Não confie nos seus olhos: manipulação de imagens na era dos DeepFakes

Johannes Langguth1 * Konstantin Pogorelov1 Stefan Brenner2 Petra Filkuková1,3 Daniel Thilo Schroeder4,5

Analisamos o fenómeno dos deepfakes, uma nova tecnologia que permite a manipulação barata de material de vídeo através do uso de inteligência artificial, no contexto da discussão mais ampla de hoje sobre notícias falsas. Discutimos a base, bem como os desenvolvimentos recentes da tecnologia, bem como as diferenças das técnicas de manipulação anteriores e investigamos contramedidas técnicas. Embora a ameaça dos vídeos deepfake com impacto político substancial tenha sido amplamente discutida nos últimos anos, até agora o impacto político da tecnologia tem sido limitado. Investigamos as razões para isso e extrapolamos os tipos de vídeos deepfake que provavelmente veremos no futuro.

Introdução

Desde a invenção da fotografia no século XIX, os meios visuais têm desfrutado de um elevado nível de confiança por parte do público em geral e, ao contrário das gravações de áudio, as fotos e os vídeos têm sido amplamente utilizados como prova em processos judiciais (Meskin e Cohen, 2008 **)** . e é amplamente aceite que os meios visuais são uma ferramenta de propaganda particularmente eficaz (**Winkler e Dauber, 2014**). Consequentemente, os incentivos para a criação de documentos visuais falsificados sempre foram elevados. O uso extensivo de imagens manipuladas para fins políticos está documentado já na década de 1920 pela União Soviética (**Dickerman, 2000** ; **King, 2014**).

Por outro lado, a manipulação de vídeo exigia especialistas qualificados e uma quantidade significativa de tempo para ser criada, uma vez que cada quadro tinha que ser alterado individualmente. A tecnologia de manipulação de vídeos foi aperfeiçoada em Hollywood na década de 1990 (**Pierson, 1999**), mas era tão caro que apenas alguns filmes fizeram uso total dele. Consequentemente, a criação de vídeos manipulados para fins de propaganda política era rara. No entanto, uma tecnologia conhecida como deepfake, que permite a manipulação de vídeos inteiros com esforço limitado e hardware de computação de nível consumidor, tornou-se recentemente disponível. Ele aproveita a inteligência artificial moderna para automatizar tarefas cognitivas repetitivas, como identificar o rosto de uma pessoa em cada quadro de um vídeo e trocá-lo por um rosto diferente, tornando assim a criação de um vídeo manipulado bastante barata.

Assim, a mudança fundamental não está na qualidade das mídias manipuladas, mas na fácil acessibilidade. Com habilidades técnicas moderadas, material de vídeo de entrada e equipamento de informática de consumo, quase qualquer pessoa pode criar vídeos manipulados hoje (**Hall, 2018** ; **Beridze e Butcher, 2019**). No entanto, até agora, os vídeos deepfake não desempenharam o papel proeminente na política que muitos inicialmente temiam (**Chesney e Citron, 2019a**), embora o ano de 2020 tenha assistido a uma enorme quantidade de outras informações erradas, especialmente relacionadas com a pandemia da COVID-19 e a Eleições presidenciais dos EUA.

Neste artigo, investigamos possíveis explicações para este desenvolvimento e discutimos os prováveis desenvolvimentos futuros e identificamos áreas nas quais se espera que os deepfakes sejam eficazes e, portanto, prováveis de aparecer no futuro. Revisamos a tecnologia deepfake, apresentamos os desenvolvimentos mais recentes de forma acessível e discutimos suas implicações no contexto da manipulação histórica da mídia. Nós nos concentramos apenas na manipulação de mídias visuais no sentido de criar ou alterar mídias diretamente, de modo que elas mostrem conteúdos que não correspondam à realidade física, e excluímos a desinformação criada por rotulagem incorreta, edição sugestiva e técnicas similares.

Manipulação de imagens individuais

Imagens individuais são relativamente fáceis de manipular através da técnica de retoque. A técnica já era amplamente utilizada na década de 1920, com maior destaque na União Soviética (**King, 2014**). Um exemplo proeminente é a remoção de Alexander Malchenko das fotos oficiais após sua execução em 1930 (**Dickerman, 2000** ; **King, 2014**) (**Figura 1**). Embora este exemplo, juntamente com muitos outros, e a possibilidade de manipulação de fotos em geral sejam comumente conhecidos há muito tempo, o esforço envolvido na criação de imagens manipuladas era comparativamente alto até recentemente. Além disso, os especialistas eram normalmente capazes de detectar tais manipulações. Consequentemente, as fotos, ao contrário das gravações de áudio, geralmente mantiveram a confiança do público. Ainda hoje, o termo “evidência fotográfica” é comumente usado, embora essa confiança pareça estar em declínio (**Meskin e Cohen, 2008**).

Figura 1

FIGURA 1 . **ESQUERDA:** Alexander Malchenko, P. Zaporozhets e Anatoly Vaneyev (em pé); Victor V. Starkov, Gleb Krzhizhanovsky, Vladimir Lenin e Julius Martov (sentado). **À DIREITA:** imagem manipulada com Malchenko removido. Fonte: Wikipédia (**Wikipédia, 2020**).

A manipulação de imagens está disponível ao público em geral desde a década de 1990. Na verdade, é tão comum hoje que o termo “to photoshop”, em homenagem ao programa Adobe Photoshop (**Adobe, 2020**), está sendo usado como verbo para o ato de manipular imagens. Normalmente, isso se refere a manipulações menores, como fazer uma pessoa se conformar a algum ideal de beleza, mas manipulações maiores para fins comerciais ou de entretenimento são abundantes (**Reddit, 2020**). Naturalmente, estas técnicas podem ser e têm sido utilizadas para fins de propaganda. Com a ajuda de um software de processamento de imagem, a manipulação mostrada na **Figura 1** , ou seja, a remoção de uma pessoa de uma imagem individual, hoje em dia requer alguns minutos de trabalho por parte de um usuário experiente.

Hoje, manipulações como a remoção de uma pessoa podem ser ainda mais facilitadas tanto para imagens individuais como para um número maior de imagens através do uso de inteligência artificial (IA). Por exemplo, NVIDIA Image Inpainting (**NVIDIA, 2020**) é uma ferramenta simples para remover pessoas ou objetos de fotos. **A Figura 2** mostra a imagem original com Alexander Malchenko (esquerda), ao lado de uma segunda versão (direita) onde marcamos Malchenko para remoção pela ferramenta NVIDIA Inpainting. O esforço exigido pelo usuário é muito pequeno. Primeiramente, deve-se especificar a parte da imagem a ser removida. Em seguida, o sistema determina como preencher a lacuna criada analisando a imagem restante.

Figura 2

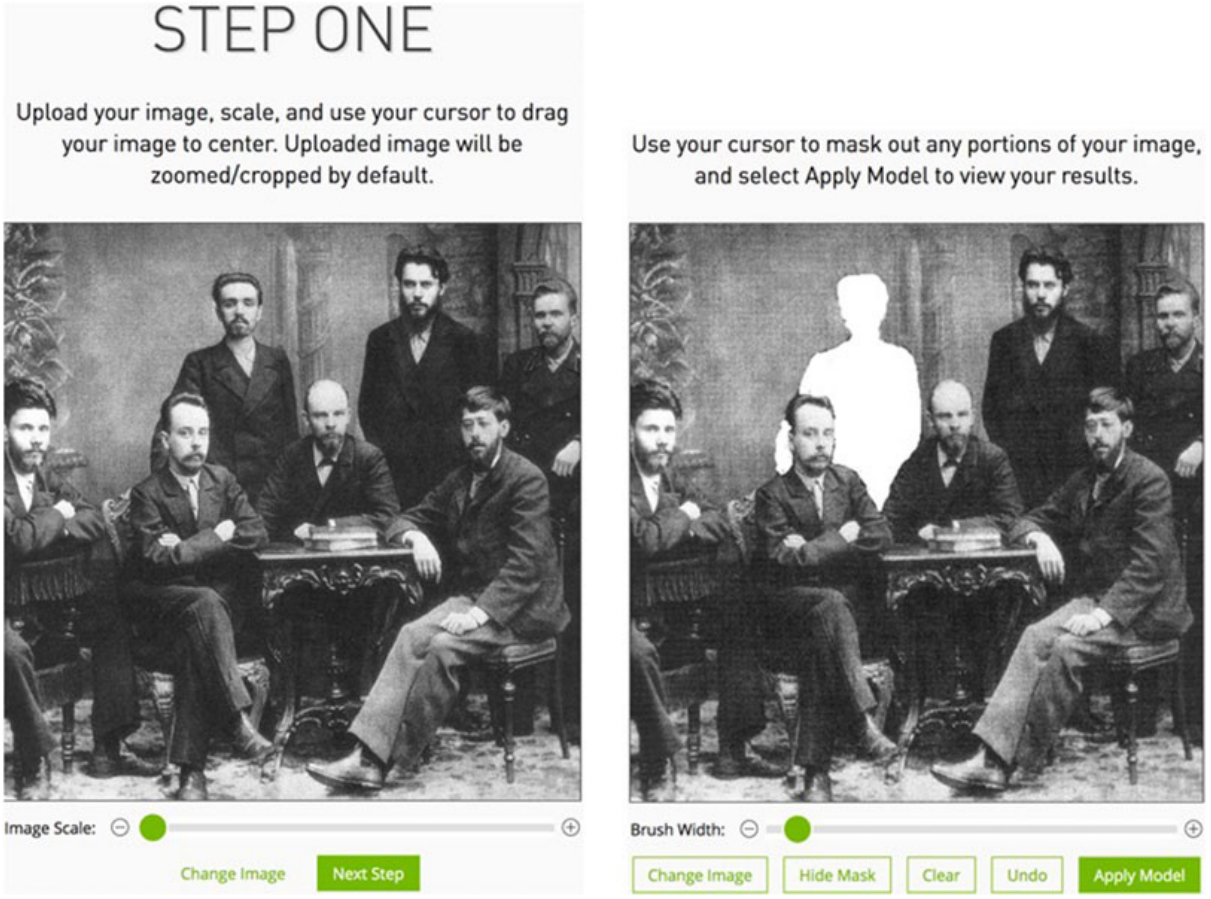

FIGURA 2 . Operação de pintura de imagem NVIDIA. **ESQUERDA:** seleção de imagens. **DIREITA:** Seleção do conteúdo a ser removido. Fonte: Criado pelos autores usando NVIDIA InPaint (**NVIDIA, 2020**). Imagem da Wikipédia (**Wikipedia, 2020**).

O resultado é mostrado na **Figura 3** . A manipulação não é perfeita, principalmente pelo fato de a parte da imagem especificada para remoção ser substituída por um fundo artificial criado do zero. Existem serviços comerciais que desempenham essencialmente a mesma função, mas proporcionam melhores resultados. No entanto, fazem uso do julgamento humano e cobram pela manipulação de cada imagem (**Business Insider, 2017**), o que limita esta abordagem a um pequeno número de imagens. De maneira semelhante, em vez de remover uma pessoa, também é possível substituí-la por uma pessoa diferente, normalmente substituindo o rosto. Ferramentas automatizadas que podem fazer isso, por exemplo, um programa chamado Reflect da NEOCORTEXT (**Reflect, 2020**), estão disponíveis, mas apresentam limitações semelhantes ao trabalhar em imagens individuais.

Figura 3

FIGURA 3 . Imagem de saída do NVIDIA Image Inpainting. O resultado mostra alguns sinais de manipulação. Fonte: Wikipédia. Criado pelos autores usando NVIDIA Inpainting (**NVIDIA, 2020**). Imagem da Wikipédia (**Wikipedia, 2020**).

Por outro lado, é muito mais fácil para os sistemas de IA remover uma pessoa de um vídeo se ela estiver em movimento. Nesse caso, o vídeo conterá imagens de fundo, e tudo o que a IA precisa fazer é pegar o fundo correto de um quadro de vídeo e colocá-lo sobre a pessoa em um quadro diferente, removendo efetivamente a pessoa. **Figura 4**mostra um exemplo disso. Foi criado usando Inpainting a partir de um vídeo que mostra duas pessoas passando. Como todo o fundo estático está contido no vídeo, as pessoas podem ser removidas perfeitamente, ou seja, um observador humano não detectaria tal manipulação. Assim, as informações adicionais do vídeo facilitam a criação automática de uma manipulação verossímil. Uma manipulação semelhante seria possível com múltiplas fotos da mesma cena de ângulos diferentes.

Figura 4

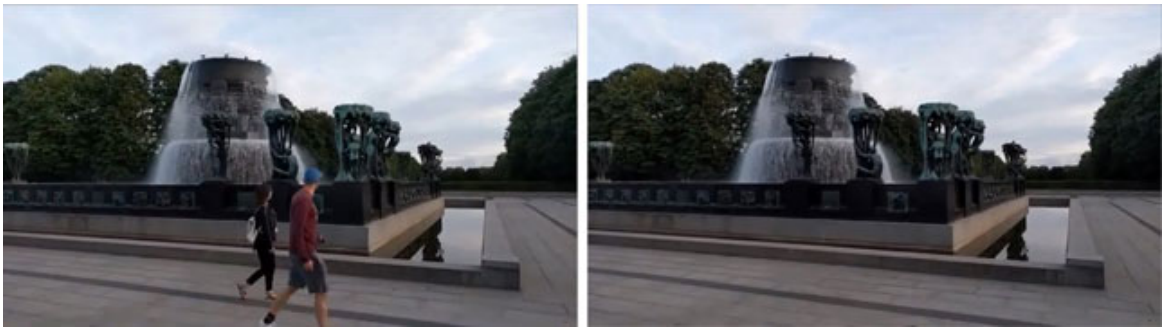

FIGURA 4 . Remoção de pessoas de um vídeo. A manipulação da imagem à direita não é visível aqui. Fonte: Still de vídeo criado pelos autores utilizando NVIDIA Inpainting (**NVIDIA, 2020**).

Na década de 1940, o número total de imagens tiradas em todo o mundo era comparativamente baixo. Assim, o número de imagens que documentavam um único evento era tipicamente tão pequeno que, ao manipular fotos individuais, era possível estabelecer uma narrativa contrafactual. Por outro lado, hoje a manipulação de uma imagem individual tem pouco peso para fins políticos. Com a proliferação da fotografia digital e dos smartphones transportados constantemente, o número de fotografias tiradas por ano explodiu para cerca de 1,2 biliões de fotografias tiradas em 2017 (**Business Insider, 2017**). Assim, para eventos importantes, normalmente já não é possível estabelecer uma narrativa através da censura ou manipulação de imagens individuais quando existem milhares de imagens originais. Fazer isso exigiria a manipulação automática de um grande número de imagens. Por outro lado, se um grande número de imagens ou vídeos de um evento estiver disponível, torna-se possível criar automaticamente manipulações críveis usando inteligência artificial, assumindo que a maior parte do material pode ser acessada e alterada. Versões simples de tais sistemas têm sido usadas há algum tempo para, por exemplo, censurar conteúdo pornográfico em plataformas online (**Gorwa et al., 2020**). Da mesma forma, o Google Street View detecta e desfoca automaticamente rostos e placas de veículos por motivos de privacidade. Mas a mesma tecnologia poderia ser usada para fins muito mais maliciosos, como remover ou substituir automaticamente pessoas ou eventos de todos os documentos de vídeo acessíveis.

Assim, desde que a mesma entidade tenha acesso a muitas ou a todas as imagens de um evento, poderá utilizar IA avançada para manipular ou censurar a maioria das imagens tiradas de um determinado evento. Pelo fato de hoje a maioria das fotos serem tiradas por smartphones permanentemente conectados a servidores de computação em nuvem, em um país que possui controle central da internet isso é tecnologicamente viável. Segundo a Universidade de Hong Kong, algumas variantes desta tecnologia foram implementadas na China (**Universidade de Hong Kong, 2020**). No entanto, muito poucos outros governos têm tanto a capacidade tecnológica como o direito constitucional para implementar tal sistema. Assim, a seguir nos concentramos em vídeos feitos por indivíduos ou pequenas organizações, e não por atores estatais.

Manipulação de Vídeos

A manipulação de vídeo tradicionalmente requer muito mais esforço e habilidade técnica do que a manipulação de fotos. O grande número de frames (isto é, imagens) que precisam ser manipulados, juntamente com a necessidade de consistência na manipulação, criam altas barreiras tecnológicas para uma manipulação de vídeo bem-sucedida. No entanto, tal como no caso da manipulação de fotografias, o custo de o fazer pode ser bastante reduzido devido à utilização de novas técnicas de aprendizagem automática.

Até recentemente, a manipulação direta do material de vídeo era rara, e uma forma mais comum de desinformação baseada em vídeo era a rotulagem incorreta, ou seja, vídeos que

afirmam mostrar algo diferente do que realmente mostram, ou edição sugestiva, ou seja, cortar material de vídeo autêntico. de tal forma que deturpa a situação filmada (**Matatov et al., 2018**). Para separar essas técnicas dos deepfakes, foi sugerido o termo deepfakes (**European Science Media Hub, 2019**). Um terceiro caso é a manipulação do conteúdo real do vídeo em um nível que não requer IA. Esses vídeos são às vezes chamados de Cheapfakes. Um vídeo que mostra um político que foi recodificado em velocidade reduzida, dando assim a impressão de fala arrastada, é o exemplo mais conhecido de uma falsificação barata (**Donovan e Paris, 2019**). Embora tais técnicas possam ser muito eficazes, elas não são novas e, portanto, não serão discutidas aqui, pois nos concentramos apenas na manipulação de conteúdo de vídeo baseada em IA.

Em grande parte, a tecnologia para manipulação direta de material de vídeo foi desenvolvida para e pela indústria cinematográfica, principalmente em Hollywood. Os marcos incluem Jurassic Park (1993), que adicionou dinossauros críveis gerados por computador às cenas filmadas, e Forest Gump (1995), onde Tom Hanks é inserido em imagens históricas de John F. Kennedy. Avatar (2009) mostrou que com um orçamento grande o suficiente, quase tudo pode ser levado para a tela. No entanto, com exceção da publicidade de campanha política, que não se enquadra na maioria das definições padrão de desinformação, existem muito poucos casos conhecidos de utilização desta tecnologia para propaganda. Uma razão provável é o facto de o custo de implementação desta tecnologia ter sido muito elevado. Filmes contendo um grande número de efeitos CGI de alta qualidade normalmente custam mais de US$ 1 milhão por minuto de filmagem. No entanto, isto mudou radicalmente devido à introdução de métodos baseados em IA para manipulação de vídeo. A seguir discutiremos os métodos que tornam isso possível.

Tecnologia de Deepfakes

Em 2017, usuários da plataforma online Reddit apresentaram vídeos de celebridades cujos rostos foram trocados por rostos de pessoas diferentes. Embora o efeito fosse novo, o software que criou essas imagens dependia de uma tecnologia desenvolvida alguns anos antes.

Num artigo marcante de 2012, a aprendizagem profunda, um refinamento das redes neurais artificiais, foi estabelecida como uma tecnologia superior para reconhecimento de imagens (**Krizhevsky et al., 2012**). A partir daí, um imenso conjunto de trabalhos surgiu nos últimos anos, propondo tanto refinamentos do método quanto extensões para outras áreas de aplicação. Além disso, este desenvolvimento teve um impacto considerável fora da comunidade científica e trouxe o tema da inteligência artificial à atenção da política, da indústria e da mídia (**Witness Lab, 2020**).

Embora os fundamentos matemáticos sejam conhecidos há décadas, o artigo de 2012 demonstrou que, quando treinadas com um grande número de imagens de entrada adequadas, as redes neurais convolucionais (CNNs) podem categorizar o conteúdo de uma imagem com alta precisão. A chave para isso reside na representação abstrata de um tipo de objeto (por exemplo, uma cadeira) nos níveis superiores da rede neural profunda, cuja

estrutura se assemelha à do córtex visual no cérebro humano. Uma CNN pode ser treinada para reconhecer pessoas específicas e diferenciá-las com segurança em uma ampla gama de imagens. Os pré-requisitos para fazer isso são um computador poderoso e um grande número de imagens com as quais uma CNN aprende.

Uma vez treinadas, as CNNs e outras redes neurais podem ser invertidas. Isso é feito especificando uma saída e, em seguida, executando as operações matemáticas inversas para cada camada na ordem inversa. A rigor, nem todas as operações podem ser invertidas, mas isso não limita a aplicabilidade do conceito. O resultado é uma imagem criada a partir dos recursos abstratos que a rede aprendeu. A camada de entrada original atua então como camada de saída. Ele produz uma imagem com a mesma resolução das imagens que foram originalmente usadas para treinar a CNN. Uma rede neural requer uma resolução de imagem fixa, mas as imagens podem ser facilmente dimensionadas. Chamamos uma CNN que foi treinada para reconhecer uma pessoa específica de rede de detectores, e a versão invertida dela de rede geradora. A tecnologia deepfake depende da combinação dos dois tipos.

Autoencoders para Deepfakes

É possível vincular um detector e seu gerador correspondente. Tal sistema é chamado de autoencoder, porque aprende a codificar uma imagem – no nosso caso, o rosto de uma pessoa – de alguma forma abstrata como resultado de seu treinamento. Com dados de treinamento suficientes, a rede também pode reconhecer um rosto em uma imagem com ruído ou em um ângulo incomum. Por representar a face internamente de forma abstrata, ele pode gerar a face sem ruído usando a rede do gerador.

Deepfakes para troca de rosto são criados treinando dois desses autoencoders. Uma rede é treinada para reconhecer a pessoa alvo cujo rosto será substituído pelo da pessoa fonte, enquanto a outra rede é treinada para reconhecer a pessoa fonte. Em seguida, o detector da pessoa fonte é vinculado ao gerador da pessoa alvo, criando assim um novo autoencoder. Quando aplicado a um vídeo da pessoa fonte, o resultado é que o rosto da pessoa alvo aparece em vez do rosto da pessoa fonte. O projeto é mostrado na **Figura 5** . Isto cria a impressão de que a pessoa alvo está fazendo tudo o que a pessoa fonte estava fazendo no vídeo de entrada, criando assim um potencial substancial para manipulação e desinformação.

Figura 5

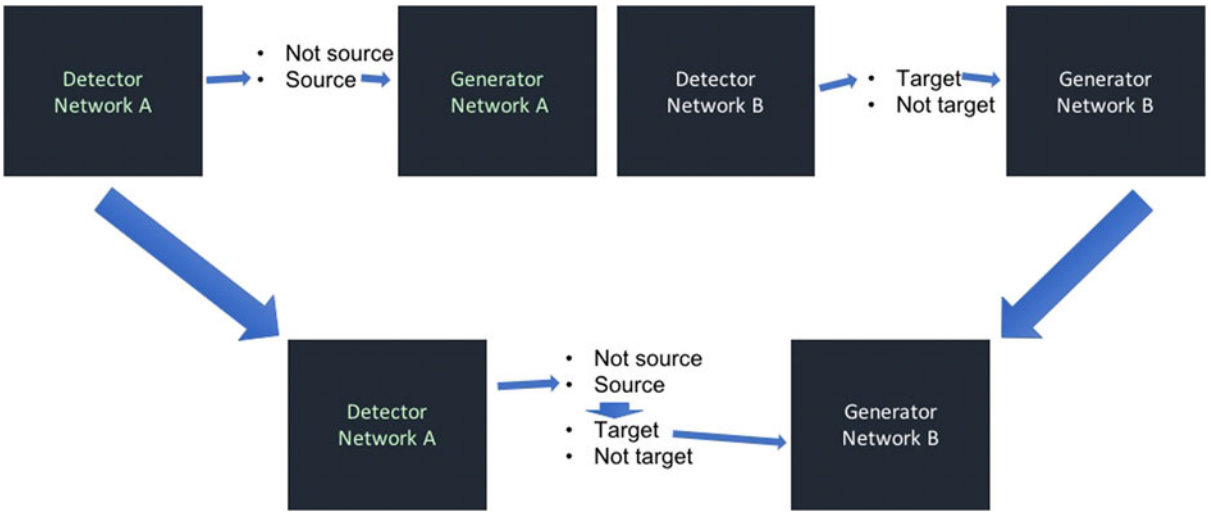

FIGURA 5 . Tecnologia básica para a criação de deepfakes. A Rede A é treinada para reconhecer a pessoa fonte mesmo em imagens distorcidas ou com ruído, e configurada de forma que produza uma imagem silenciosa da pessoa alvo. A Rede B é treinada na pessoa alvo da mesma maneira. Após o treinamento, a saída do detector da pessoa fonte é alimentada como entrada para o gerador da pessoa alvo. Assim, sempre que a pessoa fonte for detectada num vídeo de entrada, ela será substituída pela pessoa alvo.

Visão geral do software Deepfakes

A tecnologia descrita acima tornou-se disponível publicamente em 2017. Os programas que a implementam formaram a primeira geração de software deepfake, que foi seguida por mais duas gerações de crescente sofisticação. Discutiremos as gerações de software aqui. Os programas individuais são apresentados no **Apêndice Suplementar** .

Conforme discutido acima, a primeira geração de software deepfake exigia um grande número de imagens de treinamento para funcionar corretamente. Conseqüentemente, esses programas são impraticáveis para a criação de vídeos manipulados de uma pessoa comum. Por esse motivo, a maioria dos vídeos deepfake criados para fins de entretenimento apresentavam atores famosos, dos quais muitas imagens estão disponíveis publicamente. O software de primeira geração é descrito em detalhes no **Apêndice Suplementar** A.

Porém, a segunda geração de software deepfake não possui mais essa restrição. Isto se deve ao uso de redes adversárias generativas (GANs) (**The Verge, 2020**). GANs são um tipo de rede neural semelhante aos autoencoders. No entanto, em uma GAN, as redes de detectores e geradores trabalham uma contra a outra. A tarefa do gerador é criar variantes de imagens semelhantes, mas não idênticas, às entradas originais, adicionando ruído aleatório. A partir dessas imagens geradas, a rede de detectores, também chamada de discriminador neste contexto, é treinada conforme mostrado na **Figura 6** . Desta forma, a rede discriminadora torna-se muito boa no reconhecimento de variantes da mesma imagem, como um rosto visto de muitos ângulos diferentes, mesmo que haja poucas imagens originais para treinar.

Figura 6

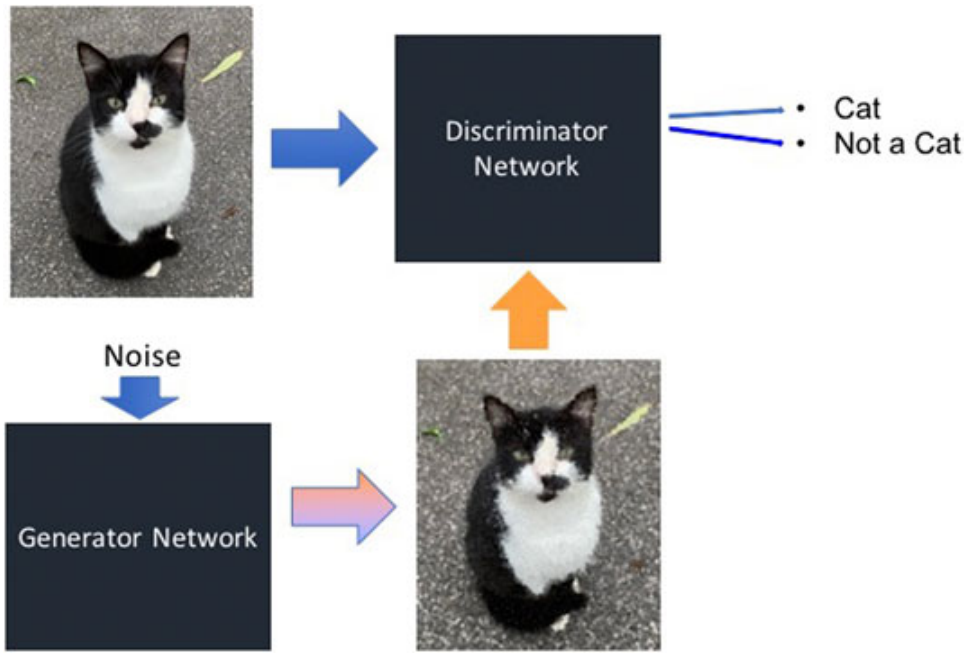

FIGURA 6 . Configuração básica de uma rede adversária generativa. O gerador cria imagens adicionando ruído aleatório às imagens de entrada, enquanto a rede discriminadora alimenta tanto as imagens de entrada originais quanto as imagens geradas. Fonte: Imagem criada pelos autores a partir de foto tirada pelos autores.

Observe que, diferentemente da primeira geração, que contém software fácil de usar, a maior parte da segunda geração de geradores de deepfake são códigos de pesquisa que exigem considerável habilidade técnica para serem usados com sucesso. No entanto, certamente seria possível criar software amigável a partir deles. Os programas mais proeminentes que utilizam este conceito para a segunda geração de software deepfake são descritos em detalhes no Apêndice **Suplementar B.**

Normalmente, a criação de um vídeo manipulativo requer a criação de uma trilha de áudio manipulativa. No entanto, comparado ao vídeo, a manipulação do áudio é uma tarefa bastante simples. Na verdade, em muitos sistemas judiciais é significativamente mais difícil estabelecer gravações de áudio como prova (**Al-Sharieh e Bonnici, 2019**). Além disso, a qualidade de áudio imperfeita pode sempre ser mascarada para aparecer como uma falta de qualidade do microfone na gravação original. Assim, para criar um deepfake convincente, a manipulação de áudio é um desafio menor. No entanto, o efeito pode ser considerável, uma vez que o áudio manipulado tem sido utilizado com sucesso para o crime cibernético (**Stupp, 2019**). Dois programas que podem gerar áudio para deepfake são descritos no **Apêndice Suplementar** C.

A última geração de software deepfake baseia-se na segunda geração, mas amplia suas capacidades de várias maneiras. Aumenta significativamente a complexidade do modelo, combina múltiplas redes de geradores em um modelo e extrai recursos de séries temporais vinculadas. Isso permite trabalhar em um espaço de recursos faciais em vez de um espaço de quadro 2D e simular alterações de quadro a quadro de maneira natural. Eles geralmente

incluem áudio diretamente para gerar movimentos naturais dos lábios. Apresentamos os dois pacotes de software mais importantes no **Apêndice Suplementar** D.

Contramedidas

Dada a ameaça potencial dos deepfakes, foram considerados diferentes tipos de contramedidas contra eles. Discutimos contramedidas técnicas e legais aqui.

Software de detecção

Reagindo à ameaça representada pelas mídias visuais manipuladas, a Agência de Projetos de Pesquisa Avançada de Defesa (DARPA) dos Estados Unidos estabeleceu o projeto forense de mídia com o objetivo de desenvolver ferramentas para reconhecer a manipulação em vídeos e imagens (Darpa, 2020), utilizando **uma** ampla variedade de ferramentas, como análise semântica. Nesse contexto, a Adobe, criadora do popular software Photoshop, anunciou um programa capaz de detectar a maioria das manipulações de imagens que o Photoshop é capaz (**Adobe Communications Team, 2018**). A tecnologia baseia-se em pesquisas de longo prazo na Universidade de Maryland (**Zhou et al., 2018**). Além disso, grandes conjuntos de dados para treinamento e teste de algoritmos de detecção tornaram-se recentemente disponíveis (**Dolhansky et al., 2019**; **Guan et al., 2019** ; **Rossler et al., 2019**).

Muitas ideias para detectores de deepfake baseados em IA foram testadas em um desafio competitivo organizado pelo Facebook em 2020 (**Dolhansky et al., 2020** ; **Ferrer et al., 2020**). O desafio resultou em diversas novas abordagens para detectores (**Mishra, 2020**). Uma visão geral sobre o campo da análise forense de mídia em relação às notícias falsas foi recentemente apresentada por Verdoliva (**Verdoliva, 2020**).

Embora estas abordagens sejam promissoras, o seu sucesso depende, em última análise, do seu modo de implementação. Além disso, pesquisas recentes (**Gandhi e Jain, 2020** ; **Neekhara et al., 2020**) utilizando estratégias adversárias (**Goodfellow et al., 2014**) indicam que mesmo os melhores detectores atuais podem ser enganados. As estratégias adversárias consistem em adicionar ruído a um vídeo ou imagem. Esse ruído é imperceptível ao olho humano, mas é suficiente para confundir um detector de notícias falsas.

Assim, é provável que muitos destes sistemas apresentem falhas na sua aplicação prática porque não oferecem uma precisão de detecção de 100% e, se estiverem disponíveis ao público em geral, também estarão disponíveis aos criadores de desinformação. Mesmo uma precisão de detecção de 99% significa que o sistema pode ser derrotado de alguma forma. Para o reconhecimento de imagens de última geração, sabe-se que existe um grande número de exemplos adversários, ou seja, pares de imagens de entrada que serão reconhecidas como o mesmo objeto pelos humanos, mas como objetos muito diferentes pelas redes neurais (**Szegedy et al., 2013** ; **Gu e Rigazio, 2014**). Assim, um detector de manipulação construído com a mesma tecnologia apresentaria as mesmas fraquezas e poderia, portanto, ser derrotado com alterações mínimas no conteúdo manipulado. **A Figura 7** ilustra a situação.

Figura 7

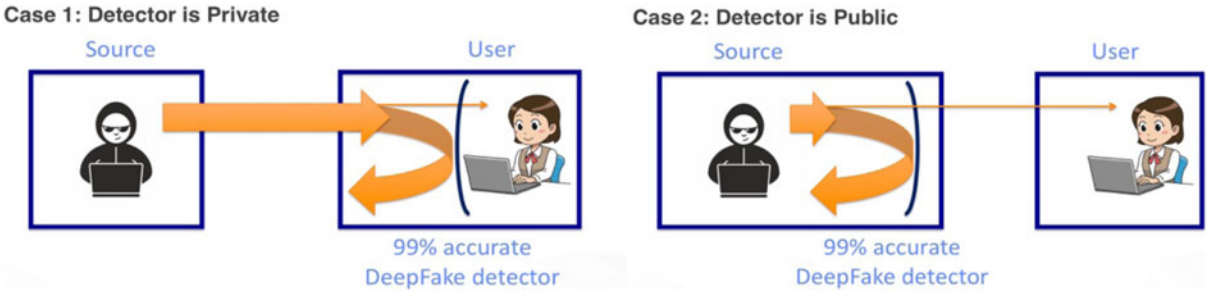

FIGURA 7 . Dois cenários diferentes para implantação de filtros de manipulação. **ESQUERDA:** o cenário assumido é que o usuário tenha um detector com 99% de precisão que filtra quase todo o conteúdo manipulado. **CERTO:** num cenário realista, a fonte da desinformação tem acesso ao mesmo detector. Consequentemente, apenas o conteúdo manipulado que passa pelo detector é divulgado ao público, tornando o detector quase inútil. Fonte: Imagem criada pelos autores do MS ClipArt.

O problema é semelhante ao do software de detecção de vírus (**Fedler et al., 2013** ; **Rastogi et al., 2013**). Nas últimas três décadas, o problema do malware não foi resolvido pelo uso de software de detecção executado nos sistemas dos usuários. Da mesma forma, é improvável que a situação seja diferente para os meios de comunicação manipulados. A declaração a seguir descreve concisamente o fenômeno: “Atualmente é trivial para os autores de malware alterar ligeiramente o malware existente, com o efeito de que ele não será detectado pelo software antivírus até que novas assinaturas sejam lançadas”. (**Fedler et al., 2013**). Para a detecção de malware e spam, instâncias centralizadas que não podem ser acessadas e testadas de maneira ilimitada pelos criadores de malware tiveram algum sucesso. Por exemplo, o software em smartphones normalmente pode ser instalado apenas através de canais controlados (Apple App Store, Google Play), e as medidas para contornar esta limitação são geralmente restritas a usuários tecnicamente experientes. O malware confirmado é removido regularmente desses canais. Esta abordagem pode ser considerada comparativamente bem-sucedida. Uma razão provável para isto é que os fornecedores de sistemas operativos para smartphones têm um forte incentivo para manter o malware longe dos seus sistemas e, assim, fornecer valor aos seus clientes.

Contudo, para a distribuição de notícias, um sistema centralizado controlado por uma empresa não é claramente uma solução desejável. A marca registrada de um cenário mediático saudável numa sociedade democrática é uma seleção diversificada de fontes independentes de notícias. Assim, o controlo central sobre o que é “verdade”, mesmo na forma benigna de detecção de manipulação, acarreta um imenso risco de abuso. Além disso, tal caso estaria provavelmente sujeito a pressão política, da mesma forma que as redes sociais como o Facebook ou o Twitter são frequentemente sujeitas a pressão de intervenientes privados ou estatais para removerem certos tipos de conteúdos dos seus serviços. Assim, uma solução preferível seria aberta, descentralizada e acessível a todos. O desafio técnico reside na construção de uma solução que seja, no entanto, eficaz.

Além do software de detecção, existem também outras abordagens tecnológicas, como o News Provenance Project do New York Times (**Koren, 2019**) que faz uso de blockchains para garantir que um vídeo não foi adulterado (**Hasan e Salah, 2019b**) . Como são baseados em criptografia e não em inteligência artificial, eles não sofrem dos problemas de precisão discutidos acima. No entanto, eles precisariam de ampla adoção para serem eficazes.

Abordagens legais e resiliência

Além das soluções tecnológicas, pode haver abordagens legais para resolver o problema do abuso de software deepfake. Poder-se-ia argumentar que a proibição da tecnologia resolveria o problema, mas uma abordagem tão severa tornaria ilegais uma grande parte da sátira, do humor, dos memes da Internet e possivelmente até dos filmes, sendo, portanto, inconcebível numa sociedade democrática. Nem a disseminação de imagens falsas nem o uso de software deepfake são inerentemente prejudiciais, e existem muitos aplicativos benignos. Pela mesma razão, é bastante improvável que as empresas de redes sociais proíbam totalmente a sua utilização ou distribuição (**Chesney e Citron, 2019b**). Ainda assim, há um impulso legislativo notável nos Estados Unidos no sentido de restringir o uso prejudicial de deepfakes (**Ruiz, 2020**).

No entanto, a experiência com novas leis contra o cyberbullying, a perseguição cibernética, o assédio online e o discurso de ódio, etc., demonstrou que a introdução de novas leis ou a alteração das leis existentes para lidar com os deepfakes pode levar a todo um novo conjunto de problemas, como auto-execuções apressadas. censura por parte de empresas de mídia social para evitar multas (**Chesney e Citron, 2019b**). Uma vez que a intenção de um deepfake é mais importante do que a própria tecnologia, pode ser possível lidar com falsificações aplicando leis já existentes contra fraude, assédio ou outras formas de declarações difamatórias (**Brown, 2020**).

Considerando as limitações das contramedidas tecnológicas e legais, parece que, antes de mais nada, as sociedades precisam desenvolver resiliência aos deepfakes, conforme formulado por Robert Chesney e Danielle Citron (**Chesney e Citron, 2019b**):

Em suma, as democracias terão de aceitar uma verdade incómoda: para sobreviverem à ameaça dos deepfakes, terão de aprender a conviver com mentiras.

Casos Deepfake conhecidos

Até o momento, todo um subgênero de vídeos de paródia deepfake se estabeleceu em plataformas de vídeo como o YouTube, sendo a troca dos atores principais de um filme a manipulação mais comum. Por exemplo, um curta-metragem chamado Home Stallone (**Face, 2020**) - uma referência ao filme de comédia de 1990, Home Alone - apresenta Sylvester Stallone gerado por IA de 8 anos em vez do ator real, Macaulay Culkin. Muitos vídeos semelhantes estão disponíveis na internet.

Em 2020, vários casos de vídeos deepfake com motivação política apareceram nos noticiários. Em 7 de fevereiro de 2020, durante as eleições para a Assembleia Legislativa em

Deli, surgiram duas mensagens de vídeo, mostrando o líder do Partido Delhi Bharatiya Janata (BJP), Manoj Tiwari, dirigindo-se aos seus potenciais eleitores em inglês e Haryanvi. Segundo a VICE Índia, os vídeos foram compartilhados em mais de 5.800 grupos de WhatsApp, atingindo cerca de 15 milhões de pessoas (**Vice, 2020**). Surpreendentemente, Manoj Tiwari não fala Haryanvi, nem nunca gravou a mensagem de vídeo em inglês ou qualquer outro idioma. Uma empresa indiana de relações públicas chamada “The Ideaz Factory” gravou uma mensagem de vídeo mais antiga de Manoj Tiwari, onde ele falava sobre um assunto totalmente diferente – em hindi. Em seguida, treinaram uma IA com vídeos de Manoj Tiwari falando, até conseguir sincronizar os lábios com vídeos arbitrários dele. Em seguida, eles usaram um dublador para gravar o inglês e o Haryanvi e fundiram áudio e vídeo (**Khanna, 2020**).

Em 14 de abril de 2020, a hashtag #TellTheTruthBelgium chamou a atenção da mídia. Um vídeo mostrou um discurso de quase 5 minutos da primeira-ministra belga, Sophie Wilmès, retratando a pandemia da COVID-19 como consequência da destruição ambiental. O movimento ambientalista Extinction Rebellion usou tecnologia deepfake para alterar um endereço passado à nação que Sophie Wilmès manteve anteriormente (**Extinction Rebellion, 2020** ; **Galindo, 2020**).

Outro exemplo é um apelo dirigido ao presidente mexicano Andrés Manuel López Obrador. No vídeo publicado em 29 de outubro de 2020, o escritor e jornalista mexicano Javier Valdez insta o presidente López Obrador e a sua administração a lutarem mais contra a corrupção e o crime organizado. Javier Valdez foi assassinado em 15 de maio de 2017, como explica seu alter ego digital no início do vídeo. Presumivelmente, ele foi morto em consequência de suas investigações sobre o crime organizado. Durante 1 minuto e 39 segundos, o “Programa de Defesa de Vozes para a Segurança dos Jornalistas” trouxe o Sr. Valdez de volta à vida usando tecnologia deepfake para exigir justiça para jornalistas mortos e desaparecidos.**Repórter Ohne Grenzen, 2020**).

Embora muitas aplicações da tecnologia deepfake sejam humorísticas ou benignas, elas abrigam uma possibilidade inerente de uso indevido. Os exemplos acima não são intencionalmente maliciosos, o caso de Manoj Tiwari mostra que é de facto possível enganar os eleitores através da utilização de meios de comunicação sintéticos. Os outros dois exemplos foram publicados acompanhados de um aviso de isenção de responsabilidade, informando que foi utilizada tecnologia deepfake. Tem sido amplamente suspeitado que deepfakes seriam usados para influenciar as eleições presidenciais de 2020 nos Estados Unidos. No entanto, no momento em que este artigo foi escrito, a manipulação de vídeo dependia mais de técnicas convencionais (**Politifact, 2020** ; **The Verge, 2020**). Em vez disso, foram divulgados vídeos deepfake que alertam sobre ameaças à democracia (**Technology Review, 2020**). Assim, até agora parece que as invenções descaradas raramente chegam à esfera pública onde podem ser desmascaradas, mas é possível que tais vídeos circulem em grupos fechados com a intenção de mobilizar apoiantes. Os vídeos que envolvem políticos estrangeiros parecem ser especialmente viáveis, uma vez que é mais difícil para as pessoas julgarem se o comportamento retratado é credível para essa pessoa (**Schwartz, 2018**).

Por enquanto, os deepfakes não visam principalmente a esfera política. A startup holandesa Sensity. ai (**Sensity, 2020**), que rastreia e conta vídeos deepfake disponíveis na internet, relata que mais de 85% de todos os vídeos deepfake têm como alvo celebridades femininas nas indústrias de esportes, entretenimento e moda. Alguns destes constituem casos da chamada pornografia involuntária, em que o rosto da pessoa alvo é colocado num vídeo de origem pornográfica. Inicialmente, isso exigia grandes quantidades de imagens da pessoa-alvo e, portanto, apenas celebridades eram alvos. No entanto, a recente tecnologia deepfake baseada em redes adversárias generativas torna possível atingir pessoas das quais existe apenas um pequeno número de imagens.

Assim, considerando a ampla disponibilidade desta tecnologia, o uso de deepfakes para cyberbullying tornou-se uma ameaça relevante. Tal evento, denominado escândalo Nth Room (**International Business Times, 2020**), aconteceu na Coreia do Sul em 2019. O evento envolveu a produção e distribuição de vídeos pornográficos deepfake usando rostos de celebridades femininas, juntamente com outras práticas de exploração. Da mesma forma, uma extensão para o popular aplicativo de mensagens Telegram, chamada DeepNude (**Burgess, 2020**), tornou-se disponível em 2019 e reapareceu em 2020. Dada uma fotografia da pessoa alvo, substitui essencialmente o corpo vestido por um corpo nu, criando assim a semelhança de uma fotografia nua da pessoa alvo. O programa é relativamente pouco sofisticado e aparentemente só foi treinado para trabalhar com mulheres brancas. No entanto, a sua concepção elimina essencialmente todas as barreiras restantes à criação de conteúdos potencialmente nocivos. Por depender de servidores externos, é concebível que programas mais sofisticados que forneçam imagens ou vídeos de alta qualidade apareçam no futuro.

Uma aplicação diferente da tecnologia deepfake é a representação de outros indivíduos na comunicação online. Em 2019, os criminosos usaram manipulação de áudio para se passar por um CEO em uma ligação telefônica e ordenar que um funcionário transferisse uma grande quantia de dinheiro para uma conta privada (**Stupp, 2019**). Recentemente, foi lançado um software que permite a personificação em videochamadas utilizando tecnologia deepfake (**Siarohin et al., 2019**). É evidente que tal tecnologia tem uma vasta gama de possíveis aplicações criminosas, sendo a mais preocupante entre elas a utilização por predadores sexuais para se passarem por menores (**Foster, 2019** ; **Hook, 2019**).

Desenvolvimento futuro projetado

Vimos que a tecnologia deepfake tem o potencial de criar vídeos gravemente prejudiciais. Por outro lado, até ao momento, o número de casos em que isso aconteceu é limitado. Isto pode ser devido à novidade da técnica. Nesse caso, devemos esperar um aumento maciço de vídeos deepfake num futuro próximo. Contudo, também é possível que substituam a desinformação tradicional apenas num número bastante limitado de situações.

Para obter uma compreensão mais profunda dos prováveis casos futuros, é necessário considerar as diferenças entre deepfakes e outras desinformações. Deepfakes são de natureza visual e contêm uma grande quantidade de informações. Isto pode aumentar a

credibilidade, mas também pode fornecer mais detalhes que desacreditam a desinformação, tais como artefactos visíveis.

A pandemia da COVID-19 foi o acontecimento noticioso dominante do ano de 2020 e foi acompanhada por uma grande quantidade de desinformação. Na primavera de 2020, o Grupo de Trabalho East StratCom do Serviço Europeu de Ação Externa salientou que “em tempos de pandemia a desinformação médica pode matar” (**UE vs DisInfo, 2020**). No entanto, até onde sabemos, não há casos impactantes de vídeos deepfake espalhando desinformação sobre a COVID-19.

Como o vírus é invisível, é difícil espalhar desinformação visual sobre ele, como a afirmação de que o COVID-19 é causado pela radiação da rede sem fio 5G (**Temperton, 2020**) (que também é invisível). A desinformação 5G-COVID foi predominantemente espalhada por vídeos do YouTube que mostram pessoas que falam sobre tais teorias de conspiração. É evidente que a eficácia dessa desinformação depende da credibilidade e da capacidade de persuasão dos oradores, que não podem ser reforçadas por deepfakes. No entanto, é possível criar imagens de um orador confiável para espalhar desinformação. Até o momento, são poucos os casos registrados. Resta saber se isto se tornará um vector significativo de disseminação de desinformação.

No entanto, os deepfakes são mais adequados para espalhar desinformação nos casos em que as gravações de vídeo são normalmente eficazes. Tais situações retratam ações de pessoas reconhecíveis, cujo significado é compreensível sem muito contexto específico e que provocam uma resposta emocional. A razão para essas condições é baseada no que a tecnologia deepfake é capaz de fazer. A troca de rostos não é necessária se as pessoas envolvidas não forem indivíduos reconhecíveis (por exemplo, polícias ou soldados uniformizados), embora seja concebível que a tecnologia deepfake seja utilizada na futura propaganda de guerra, substituindo uniformes ou marcas de nacionalidade. Isto pode ser ainda mais relevante para guerras civis e conflitos envolvendo grupos militantes que não são exércitos formais (**Camarões, 2020**).

Além disso, para provocar uma resposta racional, a desinformação textual pode muitas vezes ser igualmente eficaz e é muito mais fácil de produzir em alto volume, mas o vídeo é eficaz na criação de uma resposta emocional (Nelson-Field et al., 2013 **)** .

Além disso, como os vídeos deepfake podem conter pistas visuais de que não retratam a realidade, podem até ser menos eficazes do que o texto. Assim, é improvável que o ganho do deepfake produzido para esse fim supere o custo. Por outro lado, é provável que as respostas emocionais sejam amplificadas ao ver um vídeo. E para provocar uma resposta emocional, o próprio vídeo deve levar os espectadores a essa resposta, em vez de depender de explicações adicionais. Assim, podemos formular cinco condições que caracterizam os deepfakes eficazes para a manipulação política. Esperamos que os futuros deepfakes mais significativos cumpram estas condições:

1. Envolva pessoas reconhecíveis

2. Retrate ações humanas

3. Pode ser entendido com contexto adicional limitado

4. Provoque uma resposta emocional

5. A resposta é significativa o suficiente para afetar opiniões ou incitar ações

Claramente, vídeos envolvendo nudez ou atividade sexual de indivíduos conhecidos preenchem essas condições, e tais incidentes aconteceram (**Ayyub, 2018** ; **Walden, 2019**). Embora o caso envolvendo um ministro da Malásia (**Walden, 2019**) aparentemente não foi baseado em um deepfake real, é claro que um vídeo deepfake real teria sido ainda mais eficaz se estivesse disponível para os perpetradores. Assim, é de esperar que casos semelhantes voltem a acontecer num futuro próximo. Contudo, uma vez que o efeito de tais campanhas de desinformação depende, em grande medida, das normas culturais, o efeito só pode ser restrito a partes do mundo. Outra área provável seria o comportamento violento ou alguma outra transgressão física das normas sociais por parte de um indivíduo conhecido. Por outro lado, os vídeos que mostram apenas uma pessoa conhecida a falar têm menos probabilidade de serem enganos eficazes, pelo menos não significativamente mais eficazes do que outras técnicas de desinformação.

Um exemplo de manipulação de vídeo que cumpre todos os critérios acima ocorreu na Alemanha em 2015, onde um vídeo de um gesto obsceno e a sua subsequente negação pelo ministro das Finanças grego, Yanis Varoufakis, causou alvoroço, com os principais jornais alemães escrevendo artigos hostis até que um programa de TV satírico o apresentador revelou que o vídeo era uma falsificação que ele criou (**The Guardian, 2015**). O vídeo foi manipulado profissionalmente e foi amplamente considerado genuíno. Embora o incidente tenha sido essencialmente benigno, com o objectivo de entretenimento e também de alertar o público contra julgamentos prematuros, mostrou que os vídeos manipulados podem aprofundar as tensões existentes entre os países. Usando a tecnologia deepfake, a capacidade de realizar tais manipulações torna-se disponível para quase todos.

Além do cyberbullying e da desinformação política, outra possível área em que os deepfakes poderão ser utilizados no futuro é a manipulação de mercado. No passado, mesmo os tweets contendo desinformação tiveram um impacto significativo nos preços de mercado. Por exemplo, em Julho de 2012, uma mensagem no Twitter alegou falsamente a morte do Presidente sírio Bashar al-Assad, fazendo com que os preços do petróleo bruto subissem temporariamente mais de 1 dólar/barril (Zero Hedge, 2020 **)** . Em 2013, uma declaração de exclusão feita pelo CEO da Abercrombie and Fitch, Mike Jeffries, em 2006, sobre o público-alvo da marca Abercrombie and Fitch ganhou ampla atenção e, como resultado, fez com que o valor da marca Abercrombie and Fitch despencasse (Bradford, 2017) **.**). Assim, um vídeo deepfake que mostre um CEO fazendo uma declaração socialmente inaceitável ou realizando ações inaceitáveis (por exemplo, assédio sexual) poderia ser usado para prejudicar o valor de uma empresa. Uma vez que é possível beneficiar de um evento deste tipo através dos mercados financeiros, é suficiente que um número suficiente de pessoas acredite na desinformação após a sua divulgação, mesmo que esta seja universalmente aceite como falsa mais tarde.

Conclusão

Vimos que, embora a manipulação de fotografias e até de vídeos tenha uma longa história, foram apenas os desenvolvimentos tecnológicos recentes que minaram a fiabilidade das provas de vídeo. O software deepfake moderno pode gerar facilmente vídeos que um observador casual pode considerar perfeitamente reais. Além disso, os sistemas automatizados podem ser facilmente derrotados se o invasor puder testar e, se necessário, modificar o vídeo manipulado. Assim, sem nenhuma solução técnica fácil à vista, e tal como acontece com a desinformação e as notícias falsas em geral, a educação e a literacia mediática continuam a ser a principal defesa contra a desinformação.

No entanto, o impacto dos deepfakes não deve ser superestimado. Por um lado, a desinformação é generalizada e a desinformação convencional tem-se revelado eficaz. Assim, mesmo que a criação de deepfakes seja barata, ainda exige mais esforço do que produzir textos. Consequentemente, apenas substituirão outros métodos de desinformação onde se possa esperar um efeito reforçado.

Por outro lado, só porque um vídeo criado artificialmente não pode ser distinguido de um vídeo gravado não significa que as pessoas possam ser levadas a acreditar que um evento arbitrário aconteceu. Hoje, os criadores de conteúdo de vídeo em plataformas de vídeo como YouTube, TikTok e Reddit competem por atenção. Devido a incentivos monetários significativos, os criadores de conteúdo são incentivados a apresentar vídeos dramáticos ou chocantes, o que por sua vez os leva a encenar situações dramáticas ou chocantes, e os vídeos encenados são frequentemente rotulados como tal pelos espectadores. Assim, os utilizadores regulares destas plataformas de vídeo estão cada vez mais conscientes dos vídeos manipulativos. Da mesma forma, o aumento da proliferação de vídeos deepfake acabará por tornar as pessoas mais conscientes de que nem sempre devem acreditar no que vêem.

Contribuições do autor

JL é o autor principal deste artigo. KP investigou e formulou as tendências de pesquisa em tecnologia deepfake. SB coletou numerosos exemplos e contribuiu com um trabalho substancial de edição. PF revisou o aspecto interdisciplinar do artigo e forneceu edição adicional. DS contribuiu com ideias adicionais.

Financiamento

Este trabalho foi financiado pelo projeto norueguês SAMRISK-2 “UMOD” (nº 272019). A pesquisa apresentada neste artigo beneficiou da Infraestrutura Experimental para Exploração de Computação Exascale (eX3), que é apoiada financeiramente pelo Conselho de Pesquisa da Noruega sob o contrato 270053.

Conflito de interesses

Os autores declaram que a pesquisa foi realizada na ausência de quaisquer relações comerciais ou financeiras que pudessem ser interpretadas como potencial conflito de interesses.

Material suplementar

O material complementar deste artigo pode ser encontrado online em: **https://www.frontiersin.org/articles/10.3389/fcomm.2021.632317/full#supplementary-material**

Referências

Adobe (2020). Photoshop. Disponível em: **https://www.adobe.com/products/photoshop.html** (acessado em 1 de abril de 2020).

Google Scholar

Equipe de comunicações da Adobe (2018). Detectando manipulação de imagens com IA. Disponível em: **https://theblog.adobe.com/spotting-image-manipulation-ai/** (acessado em 1 de abril de 2020).

Google Scholar

Al-Sharieh, S. e Bonnici, JM (2019). “PARE, você está diante das câmeras: a admissibilidade probatória e o valor probatório dos registros digitais na Europa”, em *Sinergia de policiamento comunitário e tecnologia* . Editores G. Leventakis e MR Haberfeld (Cham, Suíça: Springer), 41–52.

Texto completo CrossRef | **Google Scholar**

Ayyub, R. (2018). Fui vítima de uma trama pornográfica falsa que pretendia me silenciar. Disponível em: **https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316** (Acessado em 1 de abril de 2020).

Google Scholar

Beridze, I. e Butcher, J. (2019). Quando ver não é mais acreditar. *Nat. Mach Intel.* 1 (8), 332–334. doi:10.1038/s42256-019-0085-5

Texto completo CrossRef | **Google Scholar**

Bradford, H. (2017). A reputação da Abercrombie and Fitch sofre um golpe depois que os comentários 'gordos' do CEO ressurgem. Disponível em: **https://www.huffpost.com/entry/abercrombie-reputation-ceo-comments_n_3288836** (Acessado em 1 de abril de 2020).

Google Scholar

Brown, NI (2020). Deepfakes e a transformação da desinformação em arma. *Va.* 23, 1.

Google Scholar

Burgess, M. (2020). Um bot pornográfico deepfake está sendo usado para abusar de milhares de mulheres. Disponível em: **https://www.wired.co.uk/article/telegram-deepfakes-deepnude-ai** (Acessado em 1 de abril de 2020).

Google Scholar

Business Insider (2017). As pessoas tirarão 1,2 trilhão de fotos digitais este ano – graças aos smartphones. Disponível em: **https://www.businessinsider.com/12-trillion-photos-to-be-taken-in-2017-thanks-to-smartphones-chart-2017-8?r=US&IR=T** (acessado em 1º de abril , 2020).

Google Scholar

Camarões, BH (2020). O podcast bellingcat 2ª temporada - as execuções. Disponível em: **https://www.bellingcat.com/resources/podcasts/2020/07/21/the-bellingcat-podcast-season-2-the-executions/** (Acessado em 1 de abril de 2020).

Google Scholar

Chesney, B. e Citron, D. (2019a). Deep fakes: um desafio iminente para a privacidade, a democracia e a segurança nacional. *Califórnia L. Rev.* 107, 1753. doi:10.2139/ssrn.3213954

Texto completo CrossRef | **Google Scholar**

Chesney, R. e Citron, D. (2019b). Deepfakes e a nova guerra de desinformação. Disponível em: **https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war** (Acessado em 1 de abril de 2020).

Google Scholar

Darpa (2020). Análise forense de mídia (MediFor). Disponível em: **https://www.darpa.mil/program/media-forensics** (Acessado em 1 de abril de 2020).

Google Scholar

Dickerman, L. (2000). Camera obscura: realismo socialista à sombra da fotografia. *93 de outubro* de 139–153. doi:10.2307/779160

Texto completo CrossRef | **Google Scholar**

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., et al. (2020). O conjunto de dados do desafio de detecção de deepfake. Nome do repositório de pré-impressão [Pré-impressão]. Disponível em: **arXiv:2006.07397** .

Google Scholar

Dolhansky, B., Howes, R., Pflaum, B., Baram, N. e Ferrer, CC (2019). O conjunto de dados de visualização do desafio de detecção deepfake (dfdc). Nome do repositório de pré-impressão [Pré-impressão]. Disponível em: **arXiv:1910.08854** .

Google Scholar

Donovan, J. e Paris, B. (2019). Cuidado com as falsificações baratas. Disponível em: **https://slate.com/technology/2019/06/drunk-pelosi-deepfakes-cheapfakes-artificial-intelligence-disinformation.html** (Acessado em 1 de abril de 2020).

Google Scholar

UE vs DisInfo (2020). A desinformação pode matar. Disponível em: **https://euvsdisinfo.eu/disinformation-can-kill/** (Acessado em 1 de abril de 2020).

Google Scholar

Centro Europeu de Mídia Científica (2019). Deepfakes, superficialfakes e síntese de fala: combatendo a manipulação audiovisual. Disponível em: **https://sciencemediahub.eu/2019/12/04/deepfakes-shallowfakes-and-speech-synthesis-tackling-audiovisual-manipulation/** (Acessado em 1 de abril de 2020).

Google Scholar

Rebelião da Extinção (2020). O discurso do primeiro-ministro pelos nossos rebeldes. Disponível em: **https://www.extinctionrebellion.be/en/tell-the-truth/the-prime-ministers-speech-by-the-rebels** (Acessado em 1 de abril de 2020).

Google Scholar

Cara, CS (2020). Início Stallone [DeepFake]. Disponível em: **https://www.youtube.com/watch?v=2svOtXaD3gg** (Acessado em 1 de abril de 2020).

Google Scholar

Fedler, R., Schütte, J. e Kulicke, M. (2013). Sobre a eficácia da proteção contra malware no Android. *Fraunhofer AISEC* 45.

Google Scholar

Ferrer, CC, Dolhansky, B., Pflaum, B., Bitton, J., Pan, J. e Lu, J. (2020). Resultados do desafio de detecção de Deepfake: uma iniciativa aberta para o avanço da IA. Disponível em: **https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/** (Acessado em 1 de abril de 2020).

Google Scholar

Foster, A. (2019). Como a perturbadora tecnologia de IA pode ser usada para enganar namoradores online. Disponível em: **https://www.news.com.au/technology/online/security/how-disturbing-ai-technology-could-be-used-to-scam-online-daters/news-story/1be46dc7081613849d67b82566f8b421** (Acessado 1º de abril de 2020).

Google Scholar

Galindo, G. (2020). XR Bélgica publica deepfake do primeiro-ministro belga ligando Covid-19 à crise climática. Disponível em: **https://www.brusselstimes.com/news/belgium-all-news/politics/106320/xr-belgium-posts-deepfake-of-belgian-premier-linking-covid-19-with-climate-crisis /** (Acessado em 1 de abril de 2020).

Google Scholar

Gandhi, A. e Jain, S. (2020). Perturbações adversárias enganam os detectores de deepfake. Nome do repositório de pré-impressão [Pré-impressão]. Disponível

em: **arXiv:2003.10596** . doi:10.1109/ijcnn48605.2020.9207034

Texto completo CrossRef | **Google Scholar**

Goodfellow, IJ, Shlens, J. e Szegedy, C. (2014). Explicando e aproveitando exemplos adversários. Nome do repositório de pré-impressão [Pré-impressão]. Disponível em: **arXiv:1412.6572** .

Google Scholar

Gorwa, R., Binns, R. e Katzenbach, C. (2020). Moderação algorítmica de conteúdo: desafios técnicos e políticos na automação da governança de plataformas. *Sociedade de Big Data.* 7 (1), 2053951719897945.doi:10.1177/2053951719897945

Texto completo CrossRef | **Google Scholar**

Gu, S. e Rigazio, L. (2014). Rumo a arquiteturas de redes neurais profundas robustas a exemplos adversários. Nome do repositório de pré-impressão [Pré-impressão]. Disponível em: **arXiv:1412.5068** .

Google Scholar

Guan, H., Kozak, M., Robertson, E., Lee, Y., Yates, AN, Delgado, A., et al. (2019). “Conjuntos de dados MFC: conjuntos de dados de referência em larga escala para avaliação de desafios forenses de mídia”, em aplicações de inverno de oficinas de visão computacional do IEEE (WACVW) , Waikoloa, HI, 7 a 11 de janeiro de 2019 (IEEE), 63–72.

Google Scholar

Salão, Hong Kong (2018). Vídeos deepfake: quando ver não é acreditar. *Cat* . *Tecnologia UJL.* 27, 51.

Google Scholar

Hasan, RH e Salah, K. (2019). Combate a vídeos deepfake usando blockchain e contratos inteligentes. *Acesso IEEE* 7, 41596–41606. doi:10.1109/access.2019.2905689

Texto completo CrossRef | **Google Scholar**

Gancho, C. (2019). O aviso do filtro 'bebê' do Snapchat pode ser um presente para predadores infantis online. Disponível em: **https://7news.com.au/technology/snapchat/warning-snapchat-baby-filter-could-be-a-gift-to-online-child-predators-c-114074** (Acessado em 1 de abril de 2020).

Google Scholar

Tempos de negócios internacionais (2020). Escândalo da Nth Room: aqui está como Jo Joo Bin vitimou ídolos femininos populares com pornografia deepfake. Disponível em: **https://www.ibtimes.sg/nth-room-scandal-here-how-jo-joo-bin-victimized-popular-female-idols-deepfake-pornography-42238** (Acessado em 1 de abril de 2020).

Google Scholar

Khanna, M. (2020). Como o BJP usou deepfake em um de seus vídeos de campanha em Delhi e por que isso é perigoso. Disponível

em: **https://www.indiatimes.com/technology/news/how-bjp-used-deepfake-for-one-of-its-delhi-campaign-videos-and-why-its-dangerous-506795.html** (Acessado em 1º de abril de 2020).

Google Scholar

Rei, D. (2014). *O comissário desaparece: a falsificação de fotografias e de arte na Rússia de Stalin* . Londres, Reino Unido: Tate Publishing .

Koren, S. (2019). Apresentando o projeto de proveniência de notícias. Disponível em: **https://open.nytimes.com/introduzindo-the-news-provenance-project-723dbaf07c44** (acessado em 1 de abril de 2020).

Google Scholar

Krizhevsky, A., Sutskever, I. e Hinton, GE (2012). “Classificação Imagenet com redes neurais convolucionais profundas”, em Avanços em sistemas de processamento de informações neurais , Lake Tahoe, Nevada , 3–6 de dezembro de 2012 , 1097–1105.

Google Scholar

Matatov, H., Bechhofer, A., Aroyo, L., Amir, O. e Naaman, M. (2018). “DejaVu: um sistema para jornalistas abordarem colaborativamente a desinformação visual”, no *simpósio de jornalismo Computation+* (Miami, FL: Universidade de Harvard).

Google Scholar

Meskin, A. e Cohen, J. (2008). “Fotografias como evidência”, in *Fotografia e filosofia: ensaios sobre o lápis da natureza* . Editor S. Walden (Hoboken, Nova Jersey: Wiley), 70–90.

Google Scholar

Mishra, R. (2020). “Detecção de notícias falsas usando progressão de atenção mútua de usuário para usuário de ordem superior em caminhos de propagação”, em Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops , Seattle, WA , 16 a 18 de junho de 2020 (IEEE), 652–653.

Google Scholar

Neekhara, P., Hussain, S., Jere, M., Koushanfar, F. e McAuley, J. (2020). Deepfakes adversários: avaliando a vulnerabilidade de detectores de deepfakes a exemplos adversários. Nome do repositório de pré-impressão [Pré-impressão]. Disponível em: **arXiv:2002.12749** .

Google Scholar

Nelson-Field, K., Riebe, E. e Newstead, K. (2013). As emoções que impulsionam o vídeo viral. *Austrália. Marketing J.* 21 (4), 205–211. doi:10.1016/j.ausmj.2013.07.003

Texto completo CrossRef | **Google Scholar**

NVIDIA (2020). Pintura de imagem. Disponível em: **https://www.nvidia.com/research/inpainting/** (Acessado em 1 de abril de 2020).

Google Scholar

Pierson, M. (1999). Efeitos CGI no cinema de ficção científica de Hollywood 1989-95: os anos maravilhosos. *Tela* 40 (2), 158–176. doi:10.1093/tela/40.2.158

Texto completo CrossRef | **Google Scholar**

Politifato (2020). O vídeo de 'enchimento de votos' não é de um local de votação em Flint, Michigan. é da Rússia. Disponível em: **https://www.politifact.com/factchecks/2020/nov/05/facebook-posts/video-ballot-stuffing-filmed-russia-not-flint-mich/** (Acessado em 1 de abril de 2020).

Google Scholar

Rastogi, V., Chen, Y. e Jiang, X. (2013). “DroidChameleon: avaliando o antimalware Android contra ataques de transformação”, em Anais do 8º simpósio ACM sobre segurança de informações, computadores e comunicações , Hangzhou, China , 8 a 10 de maio de 2013 (ASIACCS).

Texto completo CrossRef | **Google Scholar**

Reddit (2020). Photoshopbattles: um lugar para se divertir com o software gráfico favorito de todos. Disponível em: **https://www.reddit.com/r/photoshopbattles/** (Acessado em 1 de abril de 2020).

Google Scholar

Reflita (2020). Troca de rosto realista. Disponível em: **https://reflect.tech/** (Acessado em 1 de abril de 2020).

Google Scholar

Repórter Ohne Grenzen (2020). México javier valdez fordert gerechtigkeit für ermordete medienschaffende repórter ohne grenzen. Disponível em: **https://www.youtube.com/watch?v=ZdIHUrCobuc** (Acessado em 1 de abril de 2020).

Google Scholar

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. e Nießner, M. (2019). “Faceforensics++: aprendendo a detectar imagens faciais manipuladas”, em Proceedings of the IEEE International Conference on Computer cision , Seul, Coreia , 27 de outubro a 2 de novembro de 2019 , 1–11.

Google Scholar

Ruiz, D. (2020). Leis e propostas de deepfakes inundam os Estados Unidos. Disponível em: **https://blog.malwarebytes.com/artificial-intelligence/2020/01/deepfakes-laws-and-proposals-flood-us/** (Acessado em 1 de abril de 2020).

Google Scholar

Schwartz, O. (2018). Você achou que notícias falsas eram ruins? falsificações profundas são onde a verdade vai morrer. Guardião, novembro de 2018.

Google Scholar

Sensibilidade (2020). A primeira plataforma de detecção de deepfakes do mundo. Disponível em: **https://sensity.ai/** (Acessado em 1 de abril de 2020).

Google Scholar

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E. e Sebe, N. (2019). “Modelo de movimento de primeira ordem para animação de imagens”, em Advances in neural information processing systems , Vancouver, Canadá , 8 a 14 de dezembro de 2019 , 7137–7147.

Google Scholar

Stupp, K. (2019). Os fraudadores usaram IA para imitar a voz do CEO em casos incomuns de crimes cibernéticos. Disponível em: **https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402** .

Google Scholar

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Propriedades intrigantes de redes neurais. Nome do repositório de pré-impressão [Pré-impressão]. Disponível em: **arXiv:1312.6199** .

Google Scholar

Revisão de tecnologia (2020). Deepfake Putin está aqui para alertar os americanos sobre sua destruição autoinfligida. Disponível em: **https://www.technologyreview.com/2020/09/29/1009098/ai-deepfake-putin-kim-jong-un-us-election/** (Acessado em 1 de abril de 2020).

Google Scholar

Temperton, J. (2020). A ascensão e disseminação de uma teoria da conspiração do coronavírus 5G. Disponível em: **https://www.wired.com/story/the-rise-and-spread-of-a-5g-coronavirus-conspiracy-theory/** (Acessado em 1 de abril de 2020).

Google Scholar

O Guardião (2015). Falsifiquei o vídeo do dedo médio de Yanis Varoufakis, diz apresentador de TV alemão. Disponível em: **https://www.theguardian.com/world/2015/mar/19/i-faked-the-yanis-varoufakis-middle-finger-video-says-german-tv-presenter** . (Acessado em 1º de abril de 2020).

Google Scholar

A beira (2020). O vídeo do debate da Bloomberg violaria a política de deepfake do Twitter, mas não a do Facebook. Disponível em: **https://www.theverge.com/2020/2/20/21146227/facebook-twitter-bloomberg-debate-video-manipulated-deepfake** (Acessado em 1 de abril de 2020).

Google Scholar

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C. e Nießner, M. (2016). “Face2face: captura de rosto em tempo real e reconstituição de vídeos RGB”, em Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition , Las Vegas, Nevada , 26 de junho a 1º de julho de 2016 (IEEE), 2387–2395.

Google Scholar

Universidade de Hong Kong (2020). Projeto Weiboscópio. Disponível em: **https://weiboscope.jmsc.hku.hk/64CensoredPics/** (Acessado em 1 de abril de 2020).

Google Scholar

Verdoliva, L. (2020). Análise forense de mídia e deepfakes: uma visão geral. *IEEE J. Sel. Principal. Sinal. Processo.* 14 (5), 910–932. doi:10.1109/JSTSP.2020.3002101

Texto completo CrossRef | **Google Scholar**

Vice (2020). Veja como os deepfakes, como o usado pelo BJP, distorcem a verdade. Disponível em: **https://www.vice.com/en/article/939d4p/bjp-manoj-tiwari-deepfake-twists-truth** (Acessado em 1 de abril de 2020).

Google Scholar

Walden, M. (2019). Ministro da Malásia supostamente retratado em vídeo de sexo gay, acusado de corrupção. Disponível em: **https://mobile.abc.net.au/news/2019-06-12/gay-sex-video-sparks-scandal-for-malaysian-government/11201834?pfmredir=sm** (Acessado em 1 de abril de 2020).

Google Scholar

Wikipédia (2020). Arquivos. Disponível em: **https://en.wikipedia.org/wiki/File:Union-de-Lucha.jpg** (Acessado em 1 de abril de 2020).

Google Scholar

CK Winkler e CE Dauber (Editores) (2014). *Propaganda visual e extremismo no ambiente online* . Carlisle, PA: Instituto de Estudos Estratégicos do Army War College Carlisle Barracks Pa .

Laboratório de Testemunhas (2020). Prepare-se, não entre em pânico: mídia sintética e deepfakes. Disponível em: **https://lab.witness.org/projects/synthetic-media-and-deep-fakes/** (Acessado em 1 de abril de 2020).

Google Scholar

Cobertura Zero (2020). Tuítes supostamente falsos sobre a morte do presidente sírio Assad causam um aumento muito real no petróleo. Disponível em: **http://www.zerohedge.com/news/supposedly-fake-tweets-about-syrian-president-assads-death-cause-all-too-real-spike-crude-and-s** (acessado em 1º de abril , 2020).

Google Scholar

Zhou, P., Han, X., Morariu, VI e Davis, LS (2018). “Aprendendo recursos ricos para detecção de manipulação de imagens”, em Anais da conferência IEEE sobre visão computacional e reconhecimento de padrões , Salt Lake City, UT , 18 a 23 de junho de 2018 (IEEE), 1053–1061.

Google Scholar

Palavras-chave: deepfakes, manipulação de mídia visual, notícias falsas, desinformação, redes adversárias generativas

Citação: Langguth J, Pogorelov K, Brenner S, Filkuková P e Schroeder DT (2021) Não confie em seus olhos: manipulação de imagens na era dos DeepFakes. *Frente. Comum.* 6:632317. doi: 10.3389/fcomm.2021.632317

Recebido: 23 de novembro de 2020; **Aceito:** 25 de janeiro de 2021;

Publicado: 24 de maio de 2021.

Editado por:

Daniel Broudy

, Universidade Cristã de Okinawa, Japão

Revisados pela:

Vian Bakir

, Universidade de Bangor, Reino Unido

Michael D. High

, Universidade Xi'an Jiaotong-Liverpool, China

- **Correspondência:langguth@simula.no**

  Johannes Langguth,