежных партнеров.

• Сенсационная информация всегда должна основываться на надежных источниках. Если вы видите в Интернете срочные новости, убедитесь, что источник, сообщивший о них, также указывает свои источники. Если это не так, проверьте это при помощи Google:
1) Выделите кавычками наиболее важные ключевые слова из читаемого поста или статьи и введите их в поиск Google.
2) Нажмите на кнопку «Инструменты» в правой части поисковой строки и выберите “показать последние”.
3) Снова проверьте источники по вышеуказанным критериям.

• Спросите себя, действительно ли этому источнику не хватает доказательств и анализа. С какой точки зрения подаются новости? Может ли быть так, что события, о которых сообщается, не соответствуют действительности и что это попытка манипулировать воображением и эмоциями читателей?

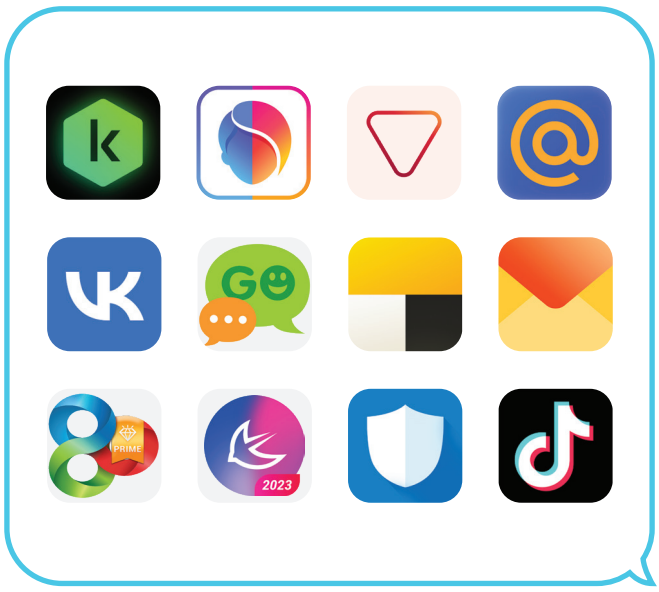
Наконец, увидев сенсационную новость на сайте одного из государственных учреждений, зайдите на авторитетный новостной портал. Одним из инструментов информационной войны является взлом правительственных сайтов и распространение на них недостоверной информации. Если государственные учреждения взломаны, СМИ обязательно сообщат об этом. Но мы не должны верить информации, которую они публикуют. В некоторых случаях взлом не требуется, и можно использовать поддельные адреса сайтов. Такие трюки трудно распознать, поэтому всегда проверяйте, действительно ли адресная строка выглядит так, как должна.

## Предлагайте надежные источники информации не только своим родителям, но и бабушкам и дедушкам.

Смотрят ли члены вашей семьи российское телевидение? Информация о войне – одна из самых популярных и обсуждаемых тем на российских телеканалах в последние несколько лет. Понятно, что для многих наших родителей, бабушек и дедушек русский язык является основным и предпочтительным языком. Поэтому мы советуем членам вашей семьи установить Netflix или другую платформу, которая включает в себя контент с русским переводом. Поскольку российское телевидение рассчитано на то, чтобы сделать зрителя равнодушным к информации, переход на платформу домашнего кинотеатра не должен быть сложным.

## Удалите российские и китайские приложения.

Удалите со своих смартфонов все приложения, созданные в России или ее потенциальными союзниками, такими как Китай. Некоторые из них имеют различные доступы к вашим данным или местоположению, что может быть использовано против вас как для распространения дезинформации, так и в военных действиях. К наиболее популярным приложениям относятся: Антивирус Касперского, CheckScan, FaceApp, MyPocket, Mail.ru, Вконтакте, Go SMS Pro, Яндекс Такси, Яндекс Почта, Go Launcher, APUS Launcher, Security Master и даже TikTok.



# Искусственный интеллект (ИИ)

ИИ связан с моделями искусственного интеллекта, которые могут генерировать разный контент, например изображения, текст, музыку и т. д. Примерами ИИ являются такие известные модели, как ChatGPT, DALL-E и другие, в которых мы можем легко вводить текстовые запросы и получать в ответ сгенерированный текст или изображения.

Большие языковые модели — это один из видов ИИ, который может понимать, распознавать, контекстуализировать и генерировать текст. Чтобы эти модели могли выполнять свои задачи, их необходимо обучать, а для этого требуется значительный объем текстовых данных. Как правило, данные получают путем чтения текстов из общедоступного интернета и преобразования их для целей обучения.

Языковые модели часто используются для ответов на вопросы, на которые мы не знаем ответа, или для создание письменного текста, когда у нас нет времени на подготовку, или даже для написания программного кода, когда мы не владеем языками программирования. Это отличный инструмент для получения информации или создания контента, для чего в противном случае потребовалось бы значительное количество времени и усилий.

Однако модели ИИ используются не только в положительных целях, но и позволяют злоумышленникам проводить вредоносные кибератаки гораздо эффективнее.

### 1. Социальная инженерия

Модели ИИ могут создать отличный текст на любом языке, и этот текст будет более точным, чем перевод, выполненный с помощью программы «Google Translate». Текст будет высокого качества, с небольшим количеством ошибок или же вообще без них. Таким образом, киберпреступник может не знать языка выбранной им жертвы, но все равно сможет написать текст, который, например, попытается обманом заставить жертву отдать деньги или перейти по ссылке, чтобы вынудить ее пре-

доставить данные для входа в систему. Отныне вредоносные электронные письма, сообщения и т. д. будут более качественными, и их будет сложнее распознать как мошеннические. Существуют специальные языковые модели, намеренно разработанные для создания таких обманчивых сообщений.

### 2. Распространение дезинформации

Ответы, генерируемые программами ИИ, не всегда точны и могут не отражать реальность, что приводит к введению пользователей в заблуждение, а полученный неточный или неправильный ответ может начать распространяться в социальных кругах пользователя. Кроме того, модели ИИ учатся на вопросах пользователей и их уточнениях, поэтому модель может представить неверные данные другим как точные, способствуя таким образом распространению дезинформации.

## Искусственный интеллект (ИИ)

# Как распознать текст, сгенерированный большими языковыми моделями ИИ?

В настоящее время не существует полноценного инструмента, который мог бы однозначно определить, написан ли текст искусственным интеллектом или нет. Самым точным инструментом распознавания на данный момент является человеческий мозг. **Например,** приведенный ниже текст написан с помощью ChatGPT. Сможете ли вы найти типичные ошибки там?

Дорогой пользователь!

Я надеюсь, что это письмо застало вас в добром здравии. Я пишу, чтобы обратить ваше внимание на важность завершения процесса регистрации и входа в систему на нашей платформе.

В рамках нашего обязательства по обеспечению бесперебойной работы, мы просим вас завершить регистрацию или войти в свой аккаунт при первой возможности. Это не только обеспечит безопасность вашей учетной записи, но и предоставит вам доступ ко всему спектру функций и преимуществ, доступных на нашей платформе.

Для завершения процесса регистрации и входа в систему, пожалуйста, перейдите по ссылке, представленной ниже: Ссылка для входа

Если вы столкнетесь с какими-либо трудностями или у вас возникнут вопросы, пожалуйста, обращайтесь в нашу службу поддержки.

Спасибо за сотрудничество, и мы с нетерпением ждем возможности служить вам.

В этом коротком тексте есть несколько мест, которые выглядят подозрительно:
• Начало текста звучит неестественно, и в наше время редко можно получить письмо с таким началом. Обычно такое начало говорит о том, что текст сгенерирован искусственно.
• Кроме того, первый абзац написан от первого лица („я“), однако остальной текст написан от лица „мы“, что также выглядит довольно подозрительно.

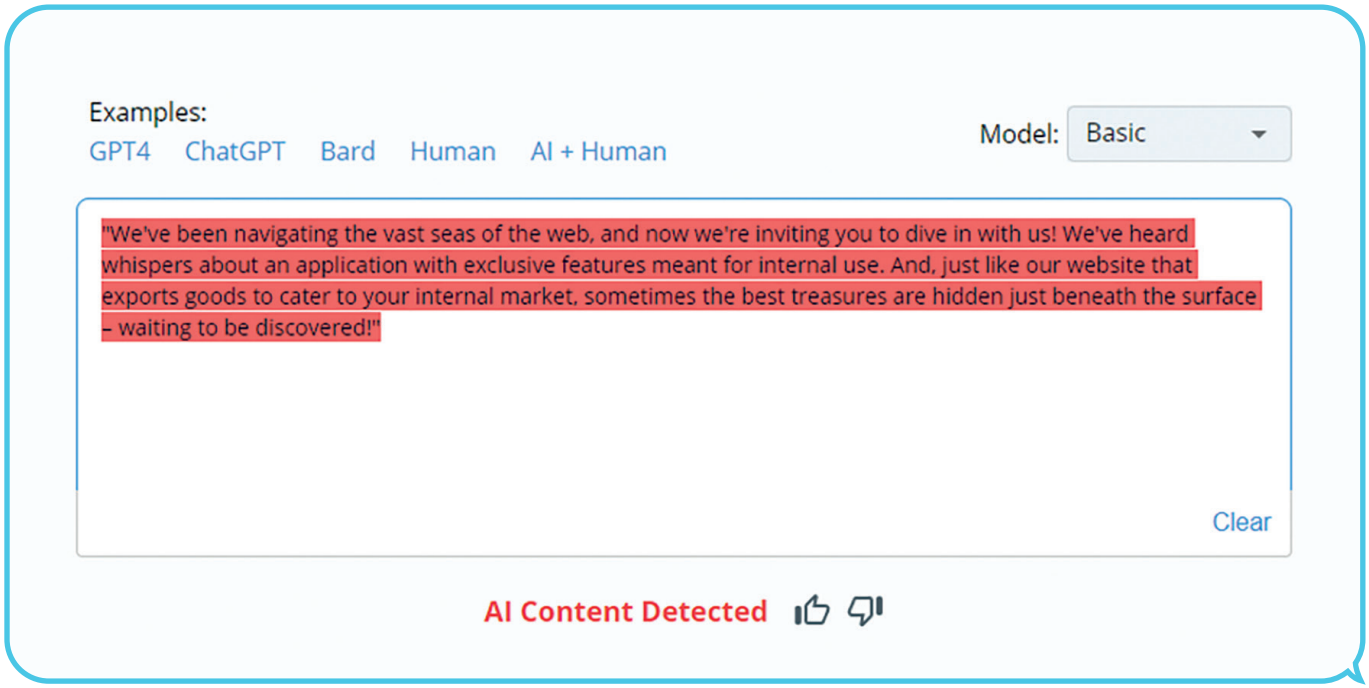
Итак, данный пример показывает, что искусственный интеллект при создании текста использует фразы, которые чаще всего встречаются в автоматически сгенерированном тексте, а использование местоимений вызывает вопросы. Другие подсказки могут быть связаны с использованием неправильных падежей или ошибками в структуре предложения. Кроме того, иногда стиль языка могут выдать работу искусственного интеллекта, даже если в тексте нет ошибок.

В Интернете есть несколько сайтов, которые пытаются определить, создан ли текст искусственным интеллектом или нет. Чтобы понять, как автоматизированные инструменты пытаются определить, был ли текст написан ИИ, нам нужно вернуться к основам того, как языковые модели генерируют текст.

В упрощенном смысле модели, генерирующие текст, пытаются предсказать следующее наиболее подходящее слово в предложении, что делает модели предсказуемыми. Чем больше данных было использовано для обучения модели, тем лучше ИИ угадывает следующее слово. Некоторые автоматизированные инструменты обнаружения пытаются определить, является ли последовательность слов в предложении статистически оптимальной, что указывает на то, что текст был написан ИИ. В то же время другие инструменты занимаются обратным процессом, рассчитывая вероятность того, что текст был написан человеком, анализируя его структуру.

Ниже приведено несколько примеров автоматизированных инструментов. В этих инструментах есть поле, куда можно вставить подозрительный текст. После нажатия кнопки анализа инструмент обычно выдает оценку вероятности того, насколько вероятно, что текст создан искусственным интеллектом. Некоторые инструменты выделяют в тексте места, которые, скорее всего, сгенерированы, а не написаны человеком.

**Copyleaks**



**Zerogpt**

**Your Text is AI/GPT Generated**

87.91%
AI GPT*

Dear [Username],

Welcome to [Your Website]! We're thrilled to have you as part of our community.

To complete your registration and unlock the full benefits of your account, please click on the following link:

[Insert Confirmation Link]

By confirming your registration, you'll gain access to exclusive features and updates. If you have any questions or need assistance, feel free to reach out to our support team at [Support Email].

Thank you for choosing [Your Website]! We look forward to providing you with a great experience.

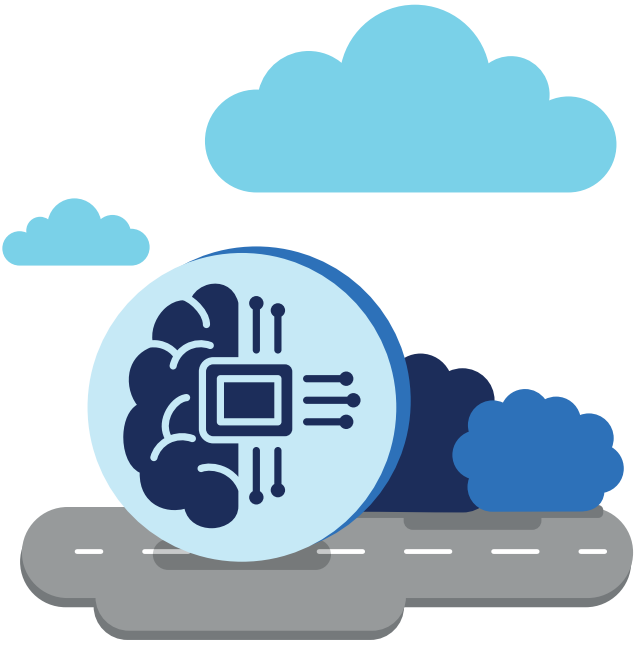
Best regards,
The [Your Website] Team

**Highlighted text is suspected to be most likely generated by AI***
**572 Characters**
**91 Words**

**Writer**

**Gptzero**

**Ilm**



## Проверка текстовой информации

# Как проверить, является ли текстовая информация подлинной?

### Google, Google, Google!

Проще говоря, поисковая система, такая как Google, - ваш лучший друг, когда речь идет о противодействии дезинформации. Всю информацию можно проверить, является ли она правдивой или ложной. Логика заключается в том, чтобы использовать ключевые слова, которые вы можете выделить из интересующей вас информации, и посмотреть, не говорят ли о ней какие-либо надежные источники.

Если вы видите что-то подозрительное или вызывающее эмоциональный отклик, велика вероятность, что кто-то уже пишет или говорит об этом. Большим преимуществом является то, что можно проверить информацию, которая не является цифровой, а также успешно идентифицировать ключевые слова.

Важно отметить, что это позволит найти надежный источник с тем же содержанием, которое вы пытаетесь проверить, но вы должны доверять этому источнику.

### Как проверить?

1. Определите ключевые слова, наилучшим образом описывающие интересующую вас информацию.
2. Воспользуйтесь одной из доступных поисковых систем для поиска той же информации.
3. Отфильтруйте источники по разделу «Новости» и выберите самые последние сообщения, опубликованные за последний час или день.
4. Определите надежные источники и проверьте информацию.

**Основные правила, которые нужно запомнить:**

- Поиск информации в Google - лучший первый шаг для проверки информации. Это очень быстро и эффективно;
- Неважно, какую информацию вы хотите проверить; текстовый поиск Google работает очень хорошо: текст, фотографии или видео;
- Главное - использовать ключевые слова, чтобы найти ту же информацию в надежном источнике.

**Для более эффективного поиска используйте основные функции Google, например:**

1. если вы ищете конкретную фразу, заключите ее в кавычки ( „...“ );
2. если в результатах поиска есть очень популярное ключевое слово, используйте минус ( - ) и это ключевое слово, чтобы исключить его из результатов поиска;
3. если вы не знаете точного написания слова, точного числа или даты, вы можете использовать * как подстановочный знак, чтобы заменить пропущенный символ в слове или все слово, число или дату.

• Если вы владеете другими языками, вы можете воспользоваться этим. Поищите, что пишут по этому вопросу другие источники на других языках.

• Если вы хорошо владеете только одним языком, попросите кого-то, кому вы доверяете, помочь вам проверить важную информацию на других языках. Обсудите с ними, как об этом пишется в других языковых пространствах. Для них это может быть так же ценно, как и для вас - в профессиональном плане.

**Полезные инструменты:**

**Google**

**Bing**

**Yandex**

*(ВАЖНО: Будьте осторожны, это российский сайт, поэтому используйте дополнительные меры безопасности на своем компьютере. Насколько известно, сайт безопасен для использования, но мы рекомендуем не пользоваться мобильным приложением, так как при установке оно запрашивает доступ к личным данным).*

# Проверка изображений

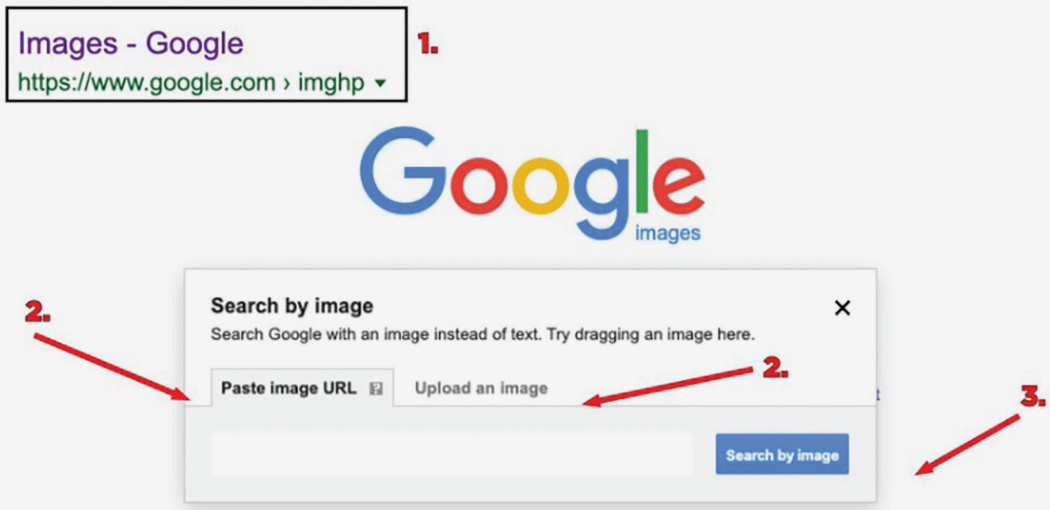
# Как проверить, является ли визуальная информация подлинной?

## Обратный поиск изображений:

Поскольку повторное использование изображений (например, размещение изображения, сделанного ранее, и утверждение, что оно было сделано недавно) остается одной из основных проблем в дезинформации, проверка того, не было ли изображение размещено ранее, является одной из лучших стратегий борьбы. В других случаях, если изображение, с которым вы работаете, было изменено, обратный поиск изображений поможет вам найти оригинальное изображение.

## Как проверить?

1. Откройте одну из поисковых систем (ссылка указана рядом с полезными инструментами);
2. Скопируйте ссылку или само загруженное изображение;
3. Проверьте, не размещались ли ранее такие же или очень похожие изображения.



## Анализ уровня ошибок

Анализ уровня ошибок – это более продвинутый метод, позволяющий выявить в изображении области с разным уровнем сжатия. В изображениях JPEG все изображение должно иметь примерно одинаковый уровень сжатия. Если на каком-то участке изображения уровень ошибок значительно отличается, то это, скорее всего, свидетельствует о цифровой модификации. На практике вам следует осмотреть изображение и определить различные высококонтрастные края, низкоконтрастные края, поверхности и текстуры. Сравните эти области с результатами анализа уровня ошибок. Если есть значительные различия, то это указывает на подозрительные области, которые могли быть подвергнуты цифровому изменению.

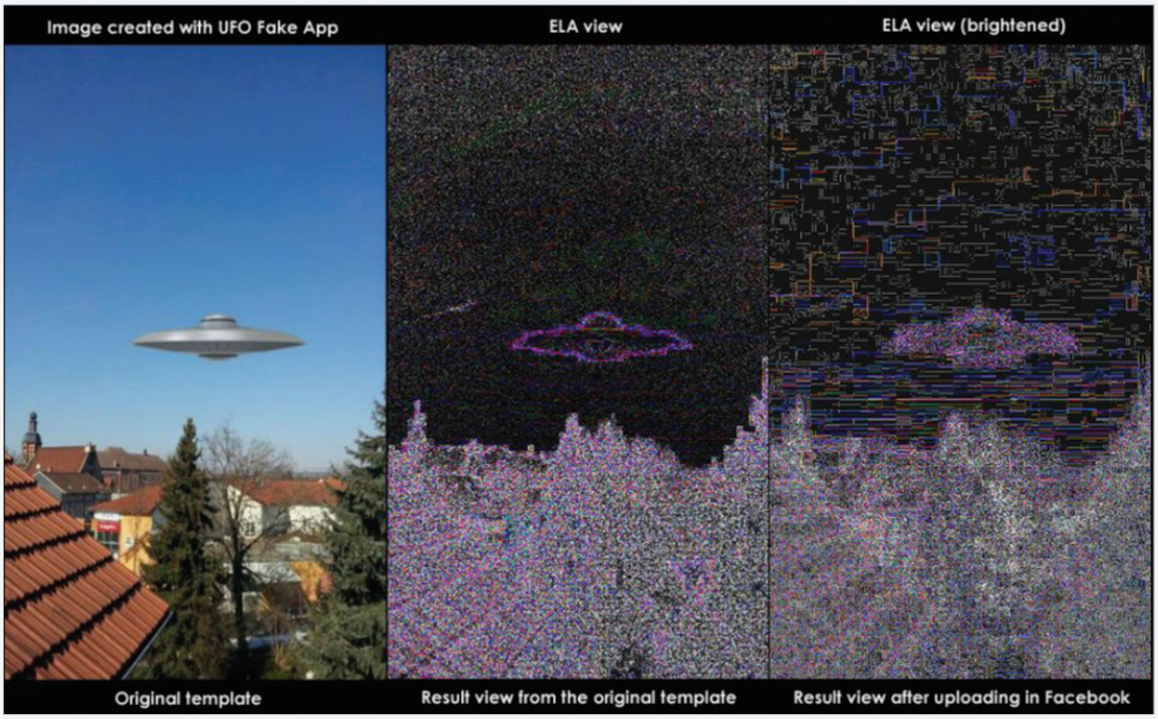
Важно отметить, что этот метод не является пуленепробиваемым, тем не менее это – надежный первый шаг к выявлению цифровых изменений, которые были внесены в изображение. Фотокриминалистика – это отдельная научная отрасль, и чтобы развенчать хорошо подделанные изображения, требуются многолетний опыт и навыки, однако если говорить о повседневных пропагандистских фейках, то обычно такие изображения не очень хорошо проработаны и легко идентифицируются.

## Основные правила, которые следует запомнить:

• Обратный поиск изображений должен быть стандартной практикой, прежде чем доверять изображению. Если изображение окажется не тем, за что себя выдает, это может отклонить подлинность сообщения;

• Анализ уровня ошибок (ELA) - не 100% надежный метод, но это отличный быстрый способ проверить на цифровые манипуляции с изображением.

• Обратный поиск изображений может быть использован для идентификации неизвестных людей на изображении.




## Полезные инструменты:

**Google** / Обратный поиск изображений

**Yandex** / Обратный поиск изображений

**Google** / Расширение „RevEye“

**Forensically**

**Foto Forensics**

## Искусственный интеллект (ИИ)

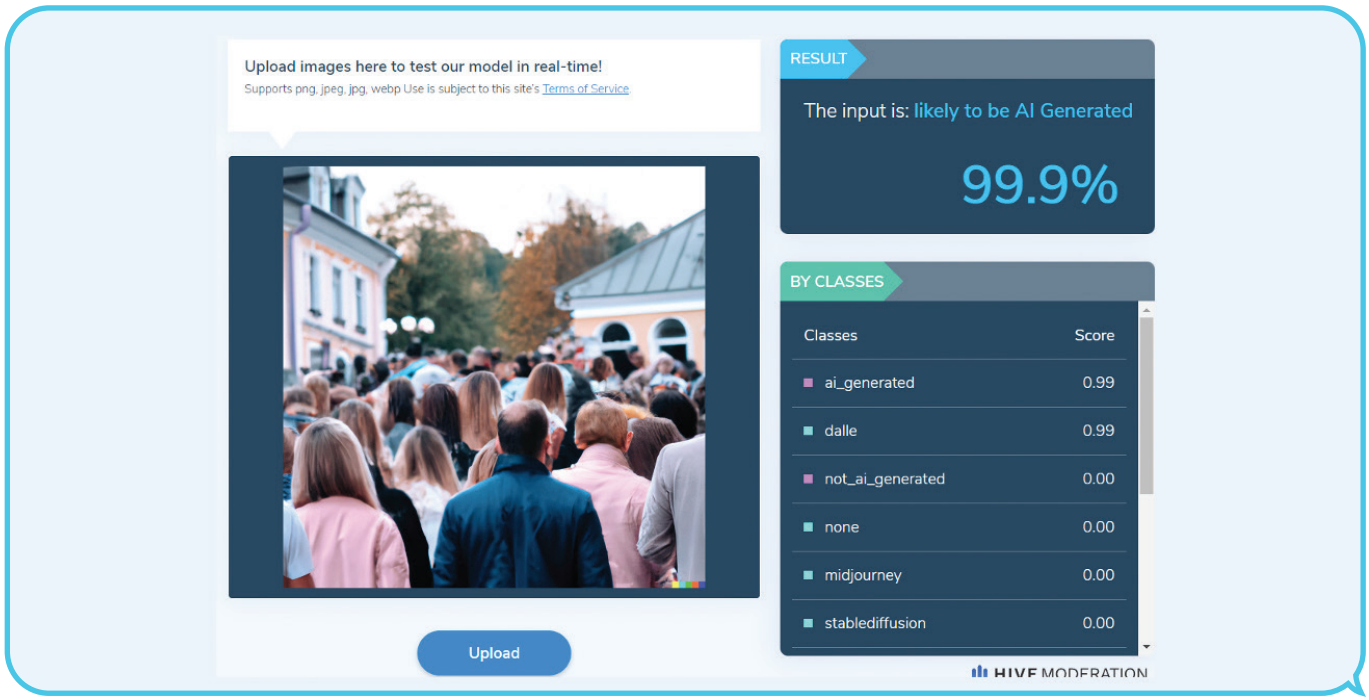
# Как определить сгенерированное ИИ изображение?

Идентифицировать созданное ИИ изображение, в отличие от сгенерированного текста, проще, особенно если ИИ пытается сгенерировать реалистичные фотографии. Ниже приведено несколько примеров. На фотографиях изображены люди в Старом городе Вильнюса. На первый взгляд кажется, что фотографии настоящие, но давайте посмотрим внимательнее.

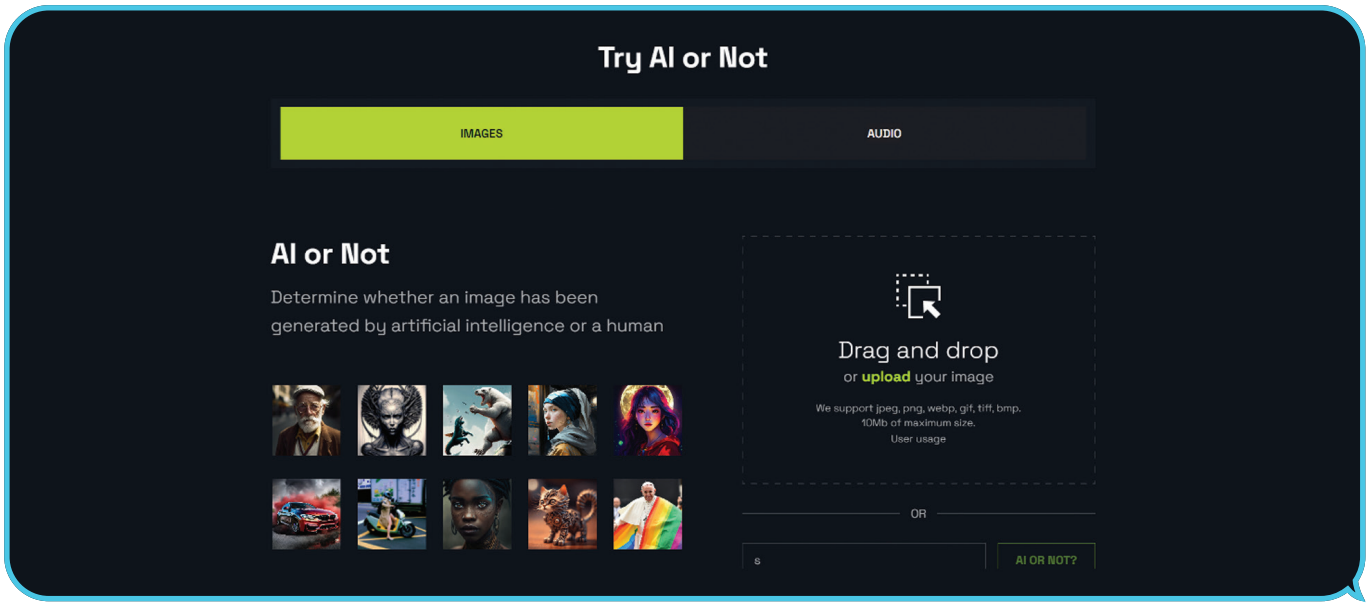
• Цвета - на левой фотографии цвета кажутся неестественными и слишком яркими, что напрягает глаз.
• Фон - часто сгенерированные фотографии имеют нечеткий фон, а линии зданий или автомобилей на заднем плане часто искажены.
• Аномалии и искажения - при увеличении фотографии вы сразу заметите, что некоторые части тела выглядят неестественно и искаженно: уши выглядят необычно, головы кажутся нарисованными кистью и так далее. Кроме того, ИИ часто добавляет более 5 пальцев или больше зубов, чем может быть у человека. Очень важно обращать внимание на детали.
• Водяные знаки – веб-инструменты, особенно бесплатные, часто добавляют на фотографии свои водяные знаки. В примерах выше вы можете заметить цветные квадратики в правом нижнем углу, указывающие на то, что изображения были созданы с помощью DALL-E. Конечно, мы можем вручную удалить этот водяной знак. Однако платформы, позволяющие создавать изображения, начали добавлять невидимые водяные знаки, которые не видны невооруженным глазом. Если загрузить фотографию в инструмент проверки, он сразу определит, что это сгенерированное изображение, поскольку выявит скрытый водяной знак.

Несколько автоматизированных инструментов могут помочь определить сгенерированные ИИ фотографии. Принцип использования этих инструментов стандартен. Просто перетащите или загрузите фотографию в определенный раздел страницы и нажмите кнопку „Проанализировать“. Через несколько секунд вы получите результат:

## Hivemoderation




## Aiornot



## Проверка видео

# Как проверить, является ли видео подлинным?

### Обратный поиск изображений

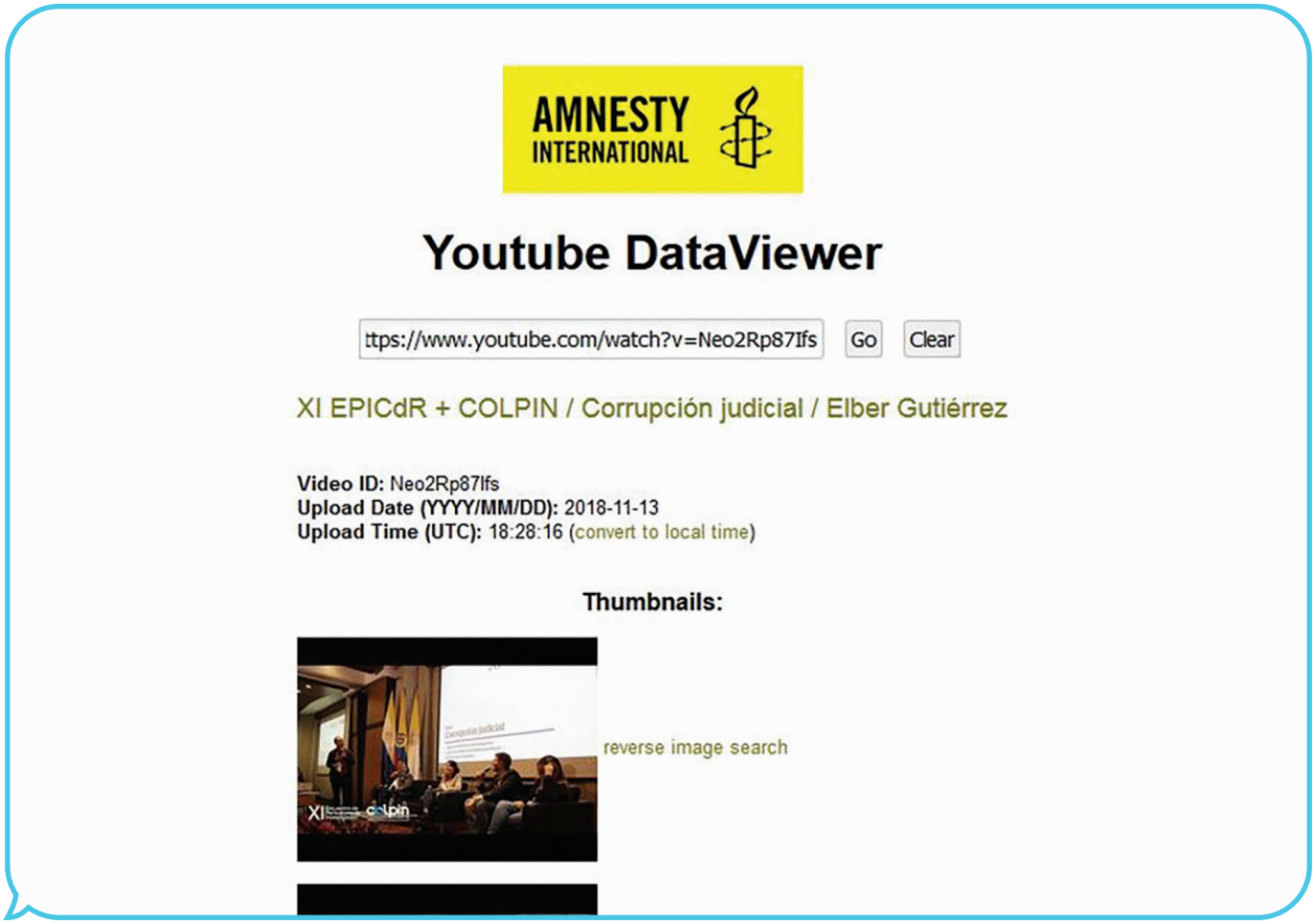
Когда речь заходит о поддельных видео, то, как и в случае с изображениями, лучшей методикой является обратный поиск изображений. Поскольку видео – это просто серия изображений, выделение одного кадра и его поиск - отличный способ сделать это. Оба инструмента - InVid и Amnesty Data Viewer - позволят вам найти похожие или идентичные видео, уже размещенные в сети, путем поиска как кадров, так и миниатюр.

### Как проверить?

1. Откройте одну из поисковых систем (например, Amnesty DataViewer или InVid);

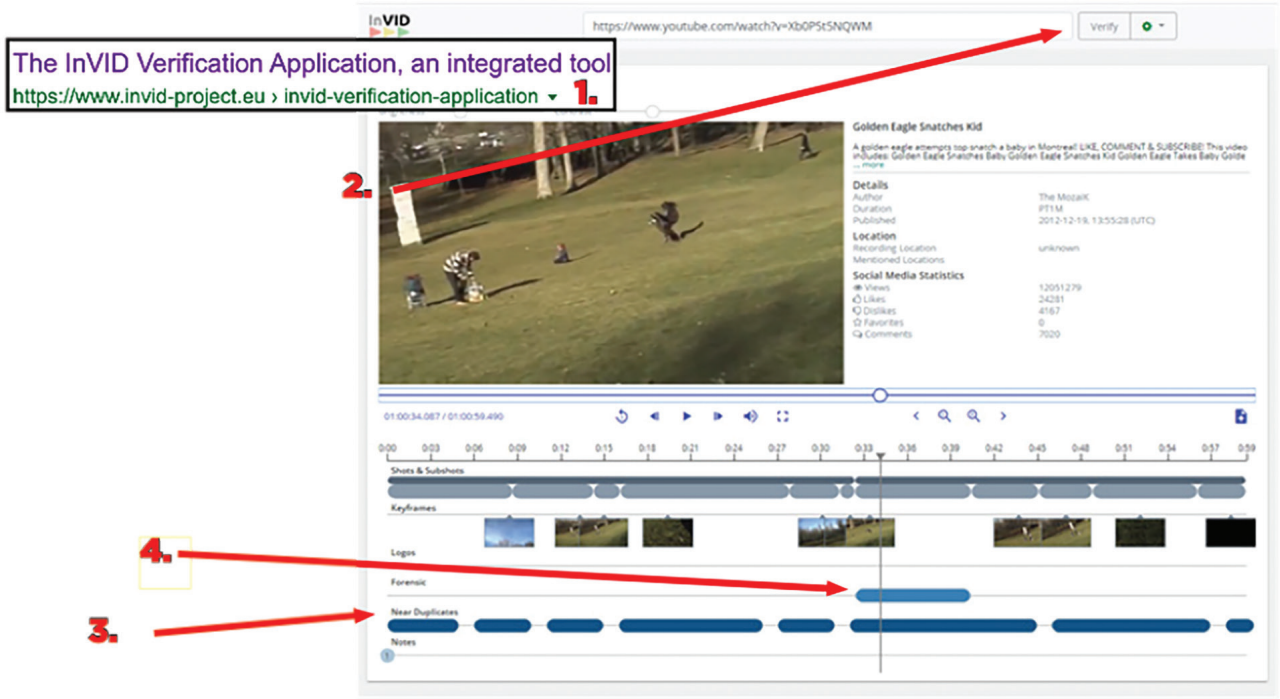
2. Вставьте ссылку на видео;

3. Проверьте, есть ли видео среди дубликатов.



## Углубленный анализ

В InVid также есть возможность углубленного анализа, когда программа выявляет потенциально измененные кадры. Потенциально измененные кадры появляются в окне анализа рядом с надписью „Forensic“. Если InVid определяет кадры как потенциально измененные, высока вероятность того, что видео поддельное.




## Основные правила, которые следует запомнить:

• Основная угроза при работе с видео та же, что и с изображениями, - переработка видео. Взятые ранее видеоролики переделываются и выдаются за фальшивые новости;

• Эти инструменты не так эффективны, как инструменты проверки изображений, но все же они могут выявить большое количество поддельных видео.

**Полезные инструменты:**

**InVid**

**Amnesty International YouTube Viewer**

Тролли

# Как распознать тролля в Интернете?

## Что такое тролль?

Тролль – это человек, который намеренно провоцирует онлайн-конфликт или оскорбляет других пользователей, чтобы отвлечь внимание и посеять раздор, публикуя подстрекательские или не относящиеся к теме сообщения в чатах, онлайн-сообществе или социальной сети. Их цель - намеренно спровоцировать других на эмоциональную реакцию и сорвать дискуссию. Тролль отличается от бота тем, что тролль – это реальный пользователь, в то время как боты автоматизированы. Эти два типа аккаунтов являются взаимоисключающими.

Распознать тролля сложнее, чем бота, так как эти аккаунты обычно более изощренные и активно притворяются обычными людьми. Ниже вы найдете ряд критериев, которые помогут вам определить тролля, но эти подсказки скорее ориентировочные, чем окончательные. Редко можно со стопроцентной уверенностью сказать, что тот или иной аккаунт принадлежит троллю, а не просто поддерживает распространение определенных злонамеренных нарративов. Прежде чем рассмотреть факторы, которые достоверно указывают на идеологического тролля, важно обратить внимание на один фактор, который таковым не является, - политический контент. Разнообразные современные реальные пользователи социальных сетей склонны к высокой ангажированности, особенно когда речь идет о политических темах.

Ниже приведены некоторые критерии, которые помогут вам выявить троллей, но учтите, что эти подсказки являются ориентировочными, а не заключительными. Редко можно быть уверенным на 100%, что аккаунт принадлежит троллю, а не простому злому пользователю.

**1. Ошибки в статьях на английском языке: A vs The**

Одним из лингвистических признаков, характерных для многих известных русских аккаунтов, является неумение правильно использовать грамматические артикли в английском - „a“ и „the“. В русском языке нет ни того, ни другого.

**2. Ошибки в формулировке вопроса**

Еще один распространенный лингвистический признак – неумение правильно сформулировать вопрос. В русском языке порядок слов в вопросах не меняется, в отличие от английского, немецкого и французского. Многие известные аккаунты российских троллей размещали вопросы, в которых сохранялся порядок слов, характерный для утверждений на русском.
Тем не менее не забывайте, что современные средства ИИ переводят тексты с все большей точностью, поэтому одного языкового фактора для обнаружения тролля недостаточно.

**3. Неясная или сомнительная личность**

Некоторые тролли используют фальшивые имена, которые очень распространены в данном языке, что затрудняет различение конкретного автора или намеренно заставляет перепутать его с другим, например, признанным журналистом. Имена, используемые троллями, также призваны восприниматься как традиционные или „правильно звучащие", чтобы любой читатель был склонен доверять такому автору статьи или комментария в социальных сетях. Также полезно проверить фотографии в профиле, если таковые имеются. Такие изображения, добавленные для достижения дополнительного доверия, могут быть стоковыми фотографиями, которые легко найти в интернете. Кроме того, такие фотографии могут быть намеренно нечеткими при ближайшем рассмотрении (фотомонтаж, предполагаемый человек в солнцезащитных очках и т. д.), что делает невозможным однозначную идентификацию личности.

**4. Усиление прокремлевских нарративов**

Российское правительство разработало особый нарратив в отношении ключевых геополитических событий последних лет. Это соответствует принципам, заложенным еще в Доктрине информационной безопасности Российской Федерации (2000 г.), по донесению государственной политики и официальной позиции по важным для российского правительства вопросам. Поскольку прокремлевские нарративы широко доступны в онлайн-источниках, таких как Министерство иностранных дел России или аккаунт RT в X (Twitter), легко проверить, появляются ли те же темы в подозреваемом аккаунте. Аккаунт, который неоднократно разделяет тезисы российского правительства по большинству или всем этим событиям, можно с полным основанием считать прокремлевским. Если аккаунт разделяет большинство или все кремлевские тезисы, допускает характерные языковые ошибки и выдает себя за американского или британского пользователя, это может быть тролль, управляемый из России.

## Другие потенциальные улики

**Тролли используют вымышленные адреса электронной почты:**

Поскольку в большинстве мест, где разрешено оставлять комментарии, требуется указать адрес электронной почты, тролли обходят эту просьбу, используя выдуманные адреса. Большинство выдуманных адресов электронной почты являются случайными, и их легко обнаружить, так как они не соответствуют настоящему имени человека.

**Задача троллей вывести людей из себя:**

Они не вежливы и не стесняются вступать в открытую борьбу. Они обзываются, выдвигают обвинения и редко говорят что-то кроме злости.

**Тролли используют анонимные прокси:**

Тролли часто используют анонимайзеры, или прокси, которые показывают другой адрес интернет-протокола (IP).

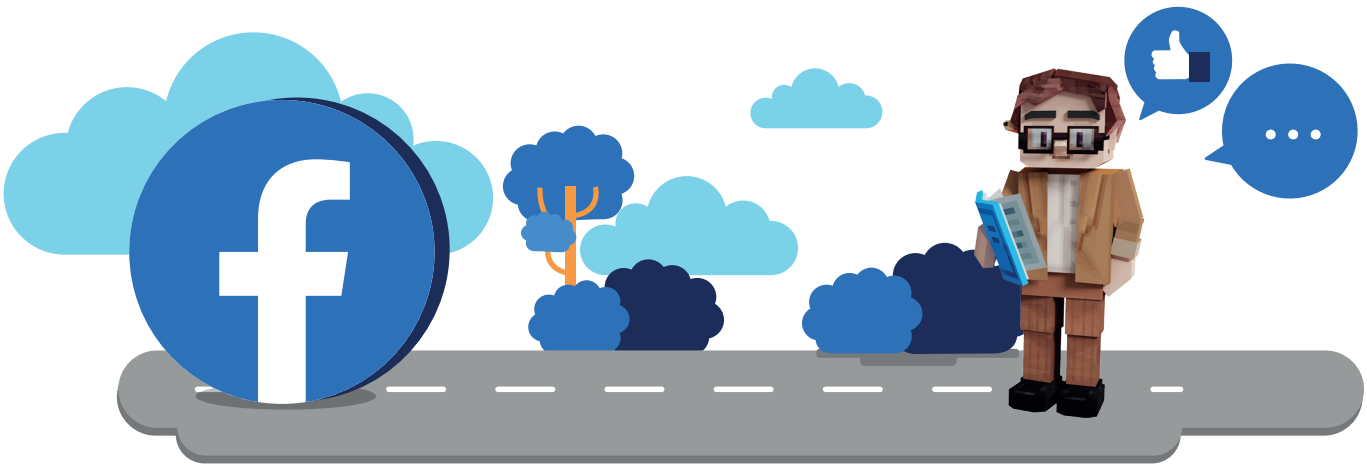
**Тролли редко добавляют что-то ценное к разговору:**

Когда тролли отвечают на обсуждение в чате или сообществе, они не добавляют ничего значимого к дискуссии. Вместо этого они шутят, ругают и оскорбляют.

## Поддельные аккаунты Facebook

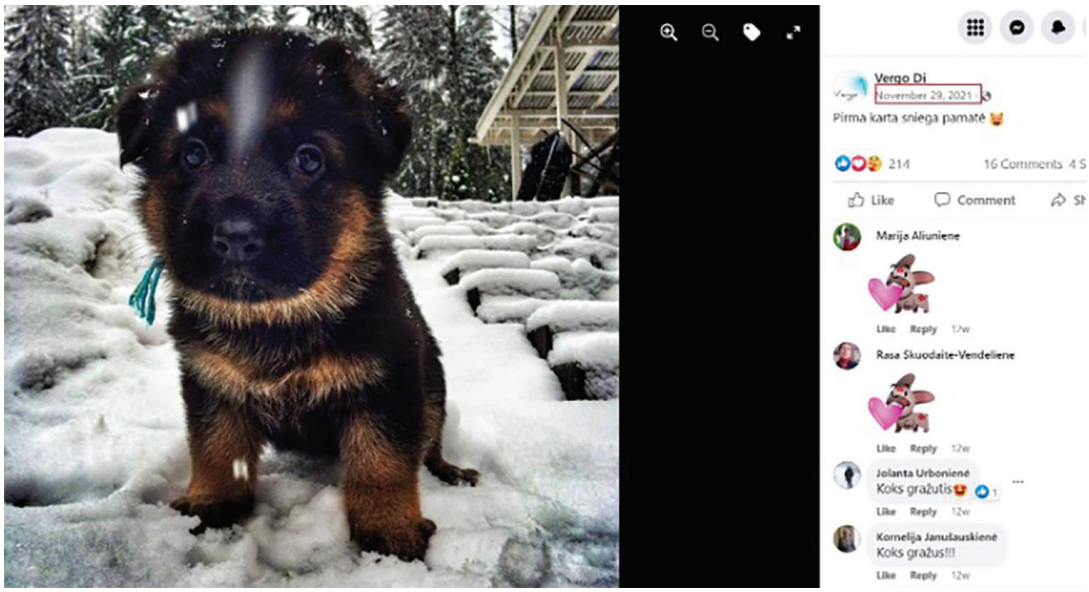
# Как определить, что аккаунт Facebook является поддельным?

Еще одна важная вещь, которую следует иметь в виду, — это фальшивые аккаунты. Обычно эти аккаунты не так активны, как тролли, и больше склонны к молчаливому наблюдению. Схожие критерии применимы к фальшивым аккаунтам на большинстве платформ, но в этом руководстве соцсеть Facebook была выбрана в качестве основного примера. Facebook кажется наиболее важным, так как пользователи обычно делятся там наиболее личной информацией. Эти аккаунты активно пытаются стать вашими друзьями по двум основным причинам: чтобы казаться более реальными, имея в друзьях несколько реальных людей, и чтобы попасть в список друзей и видеть больше личной информации. В зависимости от целей фальшивого аккаунта, он может использоваться, например, для сбора личной информации сотрудников организации.

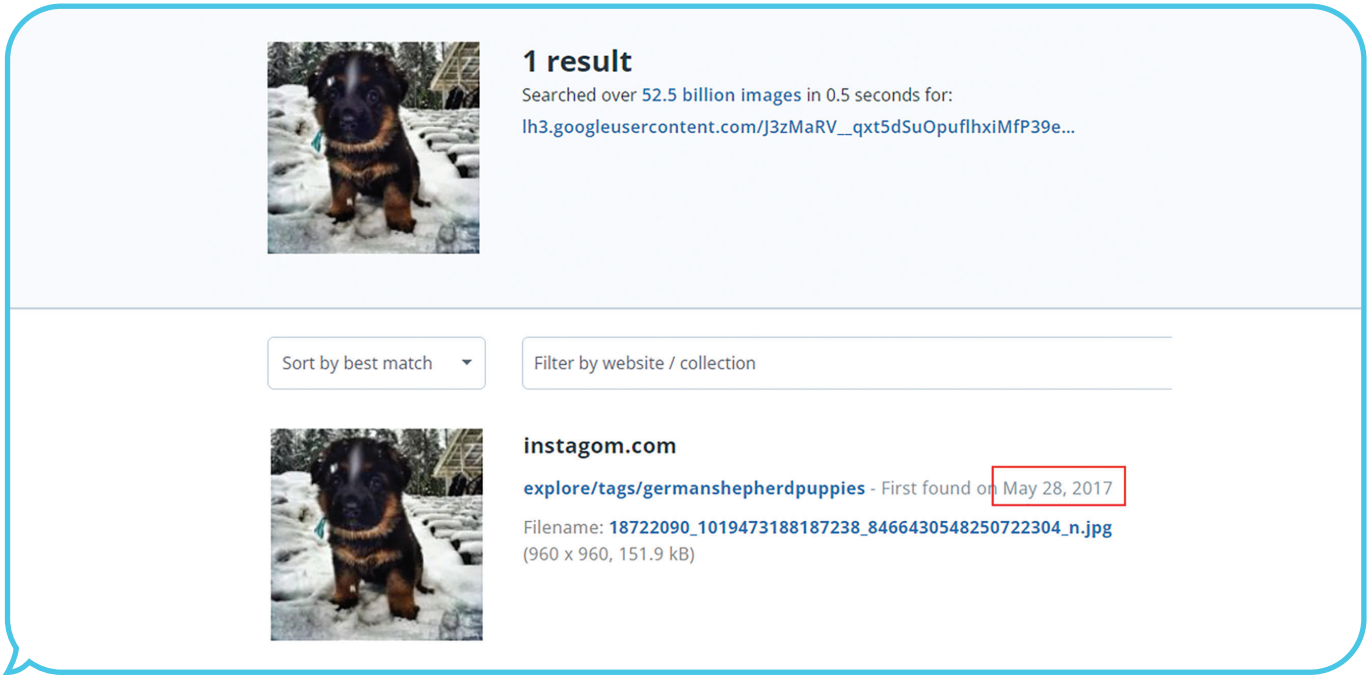


**1. Фактор привлекательности**

Аккаунты незнакомых вам визуально привлекательных пользователей, которые приглашают вас в друзья, могут быть фальшивыми.




Можно легко убедиться, что это не скопированное из интернета изображение, выбранное владельцем фальшивого аккаунта, чтобы привлечь ваше внимание и подружиться с вами.



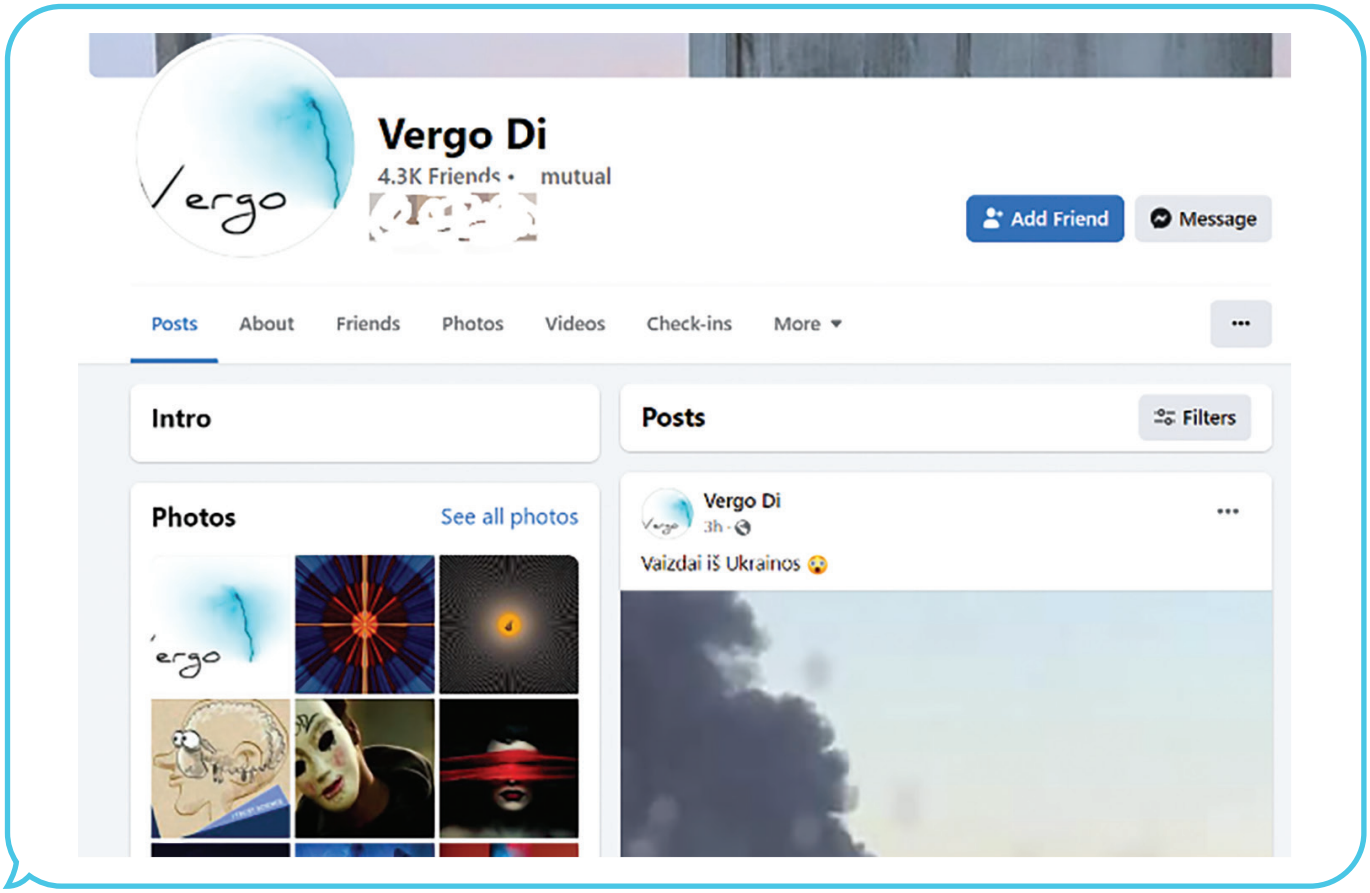

На сайте **https://tineye.com/** поиск изображений показывает дату, когда фотография была впервые опубликована. Можно сделать вывод, что профиль является фальшивой анкетой, привлекающей аудиторию, которая делится фальшивыми постами и иногда пропагандистскими сообщениями.

**2. Мало загруженных фотографий**

Большинство фальшивых аккаунтов не публикуют много фотографий - три-четыре типичные, и иногда это фотографии разных людей. Этого достаточно, чтобы создать временную иллюзию, что за аккаунтом якобы стоит реальный человек.

**3. Странные биографии**

В большинстве фальшивых аккаунтов биография содержит очень скудную информацию, либо предоставленная информация выглядит странно. Например, не исключено, но крайне маловероятно, что человек родом из Бронкса и учился в Хельсинском университете, но при этом очень молод и работает в нью-йоркской PR-фирме. Быстрая проверка имени и фамилии в Google, а также обратный поиск по картинке профиля помогут вам быстро разоблачить фальшивый аккаунт.




**4. Неотзывчивость**

Если вы обратитесь к фальшивому аккаунту, маловероятно, что он ответит даже на короткий вопрос. В идеале лучше даже не пытаться связаться с ним.

**5. Почти пустая стена Facebook**

Как правило, единственное, что вы найдете на одной из таких фальшивых стен Facebook, — это новые „лайки" на странице каких-то компании или продуктов.

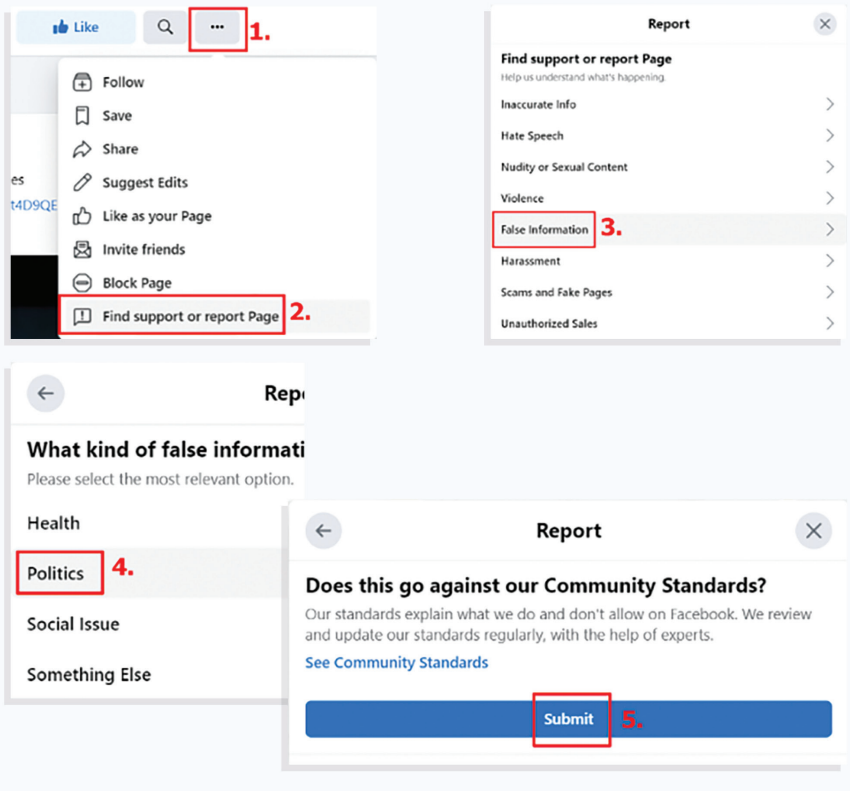
# Ответные меры

Для активной борьбы с онлайн-дезинформацией необходимы две составляющие - разоблачение и сообщение о ней.

## Обзоры медиа, Facebook / X

Очень важно поставить своих коллег в известность о возникающей дезинформации, направленной против организации. В каждой организации должен быть четкий процесс, чтобы сотрудники знали, куда отправлять сообщения о той или иной истории. Главная цель этого шага - донести до ваших коллег, что определенное циркулирующее сообщение является ложным, и предотвратить их распространение и веру в эту фейковую историю.

Второй шаг - сообщить об этом социальным сетям. Все социальные медиа-платформы имеют возможность сообщать о фейках с указанием конкретной причины, по которой о них сообщается. Если социальная платформа получит достаточно жалоб от пользователей, история или сообщение будут удалены. Именно этот метод используют общественные организации (Интернет-"Эльфы") для борьбы с дезинформацией в сети. Если СМИ распространяют фальшивые истории, в зависимости от характера СМИ (настоящие они или пропагандистские), об этом следует сообщить либо им самим, либо в национальные органы контроля за СМИ.



# Риски и опасности использования Telegram

Поскольку многие молодые люди используют Telegram для повседневного обмена сообщениями, они знают и о других возможностях этой платформы - группах и каналах, некоторые из которых могут быть приватными и анонимными. В то время как открытые коммуникационные потоки являются обычным явлением для социальных сетей, закрытые сообщества могут иметь специфический контент, включая необъективную и искаженную информацию, происхождение которой очень сложно проверить.

Было проведено несколько публичных расследований различных дезинформационных кампаний, которые использовали каналы и группы Telegram для распространения фейков, например, о пандемии COVID-19, пророссийских взглядах на войну против Украины, ультраправой риторике и теориях заговора на разных языках во многих странах.

Каналы и группы Telegram также могут использоваться для мобилизации пользователей на проведение идеологических флешмобов или политических акций протеста. Нет никаких проблем, если эти мероприятия организованы демократично и прозрачно, но иногда реальные бенефициары скрыты, а информации о реальных лицах, стоящих за акциями, мало. Например, несколько анонимных групп и каналов Telegram использовались для распространения провокационных призывов во время кампании „Я - русский“ в Эстонии:

https://eng.obozrevatel.com/section-life/news-russians-in-tallinn-threw-a-hysterical-tantrum-because-the-police-forced-them-to-remove-i-am-russian-stickers-from-their-cars-video-27-09-2023.html

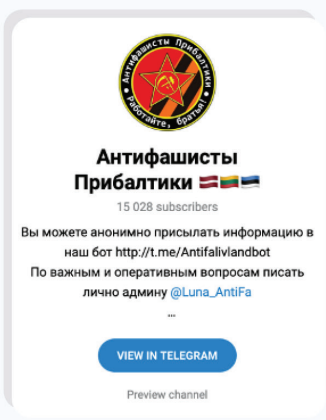
Поскольку многие дезинформаторы и пропагандисты (в том числе прокремлевские) используют собственные Telegram-каналы, их контент легко попадает в другие группы и сообщества, которые управляются анонимными администраторами. Это один из самых распространенных каналов распространения антизападной, антилиберальной и анти-балтийской дезинформации в цифровом формате.

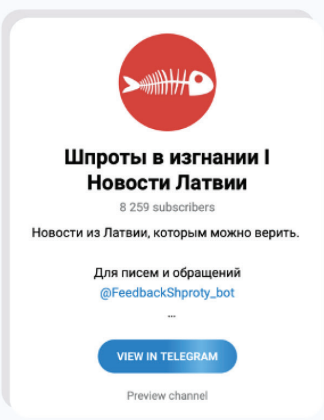
Например, Telegram-канал „Антифашисты Прибалтики" в значительной степени ответственен за раздувание числа пользователей. Некоторые из постов распространяли нарратив „русофобии". В одном из постов речь шла о якобы русофобии Латвии, которая решила запретить показ „чебурашек" в своем кукольном театре. Этот пост также подстрекал к грязной критике министра культуры Латвии, загрузив его (фальшивую) фотографию, на которой он позирует рядом с гротескными предметами, заявляя, что „это лицо латвийской культуры и национальной идентичности".

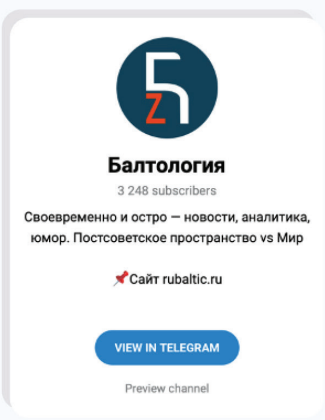
В другом посте говорится, что любой человек на „свободном Западе", который осмелится хотя бы намекнуть на право России защищать русский народ, немедленно окажется за решеткой с конфискацией имущества и запретом на любую экономическую и творческую деятельность. Эти сообщения показывают, что российская машина дезинформации активно пыталась продвинуть нарратив о том, что русская культура как будто подвергается нападкам в странах Балтии, и что русскоязычное меньшинство не может высказываться по этим вопросам, опасаясь репрессий.

**Несколько простых рекомендаций помогут укрепить вашу цифровую гигиену в Telegram:**

• прежде чем вступать в группу или канал, убедитесь, что контент на 100% вас интересует, так как название и прикрепленные посты могут быть обманчивы, посмотрите дальше и узнайте больше информации об администраторах онлайн-сообщества;
• будьте внимательны к сильно искаженной информации в анонимных группах и каналах, если тема серьезная, подлинный контент должен иметь достоверные источники и не должен содержать спекулятивных мнений, „альтернативных фактов" или упрощенных пропагандистских клише;
• Если какая-то информация в группе или канале Telegram вызвала у вас эмоции, спросите себя, почему это произошло и кому это выгодно - не спешите реагировать или делиться информацией, так как она может быть необъективной, поляризующей, оскорбительной или просто фальшивой;
• О любом тревожном контенте в группе или канале Telegram можно сообщить как администраторам, так и веб-полиции, а также сообществу по проверке фактов (например, CRI в Литве или Propastop в Эстонии). Обязательно сохраняйте оригинальный контент как можно лучше (например, делайте скриншот, копируйте тексты, изображения, теги и т. д.).



# Риски и опасности использования TikTok

Приложение TikTok, которое в последнее время набирает популярность среди молодежи, имеет китайское происхождение и характеризуется высоким уровнем дезинформации в сети. Сами создатели приложения говорят, что прилагают все усилия для борьбы с дезинформацией, радикальным экстремизмом и ненавистническим поведением, но как же обстоят дела на самом деле?

Хотя поначалу социальная сеть не казалась опасной, со временем, по мере роста числа пользователей, контент начал меняться. Теперь в сети очень много контента, распространяющего прокремлевские **нарративы***.

***Нарратив** – это систематическое и связное повествование, которое создается путем повторения сообщений на определенную тему, добавления к ним новых фактов и контекста.

**Нарратив** — это история, которая убедительно передает основную мысль и формирует мнение.

Платформа TikTok характеризуется весьма непрозрачными алгоритмами и уважением к законам авторитарных режимов. Эта социальная сеть также отличается от других своим эффектом привыкания. Он проявляется в постоянном просмотре коротких видеороликов, насыщенных эмоциональными элементами и запоминающимся саундтреком. Чем больше времени проводится в TikTok, тем лучше становится алгоритмическая подача информации, а в общем потоке потребляемого контента появляются пропагандистские сообщения.

**Можно выделить две фундаментальные проблемы, связанные с использованием TikTok:**

• **Агрессивный сбор данных.** Когда это приложение устанавливается на телефон или другое смарт-устройство, оно запрашивает дополнительный доступ к данным, которые затем собираются. Один из главных рисков использования TikTok заключается в том, что оно может увидеть контакты пользователя, посмотреть, какие еще приложения используются на устройстве, а также узнать о местоположении и определить, где находится устройство. Существует также риск, что переписка, которая ведется в приложении, может быть отслежена, просмотрена и, основываясь на определенных ключевых словах, попасть в поле зрения самой компании.

• **Теневой запрет записей** (*shadow banning* - англ.). Если пользователь публикует пост, который не нравится властям TikTok, этот пост будет скрыт, то есть будет осуществлен теневой бан.

TikTok использует алгоритм в качестве инструмента для привлечения и удержания внимания и предоставления индивидуального опыта для каждого пользователя. Алгоритм использует данные, полученные от пользователей, чтобы определить, какой контент может заинтересовать именно вас. Например, чем дольше вы смотрите видео, тем больше похожих видео вы увидите на TikTok в будущем. Он также запоминает ваши поисковые запросы о ваших интересах, чтобы в будущем предлагать вам видео в том же стиле. Кроме того, алгоритм TikTok объединяет пользователей со схожими интересами, показывая им похожий контент. Это не просто совпадение, если вы видите те же видео, что и некоторые ваши друзья.

Помните, что TikTok был создан не для передачи новостей, поскольку видео там очень короткие и ориентированы на развлекательные интересы пользователя. Прокрутка и просмотр видео на TikTok активизирует участки мозга, отвечающие за ощущение успеха, как в азартных играх, с надеждой поднять настроение. Но вместо реальной прибыли тратится огромное количество времени. Поскольку видео на TikTok короткие, развлекательные и легкодоступные, это создает своего рода зависимость, которая приводит к потере фокуса внимания.

TikTok использует ваше любопытство, а также страх пропустить что-то важное. Слышали ли вы фразу „Если вас нет в TikTok, значит, вас не существует“? Это навязчивая манипуляция, чтобы заставить молодых людей быть в постоянном контакте. Еще один трюк, используемый TikTok, - „ферма ярости“: он предлагает видео, оскорбляющие какую-либо группу людей, идеологию или движение, в надежде, что такой контент оскорбит зрителей и вызовет эмоционально острую дискуссию, которая привлечет больше внимания и сделает видео вирусным. Эти методы используются для увеличения количества кликов и подписчиков. Поскольку в TikTok существует множество случаев кибербуллинга, есть высокий риск стать жертвой преследования или разжигания ненависти. Более того, поскольку любой может создавать контент в TikTok, он может содержать необъективную или манипулируемую информацию. Поскольку в TikTok много персонализированного контента, молодым пользователям становится еще сложнее отличить чье-то мнение от фактов.

**TikTok может стать отличным средством развлечения, если следовать простым советам по его использованию:**

• ограничьте время, которое вы ежедневно проводите в TikTok - встречаться с друзьями в реальной жизни всегда веселее, чем в сети

• не забывайте критически относиться к новостному контенту в TikTok - перепроверяйте интересующую вас информацию из других источников (не из социальных сетей)

• сообщайте о кибер-буллинге, преследованиях и разжигании ненависти, если вы их видите, - избегайте распространения оскорбительных видео или вызывающего сильные эмоции контента

• если вы чувствуете, что вас смущает контент или вы им манипулируете, обратитесь к родителям или в веб-полицию.

## Расширенные меры безопасности и обращение за помощью

Когда речь идет о более сложных случаях дезинформации, возможны два основных подхода: использовать более сложные методы с открытым исходным кодом или обратиться за помощью к сообществу онлайн-аналитиков и фактчекеров. Большинство онлайн-инструментов относительно просты в использовании и содержат пошаговые инструкции по их применению. Ниже представлены два крупных и наиболее полезных инструментария:

**Bellingcat's Online Investigation Toolkit**

**Online Open Source Tool Box**

Другой вариант - обратиться к сообществу аналитиков и фактчекеров в Интернете и предоставить им информацию о той или иной истории. Большинство расследователей будут рады развенчать ложную историю и поделиться этим результатом в сети.

Вильнюс, Литва
**Тираж 500**


Спонсором этой публикации является компания

Google