Atualizado: Março 2022

A WITNESS ajuda as pessoas a usarem o vídeo e a tecnologia para proteger e defender os direitos humanos – witness.org. Para saber mais sobre o nosso trabalho com deepfakes e para se preparar melhor, acesse: wit.to/Synthetic-Media-Deepfakes

Deepfakes

Deepfakes tornam fácil manipular ou forjar vozes, rostos e ações de pessoas reais, assim como alegar que qualquer vídeo ou áudio é fake. Virou uma questão crítica para celebridades, políticos e para muitas mulheres ao redor do mundo. Enquanto deepfakes tornam-se cada vez mais fáceis de produzir, a WITNESS busca por respostas para seus atuais problemas e se prepara para ameaças futuras focadas em populações vulneráveis globalmente. Neste manual, exploramos:

- Tecnologias: Quais as principais tecnologias de deepfakes e o que podem fazer?
- Ameaças: Quais são as principais ameaças identificadas globalmente?
- Soluções: Quais são as potenciais soluções técnicas e políticas?

O que são deepfakes e mídia sintética?

Deepfakes são novas formas de manipulação audiovisual que permitem criar simulações reais do rosto, voz e ações de alguém. Permitem que se faça parecer que alguém disse ou fez algo que não fez ou que um evento que aconteceu, mas que de fato nunca ocorreu. Essas manipulações têm se tornado cada vez mais fáceis de fazer, sendo necessárias apenas poucas imagens para produzi-las. Além disso, as ferramentas para criá-las estão sendo crescentemente comercializadas.

Atualmente, os deepfakes impactam mulheres de forma desproporcional, pois são usados para criar imagens e vídeos de cunho sexual não consensuais com o rosto de uma pessoa específica. Mas teme-se que deepfakes terão um impacto mais amplo na sociedade, negócios e política, bem como em investigações sobre direitos humanos, apuração de notícias e processos de verificação.

Esse conjunto de ferramentas e técnicas permite a criação de representações realistas de pessoas falando ou fazendo coisas que nunca fizeram, de objetos e pessoas que nunca existiram, ou de eventos que nunca aconteceram.

Mídia e tecnologia sintética atualmente permitem estas formas de manipulação:

- Adicionar e remover objetos mais facilmente em um vídeo.
- Alterar as condições do contexto de um vídeo. Por exemplo, alterar o clima para fazer um vídeo filmado no verão parecer que foi filmado no inverno.
- Forjar movimentos corporais ou faciais ("marionetismo"): simular e controlar em vídeo uma representação realista dos lábios, expressões faciais ou movimentos corporais de um indivíduo específico (para, por exemplo, fazer parecer que alguém está bêbado).
- Forjar sincronização labial: combinar um arquivo de áudio com uma manipulação realista de lábios para fazer parecer que alguém disse algo que nunca disse.
- Falsificação de voz: gerar uma simulação realista da voz de uma pessoa em específico.
- Alterar o gênero da voz ou fazer soar como sendo de outra pessoa: modificar uma voz já existente com um efeito de voz de outro gênero, ou de uma pessoa específica.
- Criar uma foto totalmente falsa, porém realista, de uma pessoa que não existe. A mesma técnica pode também ser aplicada sem tantos problemas para criar falsos hambúrgueres, gatos, etc.
- Criar uma foto de um evento ou objeto a partir de uma descrição em texto.
- Transferir o rosto realista de uma pessoa para outra, a forma mais conhecida de "deepfake".

[Veja exemplos de muitas delas aqui]

Como funciona a tecnologia por trás disso tudo?

Essas técnicas, primariamente, *mas não exclusivamente*, baseiam-se numa forma de inteligência artificial conhecida como aprendizagem aprofundada e o trabalho da Rede Adversária Generativa, ou GANs.

Para gerar um item de conteúdo de mídia sintética, você começa coletando imagens ou vídeos fontes da pessoa ou item que você quer forjar. Uma GAN desenvolve a falsificação - seja ela de uma pessoa real ou trocas de rosto - usando duas redes. Uma rede gera recriações plausíveis da imagem fonte, enquanto a segunda rede trabalha para detectar essas falsificações. Esses dados detectados são enviados de volta à rede dedicada a criar as falsificações, permitindo que ela crie versões cada vez melhores.

Ainda hoje, muitas dessas técnicas – particularmente a criação de deepfakes realistas de troca de rostos – continuam a exigir poder computacional significativo, uma compreensão de como ajustar seu modelo e, muitas vezes, uma significativa pós-produção de CGI para melhorar o resultado final. Um bom exemplo de um deepfake sofisticado que usa todos esses recursos são os vídeos do "Tom Cruise" no TikTok que você já deve ter visto!

Entretanto, mesmo com limitações atuais, as pessoas continuam sendo enganadas pela mídia simulada. Como exemplo, pesquisas mostraram que pessoas nem sempre conseguiam detectar

as formas atuais de modificação de movimento de lábios que são usadas para combinar a boca de alguém com um arquivo de áudio. E pesquisas recentes descobriram que humanos não eram capazes de apontar rostos realistas de pessoas que nunca existiram.
Isso significa que as pessoas não estão inerentemente equipadas para detectar a manipulação de mídia sintética.

O cenário atual de Deepfakes e mídia sintética

Deepfakes maliciosos e mídia sintética – até agora – não são difundidos para além do contexto de imagens sexuais não consensuais. Infelizmente, deepfakes de conteúdos sexuais não consensuais estão facilmente disponíveis e são gerados envolvendo celebridades, atrizes pornôs ou pessoas comuns.

Somado a isso:

- As pessoas começaram a questionar conteúdos reais, desprezando-os como se fossem deepfakes.
- Enquanto as sátiras de deepfakes proporcionam novas oportunidades de livre expressão, também traçam uma linha tênue com fraudes.
- Imagens de "pessoas que nunca existiram" estão crescentemente sendo usadas para esconder contas falsas de desinformação.

As ameaças dos deepfakes

Em oficinas organizadas pela WITNESS, bem como em treinamentos com mais de 500 pessoas nos últimos três anos, revisamos potenciais vetores de ameaças com um amplo número de participantes da sociedade civil, incluindo mídias populares, jornalistas profissionais, pesquisadores e especialistas de OSINT (Inteligência de código aberto). Foram priorizadas áreas onde novas formas de manipulação podem expandir ameaças já existentes, trazer novas ameaças, alterar ameaças e reforçar outras. Também foram destacados os desafios ao redor do debate de "isso é deepfake" como um primo próximo de "isso é fake news".

Participantes nos nossos encontros de especialistas do Sudeste da Ásia, África Subsaariana e Brasil, e outros encontros globalmente, priorizaram suas principais preocupações sobre como novas formas de manipulação de mídia, aumento de desinformação e falta de informação afetarão seu trabalho, sociedades e comunidades.

- Jornalistas, líderes comunitários e ativistas civis terão suas reputações e credibilidades atacadas baseado em formas de assédio e violência virtual já existentes que atingem predominantemente mulheres e minorias.

- Um número de ataques usando mídia modificada já foi feito com jornalistas mulheres, como no importante caso da jornalista Indiana Rana Ayyub.
- Figuras públicas enfrentarão violência de gênero e exposição imagética sexual não consensual, bem como outras formas de uso das chamadas cópias críveis. Políticos locais podem estar particularmente vulneráveis por terem muitas imagens, porém menos estrutura institucional ao seu redor, de modo que políticos nacionais tenham que ajudar a defendê-los contra ataques de mídia sintética.
- Prejudicar as possibilidades do uso de vídeos como evidência do abuso de direitos humanos e crimes, dificultando a responsabilização e a justiça.
- Os jornalistas já sobrecarregados e com poucos recursos não terão a capacidade forense da mídia para capacitar jornalistas e apuradores para verificar deepfakes.
- Organizações de direitos humanos, coleta de notícias e de verificação serão pressionadas a provar que algo é verdade, bem como a provar que algo não é forjado. Aqueles que estiverem em posição de poder terão a oportunidade de usar negação plausível em conteúdos, declarando ser deepfake.
- Na medida em que deepfakes tornam-se mais comuns e fáceis de serem produzidos em grande escala, eles contribuirão para as estratégias de "metralhadora de mentiras", inundando as redações e agências de verificação de mídia e de apuração com conteúdos que deverão verificar ou desmentir. Isso pode distraí-las e sobrecarregá-las.
- Deepfakes cruzarão com padrões existentes de "incêndios digitais", onde imagens falsas são rapidamente compartilhadas via WhatsApp, Telegram, Facebook Messenger e outros aplicativos de mensagens.
- Videoconferências online estarão vulneráveis à manipulação.

Em todos os contextos, as pessoas que consultamos notaram a importância de enxergar os deepfakes no contexto das abordagens já existentes de verificação e apuração. Deepfakes e mídia sintética serão integrados a campanhas existentes de conspiração e desinformação, trazendo táticas em evolução (e respostas) nessa área.

A pesquisa da WITNESS em deepfakes e sátira, incluindo um relatório recente, Just Joking!, identificou um crescente uso de deepfakes para poderosa crítica política e social. Foi mostrado como deepfakes foto realistas satíricos levaram a:

- Crítica social: paródia e sátira para criticar o poder que identifica problemas políticos e sociais e que o público reconhece como satírico.
- Mau uso deliberado: alegar que algo é uma sátira quando é desinformação e culpar a audiência por "não entender a piada".
- Mau uso acidental: quando o contexto é perdido, e é compartilhado como má informação identificada como real.

Atualizado: Março 2022

Quais são as soluções disponíveis?

Há uma quantidade considerável de trabalho para se preparar melhor para deepfakes. A WITNESS se preocupa que essa busca por "soluções" não inclua adequadamente as vozes e necessidades de pessoas atingidas por problemas existentes de manipulação de mídia, violência do estado, violência de gênero e má informação ou desinformação no Sul Global e comunidades marginalizadas no Norte Global.

É possível ensinar as pessoas a identificar?

Não é uma boa ideia ensinar às pessoas que elas podem identificar deepfakes ou manipulação de mídia. Apesar de haver algumas dicas que podem ajudar a identificá-las agora – por exemplo, falhas visíveis – são apenas falhas atuais no processo de falsificação e desaparecerão com o passar do tempo. Porém, se quiser testar a sua habilidade, acesse: https://detectfakes.media.mit.edu/

Plataformas como o Facebook e empresas independentes desenvolverão ferramentas que podem detectar, mas que darão apenas algumas dicas e não estarão completamente disponíveis tão brevemente. É importante que as pessoas foquem em entender deepfakes num quadro mais amplo de literacia midiática, como a abordagem SHEEP da organização First Draft ou o quadro SIFT.

SHEEP (um acrônimo em inglês) sugere que para evitar ser enganado por má informação online, você deveria "pensar SHEEP antes de compartilhar."

SOURCE (Fonte): olhe para o que está por trás. Verifique a área "sobre" do site ou da conta, busque por qualquer informação da conta e pesquise nomes e sobrenomes.

HISTORY (Histórico): essa fonte tem segundas intenções? Descubra quais assuntos ela geralmente aborda, ou se promove somente um único ponto de vista.

EVIDENCE (Evidência): explore os detalhes da informação ou meme para descobrir se há alguma evidência confiável em algum outro lugar.

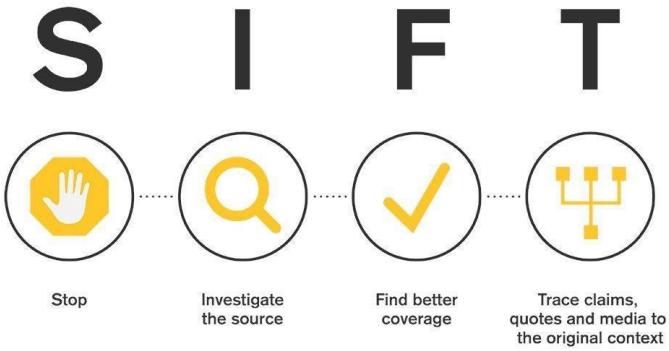
EMOTION (Emoção): a fonte se baseia em algum tipo de emoção para estabelecer um ponto? Verifique a linguagem excludente, sensacionalista e provocativa.

PICTURES (Imagens): Uma imagem vale mais que mil palavras. Identifique qual mensagem uma imagem está retratando e se a fonte está usando imagens para chamar a atenção.

WITNESS
SEE IT FILM IT CHANGE IT

Atualizado: Março 2022

SIFT proporciona outro modelo relacionado de análise de senso comum de informação suspeita:



S (STOP) - Pare
I (INVESTIGATE) - Investigue a fonte
F (FIND)- Encontre mais informações
T (TRACE) - Rastreie alegações, citações e mídia do contexto original

Atualizado: Março 2022

Como podemos nos apoiar na coordenação e na capacidade jornalística existente?

Jornalistas e investigadores de direitos humanos precisam desenvolver uma compreensão melhor sobre como detectar deepfakes usando práticas existentes de OSINT e combinando-as com novas ferramentas de mídia forense que estão sendo desenvolvidas. Saiba mais no relatório de necessidades da WITNESS e no recente trabalho de Parceria em IA.

Existem ferramentas para detecção? (e quem tem acesso?)

A maioria das grandes plataformas e muitas startups estão desenvolvendo ferramentas para detecção de deepfakes.

Algumas ferramentas estão começando a ser lançadas. Um exemplo da Sensity.AI https://platform.sensity.ai/deepfake-detection.

Contudo, recomendamos abordá-las com cautela e, ter a identificação como possível sinal de manipulação, não uma confirmação. Amplas competições recentes para detecção de deepfakes não apresentaram modelos suficientemente eficazes às técnicas conhecidas ou suficientemente aplicáveis a novas técnicas, e a maioria dos detectores disponíveis publicamente será menos eficaz do que sistemas mais fechados. Ferramentas de detecção tendem a ser menos confiáveis se você não conhece técnicas usadas para gerar mídia sintética, e menos confiáveis para a mídia comprimida e de baixa resolução que vemos online. Uma recente experiência de uma suspeita de deepfake em Myanmar mostra os desafios de confiar em detectores disponíveis publicamente sem o acompanhamento de perícia.

Mesmo que ferramentas robustas estejam sendo desenvolvidas, elas não estarão disponíveis amplamente, particularmente fora de plataformas e empresas de mídia. É provável que a mídia e organizações da sociedade civil no Sul Global sejam deixadas de fora e é importante defender mecanismos que lhes permitam ter acesso a esses recursos de detecção. A WITNESS está debatendo pela equidade em acesso às ferramentas de detecção, investimento nas habilidades e capacidade da sociedade civil global e do jornalismo e no desenvolvimento de mecanismos de escalonamento que permitam análises críticas de suspeitas de deepfakes.

Existem ferramentas de autenticação? (e quem está excluído?)

Existe um movimento crescente para desenvolver ferramentas que rastreiam melhor de onde vem vídeos e imagens – começando com o momento em que são gravados em smartfones, até quando são editados e compartilhados ou distribuídos nas redes sociais. Essa "infraestrutura de autenticidade e procedência" pode então mostrar informações de onde, se, e como uma foto ou vídeo foi modificado. Isso é relevante tanto para falsificações superficiais como vídeos fora de

contexto ou vídeos editados, bem como deepfakes. Você pode usar essa informação para decidir se pode confiar no conteúdo. Um exemplo da iniciativa nessa área é a *Iniciativa de Autenticidade de Conteúdo* (Content Authenticity Initiative), e a recentemente lançada *Coalizão para Autenticidade e Procedência de Conteúdo* (Coalition for Content Provenance and Authenticity – C2PA).

Entretanto, existe um risco de que ferramentas desenvolvidas para ajudar a rastrear as origens de vídeos e mostrar como eles foram manipulados podem também criar os riscos de vigilância e de exclusão de pessoas que não querem adicionar dados e informações extras às suas fotos e vídeos, ou não podem atribuir fotos a si mesmas por medo do que os governos e empresas farão com essa informação. A WITNESS liderou uma avaliação de Danos, Mau uso e Abuso das especificações da C2PA para identificar esses e outros potenciais danos, e para desenvolver estratégias para revertê-los e minimizá-los. Esta "infraestrutura de autenticidade e procedência" ainda será usada de forma abusiva, e agentes maliciosos encontrarão brechas, então o principal próximo passo é reforçar um quadro de direitos humanos, com barreiras contra danos, mecanismos de reparação e oportunidades para capacitar vozes críticas.

O que as plataformas deveriam fazer?

Plataformas de redes sociais como Facebook e Twitter possuem políticas sobre deepfakes e como lidarão com elas, e ainda sobre como lidarão com mídia manipulada de forma mais ampla. A WITNESS discute a política do Facebook aqui e a política do Twitter aqui.

Os principais elementos dessas políticas incluem:

- Eles cobrem apenas deepfakes ou também outras formas de mídia manipulada (ex. um vídeo desacelerado ou um vídeo que é descontextualizado)?
- Como eles definem o dano causado por um vídeo?
- A intenção por trás do compartilhamento importa?
- Eles derrubam um vídeo ofensivo? Vão sinalizá-lo? Fornecem o contexto da manipulação? Tornam o vídeo menos visível no site ou menos compartilhável?
- Se aplicam a figuras públicas?

Facebook (Meta)

A política do Facebook é específica para deepfakes em vez de outras formas de manipulação de fotos e vídeo.

O Facebook vai remover uma mídia manipulada quando

- "Ela for editada ou sintetizada – além de ajustes para claridade ou qualidade – em maneiras que não são aparentes para uma pessoa comum e provavelmente levaria alguém a pensar que o sujeito do vídeo disse ou fez coisas que não fez.

- Ela for produto da inteligência artificial ou da aprendizagem de máquina que funde, substitui ou sobrepõe o conteúdo em um vídeo, fazendo-o parecer ser autêntico.

Essa política não se estende ao conteúdo que é paródia ou sátira, ou ao vídeo que foi editado exclusivamente para omitir ou alterar a ordem das palavras. Outro vídeo manipulado enganoso pode ser referenciado ou pego por seus apuradores terceirizados. Há exemplos em que o Facebook erroneamente derrubou uma sátira de deepfake considerando ser desinformação. Um exemplo disso ocorreu em Camarões quando um acadêmico e ativista local compartilhou um vídeo claramente fabricado do embaixador francês dizendo aos camaroneses que eles nunca realmente alcançaram a independência da exploração colonial da França. Apuradores de fatos terceirizados do Facebook na emissora francesa France 24 rotularam o vídeo como parcialmente forjado, anulando assim o poder retórico da crítica.

"Áudio, fotos ou vídeos, sejam eles um deepfake ou não, serão removidos do Facebook se violarem qualquer um dos nossos outros Padrões Comunitários, incluindo aqueles a respeito de nudez, violência gráfica, supressão de eleitores e discurso de ódio."

A política do Twitter está disponível aqui.

Ela indica que "você não pode enganosamente compartilhar mídia sintética ou manipulada (não apenas deepfakes, mas também outras formas de manipulação) que são prováveis de causar danos. Além disso, podemos rotular Tweets contendo mídia sintética e manipulada para ajudar as pessoas a entender sua autenticidade e fornecer contexto adicional."

O Twitter foca em três questões-chave que determinam se eles podem ou rotular o conteúdo ou removê-lo.

1. Esse conteúdo é sintético ou manipulado?
2. O conteúdo está compartilhado de maneira enganosa?
3. O conteúdo pode impactar na segurança pública ou causar danos sérios?

Is the content significantly and deceptively altered or fabricated?	Is the content shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled.
✗	✓	✗	Content **may** be labeled.
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.*
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **likely** to be removed.

Atualizado: Março 2022

TikTok

O TikTok "proíbe falsificações digitais (Mídia sintética ou Mídia manipulada) que leva os usuários a distorcerem a veracidade dos fatos e causam dano ao sujeito do vídeo, a outras pessoas, ou à sociedade".

Nosso posicionamento

Nossas plataformas devem ser proativas em sinalização, redução – e, no pior dos casos, remoção – de deepfakes maliciosos, porque os usuários possuem experiência limitada nesse tipo de manipulação invisível aos olhos e inaudível aos ouvidos, e porque os jornalistas não têm as ferramentas prontas para detectá-los de forma rápida ou eficaz. Mas abordar deepfakes não remove a responsabilidade de também abordar ativamente outras formas de manipulação de vídeo do tipo "shallowfake", como rotular um vídeo real ou editar levemente um vídeo real.

Algumas questões em curso que se relacionam com as políticas:

- Como é que tanto o Facebook, Twitter, Tiktok e outros irão garantir uma detecção precisa de deepfakes?
- Como as plataformas farão um julgamento sobre quando uma modificação é maliciosa, ou se algo é uma paródia, ou em vez disso se disfarça como sátira ou paródia?
- Como comunicarão o que sabem aos consumidores céticos?
- Como garantirão que as decisões que tomarem estarão sujeitas a transparência e apelação por causa de erros inevitáveis?

O que os legisladores estão fazendo?

Governos estão apenas começando a legislar sobre deepfakes. Essas leis se referem tanto a imagens sexuais não consensuais bem como aos usos para fraude, má informação ou desinformação.

Nos EUA, foi proposta uma série de leis em nível estadual e federal e, na UE (União Europeia), a lei IA apoia a rotulagem de mídia sintética para proteção do consumidor.

Na região Ásia-Pacífico, dois exemplos são as leis da República Popular da China, que proíbem deepfakes e outras notícias falsas, e a recente legislação proposta nas Filipinas. Uma cautela sobre essas leis é quando elas fazem uma definição muito ampla de falsificação audiovisual e incluem formas importantes de liberdade de expressão, como sátira, ou dão ampla discrição e poder aos governos para decidir o que é 'falso'.