EFCSN

Como identificar e compreender conteúdo gerado por IA ou alterado digitalmente

Os modelos acessíveis - e de fácil utilização - de IA podem ajudar as pessoas a consumir e a criar conteúdos, **mas também podem amplificar os riscos que a desinformação representa** para as sociedades livres e democráticas. É importante evitar que os nossos espaços partilhados de informação fiquem sobrecarregados com desinformação gerada por IA ou alterada digitalmente.

Parte da solução são as novas tecnologias, como as que ajudam a descortinar a proveniência do conteúdo, e *software* de detecção. **Mas as soluções tecnológicas estão longe de ser perfeitas e precisamos do trabalho dos verificadores de factos independentes que disponibilizam à sociedade factos verificados.**

Aqui está visão geral daquilo que os verificadores de factos, independentes e profissionais, fazem para identificar e desconstruir a desinformação gerada pela IA e aquilo que podes aprender com eles.

O conteúdo gerado por IA está a aumentar

Hoje, a desinformação gerada por IA constitui uma pequena percentagem de todas as alegações analisadas pelos verificadores de factos profissionais e independentes. O conteúdo alterado digitalmente é, atualmente, mais comum.

Porém: num inquérito interno aos membros da EFCSN, **a maioria dos verificadores concordou que os conteúdos gerados por IA e alterados digitalmente serão ainda mais relevantes no futuro.** E os exemplos recentes no contexto das Eleições Europeias sustentam este prognóstico.

NOTA: O termo *alterado digitalmente* refere-se a conteúdo que tenha sido alterado significativamente para manipular ou alterar a mensagem original, incluindo edições com ferramentas de IA. Tal não inclui edições para maior clareza ou qualidade.

Gerado por IA refere-se a qualquer formato de conteúdo criado por um sistema de inteligência artificial.

Informações fornecidas por AI@EUElections, financiado pela Meta e gerido pela European Fact-checking Standards Network em parceria com os seus membros e especialistas externos.

A tecnologia avança rapidamente mas não podemos confiar só nela

Especialistas em IA e *fact-checkers* concordam: **ferramentas de detecção de IA, por si só, não são suficientes para identificar conteúdos gerados por IA.**

Antes de usar um detector, os especialistas recomendam desenvolver familiaridade com geradores e detectores de conteúdo de IA. Compreendendo a forma como os modelos são treinados - e um pouco de estatística - os *fact-checkers* podem começar a reconhecer os pontos fortes e fracos de uma ferramenta. **Mesmo assim, as ferramentas podem ser um ponto de partida útil.**

Proveniência do conteúdo
Iniciativas como as especificações C2PA podem ajudar a certificar a fonte e o histórico de um conteúdo, mas a marca d'água e a verificação não são incontestáveis.

Desinformação por IA: como nos afeta?

"Cada vez que reagimos, se assim posso dizer, com coragem, ignoramos a reflexão"
- Christine Dugoin*

PSICOLOGIA: Operações de influência são desenhadas para tirar partido de vieses psicológicos.

Perceber os nossos - e os das nossas audiências - pode ajudar a combater a desinformação.

OBJETIVOS: Porque é que alguém mal intencionado utiliza a IA para criar ou espalhar desinformação? Qual é o impacto pretendido no mundo real?

- ➔ Expandir o seu alcance para outro país ou comunidade?
- ➔ Sobrecarregar os *fact-checkers* gerando muitas variantes de alegações semelhantes?
- ➔ Influenciar pensamentos ou crenças estabelecendo credibilidade através de redes de contas inautênticas?

** Christine Dugoin é um investigador em influência informativa na Universidade de Sorbonne.*

Desmentir requer uma abordagem multifacetada e uma compreensão subtil

Se as ferramentas de detecção não funcionam, o que funcionará? É tão importante perceber o contexto de uma alegação como o seu conteúdo. Os *fact-checkers* têm competências de investigação para tal. Aqui estão algumas dicas.

> *"As ferramentas de detecção nunca funcionarão a 100% - não espero que funcionem"*
> - Henk van Ess**

CONSIDERA A FONTE: Consegues confirmar a sua identidade? Sobre o que é que fala e partilha? Quem interage com o seu conteúdo? Que efeito esse conteúdo pode ter nos leitores?

ESTABELECE CREDIBILIDADE: Verifica as informações de forma independente com fontes credíveis, como por exemplo, especialistas com experiência prática em determinada área. O que é alegado faz sentido com base no seu conhecimento?

Usa técnicas de **ANÁLISE FORENSE DIGITAL** além da investigação tradicional e da pesquisa documental. Algumas técnicas incluem: organização de dados, geolocalização, reconhecimento biométrico, análise de padrões, entre outras.

APRENDER & ADAPTAR: Os criadores de desinformação gerada por IA estão em constante adaptação. Ajusta também a tua abordagem.

PARTILHA O TEU TRABALHO

Adicionalmente à refutação de uma alegação, os especialistas recomendam adotar uma análise transparente e fornecer links para as fontes. Isto pode ajudar os leitores a acompanhar a investigação e a compreender uma narrativa mais complexa.

Em alguns casos, a investigação é mais importante do que saber simplesmente se o conteúdo foi gerado por IA.

*** Henk van Ess é especialista em OSINT e técnicas de fact-checking.*

Guia rápido de referência: O que fazer – e não fazer – e dicas

Estes indícios podem indicar que um conteúdo foi gerado por IA ou alterado digitalmente. Juntamente com as outras dicas mencionadas neste guia (contexto, técnicas de investigação e ferramentas de detecção), estes podem ajudar a compreender a verdade por trás do que vê.

Texto

- Por vezes tem **melhor gramática** do que um humano.
- Tendência para usar linguagem **excessivamente formal ou estruturada** para *social media*.
- **Uso excessivo de advérbios e adjetivos.**
- Carece de emoção humana, humor, sarcasmo e expressões idiomáticas.
- Pode ter falta de **detalhes** (nomes, datas, locais) ou ideias originais.
- Mais importante: os factos apresentados no texto estão corretos?

Vídeo

- Não uses detectores de imagens geradas por IA em fotogramas de um vídeo.
- Olha para as **expressões faciais e movimentos** como o piscar dos olhos e se o movimento da boca corresponde ao áudio.
- Pode ser caracterizado por **transições e cortes abruptos.**

Áudio

- **Compara o áudio suspeito com uma amostra autêntica** utilizando ferramentas que conseguem detectar diferenças nos padrões de discurso e da respiração...
- Ao utilizar detectores, evita amostras de áudio com pouca qualidade ou com estática e ruído de fundo.
- Pode ter **padrões de discurso mecânicos ou pouco naturais**, falta de pausas e de respiração natural.

Imagens

- Procura zonas com **detalhes pouco naturais**: pele perfeita, fundos desfocados, beleza ou luz artificial e irregularidades como dedos a mais.
- **Procura por marcas d'água** de geradores de imagem comuns.
- Toma atenção aos detalhes: É lógico? É apropriado?
- Ao usar detectores, **opta por uma versão de alta resolução da imagem que já foi partilhada e repartilhada.**