

Material de apoio à apresentação

O que são deepfakes?

Quando falamos de deepfakes, estamos a referir-nos a ficheiros de vídeo, imagem ou áudio que foram gerados ou manipulados usando inteligência artificial para mostrar pessoas a dizer ou a fazer coisas que nunca aconteceram. Não se tratam de simples "retoques" ou de filtros engraçados; são criações sintéticas que utilizam algoritmos avançados para imitar de forma convincente a aparência física e a voz de uma pessoa real. No nosso dia a dia, deparamo-nos com este conteúdo integrado nas redes sociais, por vezes para entretenimento e outras vezes com intenções mais nefastas.

O termo deriva da combinação de aprendizagem profunda (a aprendizagem profunda das máquinas) e falsa. Portanto, a chave não é apenas que o resultado pareça real, mas que tenha sido "treinado" por uma máquina que aprendeu cada detalhe das características ou da voz da pessoa imitada para enganar o nosso julgamento.

Que tipos existem?

Basicamente, podemos agrupar os deepfakes em três categorias que os alunos costumam encontrar nos seus dispositivos:

- **Vídeo:** Onde ocorre uma troca de rosto ou a boca é manipulada para corresponder a um discurso falso.
- **Imagem estática:** Fotos de situações que parecem reais, mas são criadas de raiz por uma IA.
- **Voz ou áudio:** Talvez o mais difícil de detetar atualmente; o tom, o ritmo e a forma de falar de uma pessoa são clonados para criar mensagens sonoras falsas.

É fundamental realçar que estes tipos de recursos são frequentemente combinados. Por exemplo, um vídeo pode ser mais convincente se a voz também tiver sido clonada para soar exatamente como a do protagonista.

Será que os deepfakes podem afetar pessoas comuns na nossa escola ou

Enquanto os noticiários falam de presidentes ou celebridades, as aplicações para criar deepfakes estão nos bolsos de todos os estudantes. Precisamos falar abertamente sobre como esta tecnologia pode ser utilizada para criar "piadas" que, na verdade, são formas de assédio, como vídeos que ridicularizam alguém na sala de aula ou, nos casos mais graves, imagens íntimas falsas que prejudicam profundamente a reputação e a saúde mental dos jovens.

Como professores, a nossa missão é lembrar os alunos que por detrás de cada ecrã existe uma pessoa real. Independentemente de um vídeo ser ou não um deepfake, se a sua disseminação puder humilhar ou causar sofrimento a alguém no nosso meio, temos o dever ético de não participar na sua cadeia de distribuição. Devemos fomentar uma cultura de cuidados, na qual os alunos compreendam que a sua capacidade de gerir estas ferramentas digitais está diretamente ligada à proteção e ao respeito pela dignidade dos seus pares.

Material de apoio à apresentação

Que indícios nos podem fazer duvidar da sua autenticidade?

Tal como num truque de magia, se observarmos atentamente, por vezes conseguimos identificar a presença da IA num deepfake. Estas falhas ou erros técnicos ocorrem porque a inteligência artificial ainda tem dificuldades em recriar a complexidade física do mundo real e os movimentos humanos naturais.

Para ajudar os nossos alunos a observar melhor, podemos agrupar estas falhas em diversas categorias visuais:

- **Olhar e olhos:** Um dos pontos mais difíceis para a IA. Podemos verificar se o piscar de olhos é demasiado rígido ou, pelo contrário, se a pessoa quase não pestaneja. Devemos também observar os reflexos nas pupilas; num vídeo real, o brilho deve corresponder à iluminação do ambiente, mas num deepfake, os olhos por vezes brilham artificialmente ou não olham exatamente para onde deveriam.
- **Boca e sincronização:** Muitas vezes, o movimento dos lábios não sincroniza perfeitamente com o som, como num filme mal dobrado. Podemos observar os dentes, que por vezes parecem uma "massa branca" desfocada em vez de dentes individuais, ou a língua, cujo movimento é geralmente artificial ou inexistente quando a pessoa fala.
- **Bordas e texturas:** Analisaremos a zona onde o rosto se encontra com o cabelo ou o pescoço. Se, ao virar a cabeça, a pele parecer borrada, o cabelo parecer uma mancha sólida ou houver um ligeiro "tremor" ou ondulação nas bordas do rosto, é um sinal claro de manipulação. É também comum a pele parecer demasiado perfeita, sem poros, imperfeições ou rugas naturais.
- **Física e ambiente:** Devemos observar se as sombras do nariz ou do pescoço fazem sentido de acordo com a iluminação. Também é útil observar os acessórios: os óculos apresentam frequentemente reflexos estranhos que mudam de forma, e os brincos ou colares podem "derreter" ou deformar-se quando a pessoa se move rapidamente.

É importante lembrar o grupo que estes erros estão a desaparecer à medida que a tecnologia avança. Portanto, embora encontrar uma destas falhas seja uma excelente prova, **não as encontrar não garante que o vídeo seja real**. Devemos incentivá-los a usar estas pistas visuais como ponto de partida para, de seguida, investigar o mais importante: o contexto.

O que são marcas de água e rótulos de IA?

As marcas de água e os rótulos são ferramentas de transparência utilizadas pelas plataformas e criadores de conteúdo para alertar que o conteúdo não é real. Algumas redes sociais já identificam automaticamente conteúdo gerado por IA, e alguns programas adicionam pequenas marcas ou metadados invisíveis confirmando que o ficheiro é sintético.

No entanto, devemos alertar os alunos de que estes sinais não são perfeitos. Os rótulos podem ser removidos cortando o vídeo, e nem todos os sistemas de IA são obrigados a utilizá-los. É por isso que explicamos que, embora ver um rótulo de "IA" nos dê uma resposta imediata, **não o ver não significa que o vídeo seja real**. São uma ajuda preciosa para a nossa investigação, mas nunca devem substituir o nosso pensamento crítico ou a nossa análise do contexto social.

Material de apoio à apresentação

Porque é que estas pistas visuais são úteis, mas não suficientes?

O grande desafio pedagógico é fazer com que os alunos compreendam que a tecnologia evolui mais rapidamente do que a nossa visão. O que hoje é uma falha técnica (como um dedo a mais ou um piscar de olho estranho) será corrigido amanhã por uma versão mais poderosa da IA. Se basearmos a nossa educação apenas em "detetar a falha", os nossos alunos ficarão completamente desprotegidos quando a IA se tornar perfeita.

Além disso, existe aquilo que conhecemos como o "**dividendo do mentiroso**": se habituarmos as pessoas a acreditar que tudo o que parece um pouco estranho é falso, quem mente poderá alegar que um vídeo real é um deepfake para escapar às suas responsabilidades. Ao focarmos a nossa atividade no contexto social e não apenas no pixel, proporcionamos aos alunos uma capacidade analítica que não se esgota, independentemente do quanto a tecnologia de geração de imagens melhora.

De que outra forma podemos estudar a fiabilidade de um possível deepfake?

Em vez de procurarmos apenas falhas técnicas, devemos ensiná-los a procurar **pistas contextuais**. Uma pista fundamental é a **intenção emocional**: se um vídeo provoca raiva súbita, choque ou uma urgência extrema em partilhá-lo, estamos provavelmente perante uma manipulação. Devemos também observar a **coerência**: é lógico que aquela pessoa diga aquilo naquele local? Se o ambiente não se coaduna com a situação ou se a conta que publicou o conteúdo for anónima, recém-criada ou dedicada à publicação de conteúdo sensacionalista, as nossas suspeitas devem começar a aumentar.

Outra pista importante para os alunos é a **proveniência**: se o vídeo aparecer como um "reencaminhamento" sem uma fonte original clara ou se não houver outros meios de comunicação fidedignos a noticiar o assunto, temos uma forte razão para duvidar.

O que devemos fazer quando não temos a certeza?

A resposta mais importante que podemos dar é: **pare**. Num ambiente digital que nos impele para a instantaneidade, o simples gesto de não partilhar algo de que duvidamos é um ato de cidadania digital. Devemos transmitir a mensagem de que não somos obrigados a saber instantaneamente se tudo é real ou falso; o "travão de emergência" consiste em pausar a nossa reação emocional e impedir que o potencial engano chegue a mais pessoas através de nós.

Após a interrupção, o passo seguinte é a verificação ativa. Podemos sugerir que os alunos procurem a fonte original, façam uma pesquisa por palavras-chave para verificar se outros meios de comunicação estão a abordar o assunto ou utilizem ferramentas de pesquisa inversa de imagens. E, acima de tudo, devemos reforçar a ideia de que, em caso de dúvida, o melhor é consultar um adulto de confiança. Pedir ajuda para compreender um conteúdo confuso não é sinal de fraqueza, mas sim de maturidade digital.