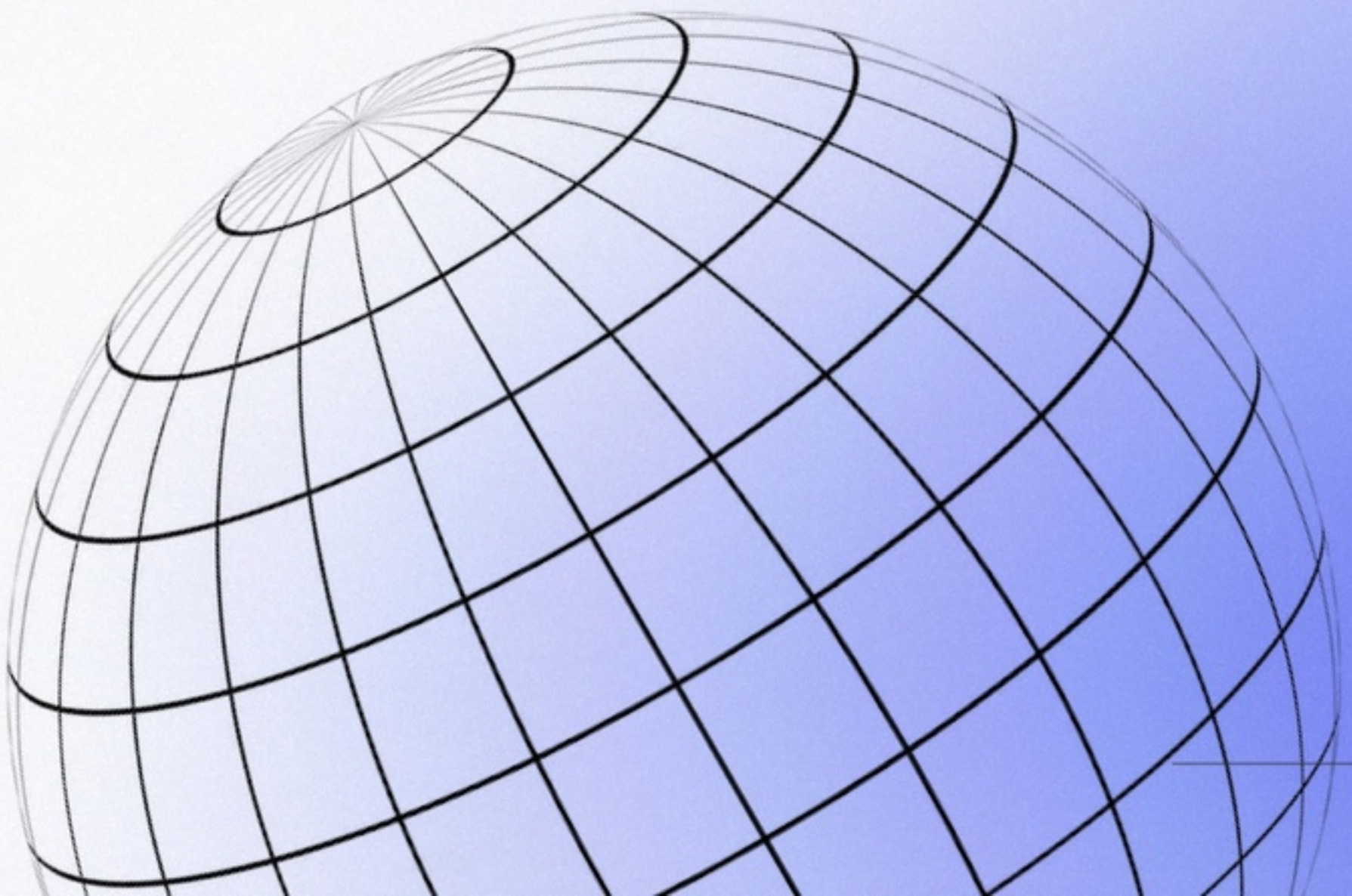
КАК НЕ ДАТЬ НЕЙРОСЕТЯМ СЕБЯ ОБМАНУТЬ?

#нцпти_медиаграмотность

Что нейросети вообще могут?

Полностью создавать контент: писать тексты, генерировать картинки по заданным параметрам и преобразовывать их в видео, изменять голос или изображения в кадре.

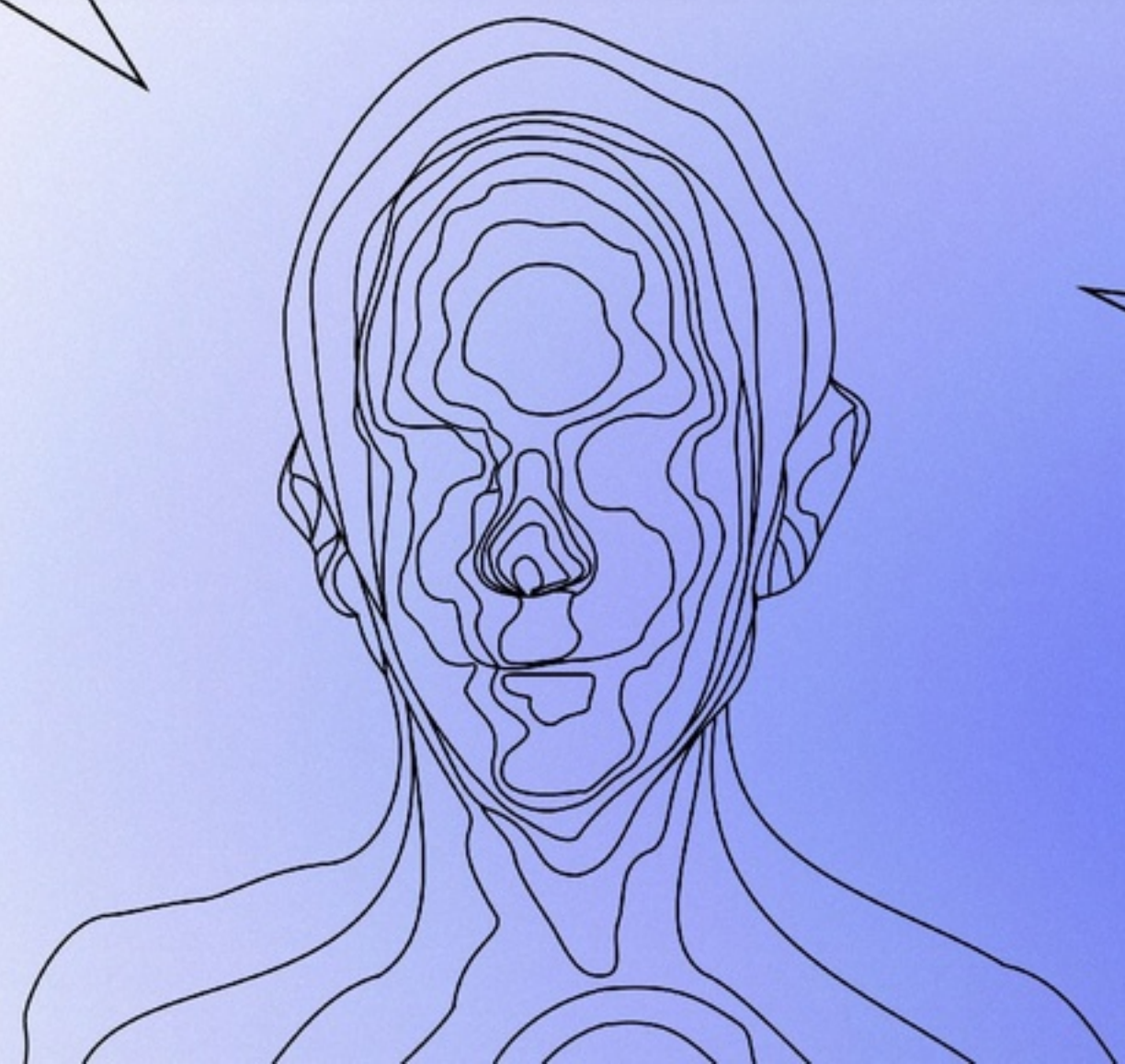
Некоторые из этих функций пока работают плохо. Но стоит помнить, что всего несколько лет назад технология «deepfake» (подмена лиц в видео) была фантастикой.



То есть нейросети всемогущи, и люди теперь не нужны?

Нейросети не могут создавать новые материалы. Всё, что они генерируют, пока является продуктом анализа уже существующих элементов.

Не весь контент, созданный нейросетями, удачен, но зачастую пользователи загружают в социальные сети самый приемлемый и интересный результат — это и создает иллюзию неограниченных возможностей нейросетей.



Как нейросети могут нас обмануть?

Открытые для большинства пользователей нейросети функционирует в основном для развлечения и рекламы технологических возможностей их разработчиков. Однако этой технологией стали активно пользоваться злоумышленники.

Исследователи McAfee опросили более 7 тыс. человек из разных стран. Около 25% респондентов заявили, что сталкивались с голосовым мошенничеством, в результате которого перевели деньги злоумышленникам.

Мошенники подделывают голоса и «внешность» чтобы выманить у него деньги, конфиденциальную информацию или побудить на совершение противоправных действий.

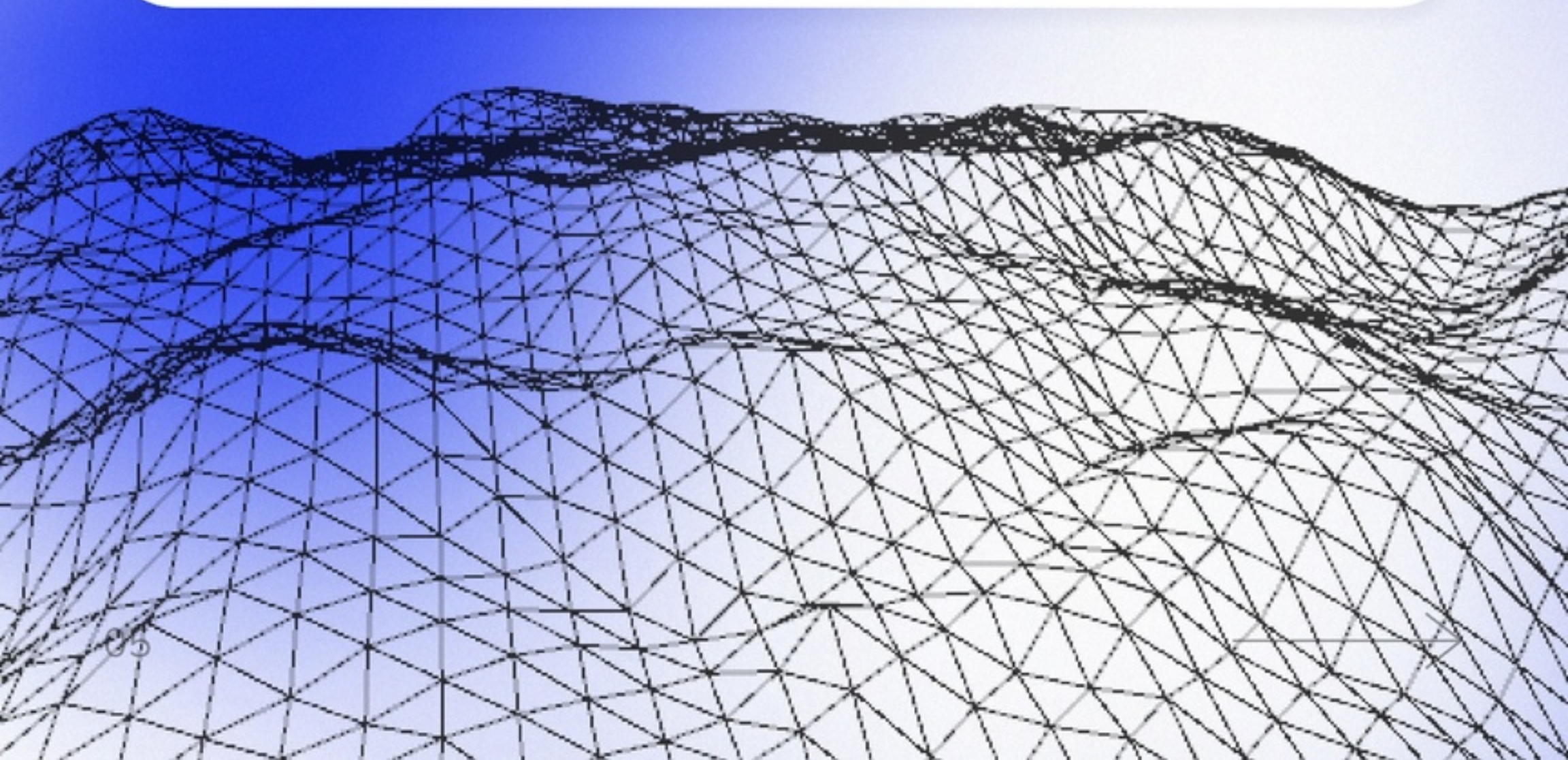
Также сгенерированные картинки часто используют в подтверждение фейковых (ложных) новостных сюжетов.

#нцпти_медиаграмотность

Как можно определить, что материал создан нейросетью?

Созданное нейросетью изображение может содержать различные изъяны (визуальные артефакты), заметные при внимательном изучении материала. Это могут быть лишние части тела, неподходящие элементы или неестественное освещение.

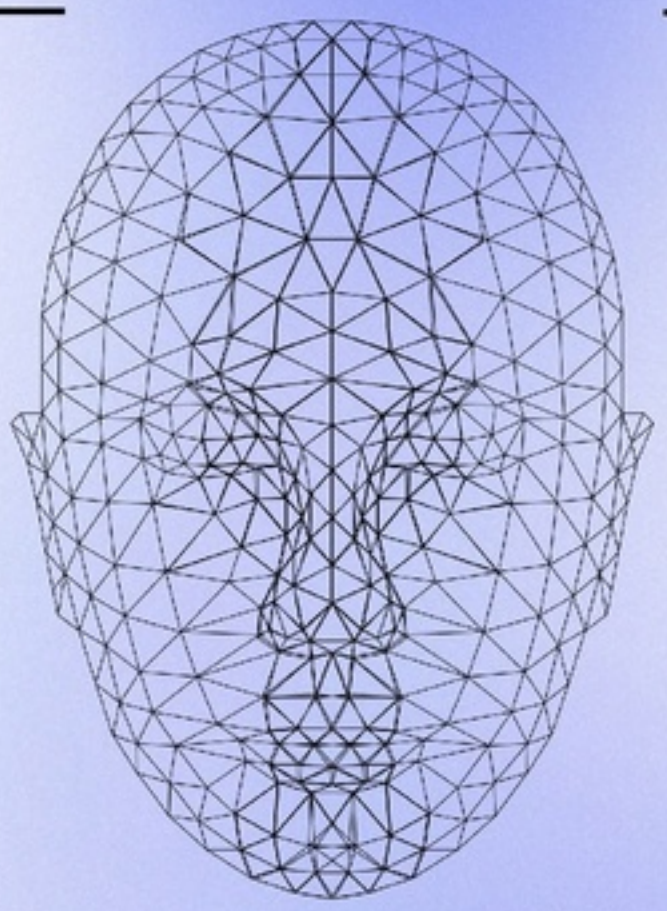
Видео и звук также могут содержать ошибки. Однако они могут быть устранены при пост-обработке сгенерированного контента.



А есть какие-то программы, чтобы быстро определить след нейросети?

Для проверки текста можно использовать GPTZero, Crossplag.com, последние версии «Антиплагиата» и даже сами нейросети (например, ChatGPT).

Для проверки изображения: Hive Moderation, Ai Or Not (и одноимённый бот в Telegram).



А есть какие-то программы, чтобы быстро определить след нейросети?

Для проверки видео универсального инструмента пока не существует. В случае с онлайн-звонком хороший способ распознать наложение чужого лица — попросить собеседника повернуть голову в профиль. Большая часть нейросетей не может справиться с этим и «маска» может пропасть.

Программная проверка звука пока недоступна для рядовых пользователей. Впрочем, внимательный пользователь может определить подделку самостоятельно.