PRADO, Magaly Pereira do. Deepfake de áudio: manipulação simula voz real para retratar alguém dizendo algo que não disse. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 45-68.

Recebido em: 9 mar. 2021
Aprovado em: 3 abr. 2021

dx.doi.org/10.23925/1984-3585.2021i23p45-68

Deepfake de áudio:

manipulação simula voz real para retratar alguém dizendo algo que não disse

Magaly Parreira do Prado[1]

Resumo: "Deepfake de áudio" faz parte do complexo de deepfakes, que é uma das formas das fake news (FN) que assolam o planeta no intuito de enganar os incautos. O objetivo do estudo é entender de que maneira o deepfake de áudio contribui a propagar e exercer uma ascendência sobre o público, desvirtuando sua maneira de pensar, sendo a incógnita por trás dos algoritmos complexos que as inflamam. A proposta é apontar como se dá a relação entre os dados (usurpados) e a análise e monitoramento das mídias para "melhor" direcionar quem receberá cada tipo de desinformação. A hipótese central é que a falta de proteção dos nossos dados pessoais faz com que eles virem a matéria-prima do uso indiscriminado pelos produtores de informações fraudulentas. O poder dos recursos de técnicas de Inteligência Artificial e um rol de ferramentas para fabricá-las e detectá-las são levantados. Ao escrutinar as deepfakes de áudio embutidas ou não nas de vídeo – mas fazendo tanto estrago quanto, em sua disseminação descontrolada –, examinamos um caso danoso de clonagem e manipulação de voz como relato de risco. Em conclusão, a afronta, à ética da informação é discutida.

Palavras-chave: Deepfake. Deepfake Áudio. Inteligência Artificial. Algoritmos.

1 Pesquisadora de pós-doutorado na Cátedra Oscar Sala, do Instituto de Estudos Avançados, da Universidade de São Paulo e na Escola de Comunicações e Artes (ECA). Doutora em Comunicação e Semiótica e mestra em Tecnologias da Inteligência e Design Digital, ambos pela Pontifícia Universidade Católica São Paulo (PUC-SP). Graduada em Jornalismo, e pós-graduada em Comunicação Jornalística pela Faculdade Cásper Líbero. ORCID: orcid.org/0000-0003-2792-0264. CV Lattes: lattes.cnpq.br/7192215883585882. E-mail: magalyprado@usp.br.

Audio deepfake: manipulation simulates real voice to portray someone saying something they did not say

Abstract: "Audio deepfake" is a part of the complex of deepfake, which is a form of the fake news (FN) produced with the purpose of deceiving the unwary. The study aims at understanding how audio deepfake spreads and exerts an ascendancy over the public, distorting their way of thinking through unknown algorithms. The idea is to point out how the relationship between the (usurped) data and the analysis and monitoring of the media takes place to "better" direct those who receive the various kinds of misinformation. The main hypothesis is that the lack of protection of our personal data turns them into the raw material for the indiscriminate use of fraudulent information. The power of the resources of Artificial Intelligence techniques is examined, and a list of tools to manufacture and detect them is set up. In scrutinizing audio deepfakes inserted or not in video deepfakes and the damage they cause in their uncontrolled dissemination, the paper analyzes a malicious case of voice cloning and manipulation camouflaged as a risk report. The challenge to the ethics of information is discussed.

Keywords: Deepfake. Audio Deepfake. Artificial Intelligence. Algorithms.

Introdução: das fake news às deepfakes de vídeo e de áudio

A sociedade dataficada passou a ouvir falar das denominadas fake news (FN) mais extensivamente nos últimos cinco anos, quando se deu tamanho espalhamento pernicioso em sites impostores, nas redes sociais e em mensageiros instantâneos. Eis o problema de fundo desta pesquisa. A propagação viral sobreveio por meio de textos com informações inverídicas e, em dadas ocasiões, mal intencionadas. Quase imediatamente, as imagens (de modo geral, acompanhando textos, para melhor atrair a leitura), que nem sempre seguiam a linha da fraude, também passaram, cada vez mais, a reforçar o mesmo intuito: o de agir de maneira dissimulada, iludir ou mesmo tapear, afinal, a falsidade não é evidente aos olhos comuns da maioria.

Como um extremo das FN na era cibernética, sucedeu-se o tipo de FN no formato de vídeo, a chamada deepfake (DF), tecnicamente mais complicada de produzir. Nela, como em qualquer material videofônico, une-se o texto, a imagem (estática ou em movimento) e o áudio. Edita-se de forma a deturpar, tirar do contexto, degenerar etc., na intenção maior de provocar ainda mais a já instalada desordem. Contudo, paralelamente à DF de vídeo, a mais conhecida e disseminada, surgiu a DF de áudio (objeto definido para este estudo), cujo foco são as manipulações de voz (pré-gravadas) disponibilizadas na rede, com a possibilidade de emparelhar a ruídos (burburinhos para simular ambientes, lugares, momentos etc.), colhidos exclusivamente ou retirados de bancos de som digitais. Deste modo, transitaram entre os formatos que deterioram e confundem a audiência, junto à qual as FN superabundam para se juntar ao obscurantismo comunicacional e todo o desvio e riscos que ele causa.

Humanos, máquinas e coisas deflagram a miríade das FN no espaço numérico. Motivados por crenças e com aversão à irritação, muitos acabam aceitando como verdade tudo o que lhes é dito. Assim sendo, é axiomático, para quem é das áreas das ciências da comunicação e das ciências da informação, escrutinar o problema no modo contínuo.

Na era digital, com a linguagem codificada e a avalanche de sistemas lesivos, como o do *spam*, dos intrusos *cookies* de rastreamento (para o eufemismo de afirmar ser "melhor" para "compreender" as pessoas) e dos mecanismos de buscas (considerados inofensivos em seu início e até amigos por nos "ajudar", refinando pesquisas), a retórica da aplicação da Inteligência Artificial (IA) é usada para a mais alta e perspicaz enganação.

Os problemas são inúmeros, mas os principais recaem em: o que dizer da computação cognitiva, que tenta imitar os humanos, como, por exemplo, em estilos de escrita e fala? Onde querem chegar, além de mitigar situações ou atrair imprudentes com promessas fictícias? O que alegar a quem extrapola e cria *deepfake news*? Quede a ética?

Antes de tratarmos das deepfakes – e, especificamente, das DF de áudio, *corpus* deste estudo –, é preciso incorporar à discussão a intervenção das FN e a capacidade da *deep learning* (DL), isto é, da *aprendizagem profunda, por meio de uso maciço de dados*.

> *Deep learning* é uma tecnologia disruptiva de aprendizado de máquina com alto desempenho na resolução de problemas complexos e flexibilidade de aplicação de seus algoritmos. Dentre as principais aplicações estão o reconhecimento de imagens, voz e texto; a previsão de eventos e desenvolvimento de sistemas de recomendação, por exemplo para tornar a experiência do cliente única; e a detecção de anomalias, por exemplo para detecção de fraudes. (DEEP LEARNING, 2017)

Grosso modo, as FN atingem, diariamente, milhões de pessoas, tumultuando a cultura democrática, desacreditando o jornalismo e atrapalhando o livre saber da esfera pública. É importante frisar que, na condição da imprensa, se elas são *fakes*, não são *news*, embora sejam assim conhecidas as informações fraudulentas proliferadas na atual era da pós-verdade,[2] a qual o mundo vem atravessando de forma descontrolada. O adjetivo *fake* (falso) sequer coaduna com o substantivo *news* (no caso, notícias). Portanto, por motivos óbvios: para um fato se tornar notícia, a prioridade, entre as várias regras éticas da imprensa, é que ele seja verdadeiro, ou melhor, uma verdade factual.[3] A "notícia falsa", logo, não é notícia;

2 Ignácio Ramonet (2018) trata de uma das vertentes da pós-verdade, ao dizer que a vitória de Donald Trump nas eleições norte-americanas de 2016 também demonstrou que "a verdade não é mais necessária. Para ganhar a eleição, você não precisa se apoiar na verdade. A verdade não é relevante, não é mais pertinente, e por isso se colocou esse conceito de pós-verdade ou verdade alternativa: você tem a sua, eu tenho a minha."

3 Eugênio Bucci (2019, p. 22) nos lembra: "Hannah Arendt ressalta que a verdade factual é pequena, frágil, efêmera. Como um primeiro registro dos acon-

apenas tenta ser um simulacro de notícia – mas isso não a impede de circular e nem de ter consequências desastrosas.

Como método para abordar a problemática, partiremos por categorizar as FN e as DF. Mostraremos a relação entre os dados, a análise e o monitoramento de mídias; o poder da IA nos experimentos e na fabricação das DF e, em específico, as DF de áudio e as ferramentas de confecção e de verificação de FN e DF. Apresentaremos o relato de um caso de detecção de DF de áudio e a ética colocada à prova.

Categorias de fake news

A expressão fake news abrange diversas categorias: notícias fraudulentas ou frágeis; informação falsa (em geral, com fontes inventadas), manipulada, adulterada ou fabricada (com a intenção de ludibriar); desinformação (criada para prejudicar) ou má informação (sem apuração ou mal apurada [*misinformation*], ou mesmo usando a verdade, muitas vezes fora de contexto, para causar danos [*mal-information*]); notícias antigas requentadas; sensacionalismo (próprio dos tabloides); mentiras, maquiagens, boatos, fatos alternativos etc. Todas ameaçam a qualidade do jornalismo e, por conseguinte, a formação da opinião coletiva.

O que vemos é que as FN, também conhecidas contraditoriamente como notícias, mesmo com o adendo de que são mentirosas, assolam a comunicação e, infelizmente, ainda não foram definidas de forma clara – quer na maneira como são ditas à boca pequena, quer na forma da lei – e tampouco receberam uma solução plausível. Ainda não se desenvolveu um mecanismo efetivo para contrastar com o conteúdo de qualidade. Trabalham como uma máquina de propaganda.

O conceito de "mídia recontextualizada" vislumbra outra dimensão:

> Mídia recontextualizada é qualquer imagem, vídeo ou clipe de áudio que foi retirado de seu contexto original e reformulado para um propósito ou quadro narrativo totalmente diferente. Enquanto falsificações baratas, mais amplamente, alteram a mídia, a mídia recontextualizada usa imagens, vídeo ou áudio

tecimentos, um primeiro – e precário – esforço de conhecer o que se passa no mundo, a verdade factual é mais vulnerável a falsificações e manipulações. Mesmo assim, a verdade factual é facilmente reconhecível por todos, pelos homens e mulheres normais, comuns [...]. No nível dos fatos, dos acontecimentos, dos eventos que todos vemos e que todos temos condições de verificar e comprovar no uso das habilidades e das faculdades comuns dos seres humanos comuns, não há ninguém que não saiba divisar as distinções entre a verdade factual e a invenção deliberada de falsidades com o objetivo de esconder os fatos."

> inalterados, mas os apresenta em um contexto novo ou falso de acordo com a agenda dos manipuladores. Durante os primeiros protestos contra o assassinato de George Floyd, em junho de 2020, muitas imagens recontextualizadas se espalharam nas redes sociais. Um [vídeo] mostrava uma imagem do programa de TV Designated Survivor, mas alegava que era de um protesto Black Lives Matter; outra foto de um McDonald's queimando em 2016 foi reformulada como se fosse um protesto atual. (RECONTEXTUALIZED..., 2021, tradução nossa)

Ao fim e ao cabo, importa descobrir de que forma e com que intensidade o fenômeno dos algoritmos de FN afeta a cultura democrática, bem como as consequências desse fenômeno – difícil de imaginar seu esgotamento, inclusive. O objetivo é entender de que maneira as FN se propagam estrondosamente e exercem uma ascendência sobre o público, desvirtuando sua maneira de pensar, é a incógnita por trás dos algoritmos complexos que as inflamam. A velocidade da ação algorítmica a nos trazer ilações é fato.

Não é de surpreender que, cada vez mais, vamos nos deparar com as *big techs*, as plataformas, os buscadores, enfim, especialmente empresas de mensageria instantânea, com ou sem redes sociais embutidas, mostrando a veleidade inerente em seus usuários (como viciados mesmo), em uma espécie de demonstração da circularidade de algoritmos de IA na confecção de peças inautênticas de toda ordem (ou melhor, desordem), das mais simples, como as mensagens de texto, às mais elaboradas, como em áudio ou em audiovisual, que requerem edição para a deformidade intencional. Mais à frente, neste texto, entraremos de cabeça em aspectos que ultrapassam o humano e o racional. Por enquanto, ainda tateamos na vagueza de se tentar entender a IA quando da algoritmização por trás disso tudo. Mas é possível constatar algumas das intenções com a ajuda de ordem teórica. Karen Hao (2021) cita Hany Farid, que colabora com o Facebook, para entender a desinformação baseada em imagem e vídeo na plataforma: "Quando se está no negócio de maximizar o engajamento, não se está interessado na verdade."

Relação entre dados

A dúvida que paira é saber até quando teremos os dados armazenados, ou seja, até que ponto haverá espaço suficiente para esse armazenamento e, de quebra, com cibersegurança. Outra questão é como garantir o direito à nossa privacidade. A hipótese central é que a falta de proteção dos nossos dados pessoais faz com que eles virem a matéria-prima do uso indiscriminado pelos produtores de FN e DF.

Os sistemas de IA funcionam com dados coletados de várias fontes, como *cookies* de relacionamento, e-mail, dados online etc., que podem ter várias formas, como, por exemplo, áudio, vídeo ou texto.

> O trabalho do cientista de dados é coletar, armazenar e entender os dados (tornando simples a análise via visualização e estatísticas descritivas) para preparar os dados para modelos de IA. A qualidade do trabalho dos cientistas de dados é essencial para que os sistemas de IA funcionem corretamente. Na verdade, há um ditado que diz que "sua IA é tão boa quanto seus dados" e os dados terão uma influência direta nas ações ou decisões produzidas pelos sistemas de IA. (SHNURENKO; MUROVANA; KUSHCHU, 2020, p. 5, tradução nossa)

Os pontos fortes da relação entre os dados vêm principalmente "das técnicas de aprendizado de máquina [*machine learning*], seja de reforço, aprendizagem supervisionada ou não supervisionada, usando grandes conjuntos de dados – verbais, textuais, imagens ou fluxos de vídeo. Talvez o mais importante é que alguns dos sistemas de IA podem estar trabalhando em tempo real" (SHNURENKO; MUROVANA; KUSHCHU, 2020, p. 5, tradução nossa). Contudo, há de se levar em consideração o fato de que os mecanismos inerentes de IA mais amplamente usados "estão relacionados à máquina com *deep learning* e esses mecanismos tornam possível: classificar (medir relevância ou relacionamentos), prever (fazer afirmações sobre o que vem a seguir ou o que vai acontecer no futuro) e priorizar ou otimizar, especialmente por meio de métodos evolutivos de IA, como algoritmos genéticos" (SHNURENKO; MUROVANA; KUSHCHU, 2020, p. 18).

Machine learning tornou-se uma área de investimento e de pesquisa proeminente com o propósito de oferecer aos computadores a capacidade de aprender com base em exemplos e experiências, é o que diz Jones (2017):

> Após pesquisas investigativas sobre IA e aprendizado de máquina, por volta do ano 2000, surgiu o *deep learning*. Os cientistas da computação usavam redes neurais em várias camadas com novas topologias e métodos de aprendizado. Essa evolução das redes neurais resolveu com êxito problemas complexos em vários domínios. Na década passada [anos 2000], surgiu a computação cognitiva, cujo objetivo é construir sistemas que possam conhecer e interagir naturalmente com humanos.

O DL transforma o reconhecimento de fala e imagem de forma mais precisa; "é um conjunto relativamente novo de métodos que está mudando o aprendizado de máquina de formas fundamentais. O *deep*

learning não é um algoritmo propriamente dito, mas uma família de algoritmos que implementam redes profundas com aprendizado sem supervisão" (JONES, 2017).

Ao aprender – profundamente ou não – com a máquina, *bots*, *chatbots* e ciborgues incrementam esses desenvolvimentos invasores, ajudando a falsear, replicar e, sobretudo, viralizar no ciberespaço um conteúdo de interesse específico, produzido com rigor minucioso, de acordo com o resultado da análise dos dados, para direcionar (e modular) de forma algorítmica os incautos, os solitários na internet ou os agregados às comunidades virtuais, desde que sejam influenciáveis, indecisos, crentes e propícios à transdução (PRADO, 2019, p. 70).

Claire Wardle (2017, tradução nossa) discorre sobre como esse conteúdo fraudulento da desinformação é divulgado. De acordo com ela, as pessoas compartilham FN porque não verificam seu conteúdo: "Parte disso está sendo promovido por grupos que estão deliberadamente tentando influenciar a opinião pública, e outra está sendo disseminada como parte de sofisticadas campanhas de desinformação, por meio de redes de *bots* e fábricas de *trolls*". A autora compreende que "o termo '*troll*' é mais frequentemente usado para se referir a qualquer pessoa que assedia ou insulta outros online. No entanto, também foi usado para descrever contas controladas por humanos que executam atividades semelhantes a *bot*" (WARDLE, 2018, tradução nossa).

> *Trolls* independentes são amadores que espalham informações inflamatórias para causar distúrbios e reações em sociedade brincando com as emoções das pessoas [...]. Por exemplo, postagem audiovisual manipulada com conteúdo racista ou sexista podem promover o ódio entre os indivíduos. Opostos a *trolls* independentes que espalham informações falsas para sua satisfação, os *trolls* contratados farão o mesmo trabalho para obter benefícios monetários. Diferentes atores, como partidos políticos, empresários e empresas contratam rotineiramente pessoas para forjar notícias relacionadas a seus concorrentes e divulgá-las [...]. Por exemplo, de acordo com um relatório publicado pela inteligência ocidental [...], a Rússia está executando "fazendas de *trolls*", onde os *trolls* são treinados para afetar as conversas relacionadas a questões nacionais ou internacionais. De acordo com estes relatórios, vídeos deepfake gerados por *trolls* contratados são a mais nova arma na guerra de notícias fabricadas em curso que pode trazer um efeito mais devastador para a sociedade. (MASOOD *et al.*, 2021, p. 3, tradução nossa)

Os *bots* são um exemplo claro de ferramenta usada pelos esquemas de FN, nas redes sociais, para espalhar conteúdos errôneos. Em termos de velocidade de propagação de conteúdo, é impossível competir com os

bots, que, dessa forma, prejudicam os legítimos e espontâneos debates democráticos entre os cidadãos, atingindo negativamente a esfera pública.

> No estudo de desinformação e manipulação de mídia, os *bots* normalmente se referem a contas de mídia social que são automatizadas e implantadas para fins enganosos, como para amplificar artificialmente uma mensagem, jogar uma tendência ou algoritmo de recomendação ou aumentar as métricas de engajamento de uma conta. Essas contas são normalmente controladas centralmente ou em coordenação umas com as outras. (BOTS, 2021, tradução nossa)

As FN – tanto acionadas por humanos quanto por *bots*, programados por humanos, obviamente – se alastram exatamente onde a excessiva maioria do público-alvo (aquele que deverá ser atingido) está: nas redes sociais e em grupos de mensageria instantânea. Logo, para ajudar na viralização das FN e atingir mais pessoas, inclusive indo além das previamente escolhidas, a fábrica das FN reforça sua atuação com o uso de *bots*, como arautos da informação. Assim, faz com que a propagação própria da internet, de todos para todos (*ipsis litteris*, pessoas, ciborgues, dispositivos, coisas), seja acelerada e que a alta proliferação da sabotagem tenha alcance em uma progressão desmesurada.

"A existência de robôs e a participação deles na vida cotidiana era, há pelo menos duas décadas, matéria de experiência científica, estava no âmbito da imaginação, da ficção científica", dizem Luziane Leal e José Filomeno de Moraes Filho (2019, p. 344). Contudo, a evolução das tecnologias mudou esse cenário e, conforme os autores, colocou os robôs em ambientes inimagináveis, como na construção da opinião pública, na escolha subjetiva do eleitor por seus candidatos e, assim, na participação direta dos rumos da democracia.

Para se direcionarem ao público específico, as plataformas captam, antes, por rastreamento, a matéria-prima, ou seja, os dados das ações das pessoas, que revelam suas características e seu comportamento nas redes. Com a extração desses dados (emocionais, biométricos etc.), a análise prediz padrões comportamentais e, assim, como rastros de dados provocam outras camadas de dados, é possível fazer correlação para influenciar as próximas ações do público-alvo, uma forma de modular o pensamento das pessoas escolhidas.

Análise e monitoramento de mídia

Uma das formas contemporâneas de conhecer melhor o público que se pretende atingir é a análise de mídia. Monitorar para recolher os

dados e os rastros de determinados perfis, para escolher quem interessa que caia no manuseio do direcionamento de FN – seja ele de texto, fotografia, audiovisual, DF etc. –, ficou bem mais fácil com a quantidade cada vez maior de dados disponibilizados pelos próprios usuários, sejam dados vazados, sejam dados comprados pela indústria de FN.

Lev Manovich (2018, tradução nossa) acredita que a análise de mídia tecnológica é como um novo estágio no desenvolvimento da moderna mídia tecnológica. O autor diz que "nós, como pesquisadores acadêmicos, vivemos na 'sombra' de um mundo de redes sociais, recomendações, aplicativos e interfaces que usam análises de mídia. [...] E esta etapa é caracterizada pela análise algorítmica em larga escala das interações". Trata-se de interações entre "mídia e usuário e o uso dos resultados na tomada de decisão algorítmica, como publicidade contextual, recomendações, pesquisa e outros tipos de recuperação de informação, filtragem de resultados de pesquisa e postagens de usuários". E mais: "classificação, detecção de plágio, impressão digital de vídeo, categorização de conteúdo de fotos de usuários, produção automática de notícias etc."

Manovich (2018, tradução nossa) ressalta, ainda, que estamos apenas no começo deste estágio. "Dada a trajetória da automação gradual de mais e mais funções na sociedade moderna usando algoritmos", o autor espera que "a produção e a personalização de muitas formas de 'cultura comercial' (caracterizadas por convenções, expectativas de gênero e modelos) também sejam gradualmente automatizadas". Assim, no futuro, "as plataformas de distribuição digital já desenvolvidas e a análise de mídia serão acompanhadas pela terceira parte: a geração de mídia algorítmica". Em outras palavras, o modelo matemático inserido no cotidiano da informação.

O poder da IA para uma série de experiências dos problemas do mundo real

Kai-Fu Lee (2018, p. 18, tradução nossa) aponta que redes neurais e DL (termos que podem ser compreendidos como "imitação do cérebro", numa tradução popular) requerem "grandes quantidades de duas coisas: poder de computação e dados. Os dados 'treinam' o programa para reconhecer padrões, fornecendo muitos exemplos, e o poder de computação permite que o programa analise esses exemplos em altas velocidades." Ao lembrar que tanto os dados quanto o poder de computação eram "escassos no início do campo [da computação] na década de 1950", o autor

reforça que "nas décadas seguintes, tudo isso mudou". As próprias redes ainda eram severamente limitadas no que era possível fazer. "Resultados precisos para problemas complexos exigiram muitas camadas de neurônios artificiais, mas os pesquisadores não encontraram uma maneira de treinar com eficiência essas camadas à medida que foram adicionadas", conta Lee, que ainda complementa: "A grande ruptura técnica do *deep learning* finalmente chegou em meados dos anos 2000, quando o principal pesquisador Geoffrey Hinton descobriu uma maneira de treinar com eficiência essas novas camadas em redes neurais."

> O resultado foi como dar esteroides às velhas redes neurais, multiplicando seu poder de realizar tarefas como reconhecimento de fala e objetos. Em breve, essas redes neurais aprimoradas – agora rebatizadas como "deep learning" – poderiam superar os modelos mais antigos em uma variedade de tarefas. Depois de décadas passadas à margem da pesquisa de IA, as redes neurais atingiram o *mainstream* durante a noite, desta vez na forma de *deep learning*. (LEE, 2018, p. 18, tradução nossa)

No entanto, é bom frisar que o DL se desenvolve continuamente. Pesquisadores, futuristas e CEOs de tecnologia já começaram a falar sobre "o enorme potencial do campo para decifrar a fala humana, traduzir documentos, reconhecer imagens, prever o comportamento do consumidor, identificar fraudes, tomar decisões sobre empréstimos, ajudar os robôs a 'ver' e até mesmo dirigir um carro" (LEE, 2018, p. 19, tradução nossa).

> Fundamentalmente, esses algoritmos usam grandes quantidades de dados de um domínio específico para tomar uma decisão que otimiza para um resultado desejado. Ele faz isso treinando a si mesmo para reconhecer padrões profundamente enterrados e correlações conectando os muitos pontos de dados para o resultado desejado. (LEE, 2018, p. 19, tradução nossa)

É sempre prudente questionar: resultado desejado para quem? Quem está por trás de tais sistemas de IA? "Fazer isso requer uma grande quantidade de dados relevantes, um forte algoritmo, um domínio estreito e um objetivo concreto", sinaliza Lee (2018, p. 19, tradução nossa). Ele argumenta que, se houver "falta de qualquer um destes, as coisas desmoronam. Poucos dados? O algoritmo não tem exemplos suficientes para descobrir correlações significativas. Um objetivo muito amplo? O algoritmo carece de *benchmarks* [avaliações corporativas] claros para atingir na otimização". Como vemos, não parece ser tão fácil lidar com os dados, tirar sentido deles, entrevistá-los.

Deepfakes: mídia fabricada produzida com Inteligência Artificial

No fundo, é possível resumir a descrição e crítica de DF: trata-se da heurística ao avesso, quando se descobrem os não-fatos.

> Em dezembro de 2017, um usuário do *Reddit* utilizando ferramentas de Inteligência Artificial e Aprendizado de Máquina de código aberto, como o *Keras* e o *TensorFlow* (esse último, do Google), criou um algoritmo para treinar uma rede neural a mapear o rosto de uma pessoa no corpo de outra, *frame* por *frame*. Ao invés de usar edição manual como antes, o usuário através da ferramenta (que recebeu o nome de *Deep Fake*) precisa apenas de uma fonte para reconhecer o modelo do rosto da "vítima", mapear a estrutura da cabeça-destino e fazer a sobreposição. O software é capaz de ajustar a movimentação do vídeo original ao novo rosto e isso inclui expressões faciais e movimentos labiais. (GOGONI, 2018)

Conforme definição de Michael K. Spencer (2019), DF são, essencialmente, identidades falsas criadas com o DL, "por meio de uma técnica de síntese de imagem humana baseada na IA", a qual é "usada para combinar e sobrepor imagens e vídeos preexistentes e transformá-los em imagens ou vídeos 'originais', utilizando a tecnologia de GAN (*Generative Adversarial Network*, ou rede geradora antagônica)". O autor acrescenta que, desde 2019, "também estamos vendo uma explosão de faces *fake*, através das quais a IA é capaz de conjurar pessoas que não existem na realidade, e que têm um certo fator de influência". O assunto, que causa assombro, pode ficar, certamente, no escopo de outro artigo.

Os DF podem ser categorizados nos seguintes tipos: "i) troca de rosto; ii) dublagem; iii) fantoche-mestre; iv) rosto síntese e manipulação de atributos; e v) deepfakes de áudio", conforme Momina Masood *et al.* (2021, p. 1, tradução nossa), nosso principal quadro de referência. Especificamente sobre os DF de áudio, "também conhecido como clonagem de voz", constata-se que "se concentra na geração da voz do locutor usando técnicas de DL para retratar o locutor dizendo algo que não disse" (MALIK; MALIK; BAUMANN apud MASOOD *et al.*, 2021, p. 2, tradução nossa).

O uso estratégico de recursos visuais na desinformação é "provavelmente motivado pela premissa de que as imagens são uma representação direta da realidade e, como tal, são percebidas como mais credíveis do que formas de comunicação mais abstratas, como as palavras", explicam Paul Messaris e Linus Abraham (2001 apud HAMELEERS *et al.*, 2020, p. 297, tradução nossa). Os autores vão adiante:

> Essa qualidade realista das fotos significa que o público pode desconfiar menos da desinformação na forma multimodal do que na forma textual. A desinformação multimodal pode, portanto, ser percebida como mais confiável do que a desinformação textual. Testar tal proposição é especialmente importante nos dias de hoje, uma vez que a manipulação de imagens (e até mesmo a manipulação de vídeos) está se tornando mais fácil com a ampla disponibilidade de softwares de edição de imagens.
> Na verdade, a comunicação visual tem uma longa história como ferramenta de propaganda (Bagchi, 2016), e um crescente corpo de pesquisas aponta para o papel crucial dos recursos visuais ao lado do texto na comunicação política multimodal (Graber, 1990). Muito deste trabalho está relacionado ao enquadramento visual e multimodal – a capacidade integrativa de imagens ao lado do texto para destacar um aspecto saliente de uma questão (de Vreese, 2005; Entman, 1993; Grabe & Bucy, 2009) – que pode ter um impacto ainda mais forte no público do que apenas dicas textuais (Powell, Boomgaarden, de Swert, & de Vreese, 2015). (HAMELEERS *et al.*, 2020, p. 283, tradução nossa)

Definimos desinformação visual com base em Michael Hameleers *et al.* (2020, p. 283, tradução nossa): "uso de imagens por agentes de desinformação para apresentar deliberadamente uma imagem enganosa ou fabricada da realidade. Como as pessoas tendem a ser menos críticas aos recursos visuais (Wardle, 2017), é importante avaliar o impacto da desinformação multimodal".

Seguimos as conceituações existentes sobre falsidade comunicativa com base em intenções e facticidade – por exemplo, as classificações de Tandoc Jr. *et al.* (2017) e Wardle (2017) –, adicionando o componente multimodal, para distinguir diferentes formas de desinformação visual:

- emparelhar imagens reais com textos enganosos (descontextualização);
- cortar ou descontextualizar os recursos visuais para tornar certos aspectos das questões mais salientes de uma forma direcionada a um objetivo (ressignificação);
- manipular recursos visuais para apresentar uma realidade diferente (tratamento visual);
- fabricar conteúdo combinando imagens manipuladas com texto manipulado (manipulação multimodal). (HAMELEERS *et al.*, 2020, p. 281, tradução nossa)

Central para o papel dos recursos visuais na desinformação é "sua indicialidade (Messaris & Abraham, 2001). Isso descreve a qualidade real dos recursos visuais, pois eles são uma representação direta de objetos físicos e eventos no ambiente não mediado, enquanto as palavras são símbolos abstratos que não têm nenhuma semelhança física com seus referentes (Grabe & Bucy, 2009)" (HAMELEERS *et al.*, 2020, p. 284, tradução nossa). "Ao ler, deve-se extrair significado semântico dos símbolos escritos e, em seguida, criar uma reconstrução imaginária de um evento. Em contraste, a adição de uma imagem a um texto fornece um 'índice' da realidade e empresta uma qualidade evidencial inerente a uma história", destacam Messaris e Abraham (2001 apud HAMELEERS *et al.*, 2020, p. 284, tradução nossa). Contudo, na visão de Dolf Zillmann, Rhonda Gibson e Stephanie Sargent (1999 apud HAMELEERS *et al.*, 2020, tradução nossa), a explicação é possível influenciar as percepções da audiência sobre os acontecimentos noticiosos e, assim, induzir os leitores a ignorar o fato de que as imagens são construções artificiais feitas pelo ser humano. Nesses casos, quando usados na desinformação, os recursos visuais adquirem um poder propagador de falsidades, porque são vistos como mais confiáveis e do que textos.

Deepfake de áudio: mídia sintética de clonagem e geração de voz usa técnicas de DL

Ao acompanhar o avanço do hábito de ouvir (e gravar) áudios, que circulam em abundância nos mensageiros instantâneos e nos *audiocasts*, era de se esperar que a audiofonia ganhasse corpo no espectro das DF. "Ao contrário dos vídeos deepfake, menos atenção foi dada à detecção de deepfakes de áudio. Nos últimos anos, a clonagem de voz também se tornou muito sofisticada", consideram Masood *et al.* (2021, p. 2, tradução nossa). Eles acrescentam que "a clonagem de voz não é apenas uma ameaça à verificação automática de sistemas de *speakers*, mas também para sistemas controlados por voz implantados nas configurações da Internet das Coisas." Dizem, ainda, que a clonagem de voz tem "tremendo potencial para destruir a confiança pública e capacitar criminosos para manipular negociações comerciais ou privadas".

A justificativa é que não há pesquisas publicadas recentemente sobre geração e detecção de DF com foco em geração e detecção de modalidades de áudio, conforme Masood *et al.* (2021, p. 3, tradução nossa) sinalizam: "A maioria das pesquisas existentes se concentra apenas em revisão de imagens DF e detecção de vídeo." Embora todas as categorias

de multimídia falsa "(ou seja, notícias falsas, imagens falsas e áudio falso) possam ser fontes de desinformação, espera-se que DF baseados em audiovisual sejam muito mais devastadores. Este dano não se limita a visar indivíduos; em vez disso, DF podem ser usados para manipular eleições ou criar situações belicistas."

Ferramentas de IA

Joaquin Quiñonero Candela, líder da equipe de IA do Facebook, foi quem transformou tal rede social em empresa movida a IA e em uma potência no uso dessa tecnologia. "Em seis anos, ele criou alguns dos primeiros algoritmos para direcionar os usuários com conteúdo precisamente adaptado aos seus interesses, e então difundiu esses algoritmos por toda a empresa", relata Karen Hao (2021, tradução nossa), em reportagem à *MIT Technology Review*.

Nos últimos dois anos, a equipe de Quiñonero desenvolveu a ferramenta original de *Kloumann*, chamada *Fairness Flow*. Ela permite que os engenheiros meçam a precisão dos modelos de *machine learning* para diferentes grupos de usuários. "Eles podem comparar a precisão de um modelo de detecção de rosto em diferentes idades, gêneros e tons de pele, ou a precisão de um algoritmo de reconhecimento de voz em diferentes idiomas, dialetos e sotaques." (HAO, *ibid.*). O *Fairness Flow* também vem com um conjunto de diretrizes para ajudar os engenheiros a entender o que significa treinar um modelo "justo". Todavia, "um dos problemas mais espinhosos em tornar os algoritmos justos é que existem diferentes definições de justiça, que podem ser mutuamente incompatíveis" (HAO, 2021, tradução nossa). Outras ferramentas utilizadas para verificar imagens são elencadas pelo site *datajournalism.com* e servem como elementos que se correlacionam à exposição de como o nosso problema é feito, produzido:

> Uma imagem em particular é uma representação real do que está acontecendo?
> *Foto Forensics* [fotoforensics.com]: este site usa análise de nível de erro (ELA) para indicar partes de uma imagem que podem ter sido alteradas. O ELA procura diferenças nos níveis de qualidade da imagem, destacando onde as alterações podem ter sido feitas.
> *Pesquisa Google por imagem* [support.google.com/websearch]: ao enviar ou inserir o URL de uma imagem, os usuários podem encontrar conteúdo como imagens relacionadas ou semelhantes, sites e outras páginas usando a imagem específica.

> *Jeffrey's Exif Viewer* [exif.regex.info/exif.cgi]: uma ferramenta online que revela as informações do Exchangeable Image File (EXIF) de uma foto digital, que inclui data e hora, configurações da câmera e, em alguns casos, localização GPS.
> *JPEGSnoop* [sourceforge.net/projects/jpegsnoop]: um aplicativo gratuito apenas para Windows que pode detectar se uma imagem foi editada. Apesar do nome, ele pode abrir arquivos AVI, DNG, PDF, THM e JPEG embutidos. Ele também recupera metadados, incluindo: data, tipo de câmera, configurações de lente etc.
> *TinEye* [tineye.com]: um mecanismo de busca reversa de imagens que conecta imagens a seus criadores, permitindo que os usuários descubram a origem de uma imagem, como ela é usada, se existem versões modificadas e se existem cópias de maior resolução. (VERIFICATION..., 2021, tradução nossa)

WaveNet, Tacotron e *deep voice*

Como dito anteriormente, em outras palavras, a manipulação de áudio sintetizado por IA "é um tipo de deepfake que pode clonar a voz de uma pessoa e representar essa voz dizendo algo ultrajante, que a pessoa nunca disse. Avanços recentes em algoritmos sintetizados por IA para síntese de fala e clonagem de voz mostraram um potencial para produzir vozes falsas realistas que são quase indistinguíveis do discurso genuíno" (MASOOD *et al.*, 2021, p. 15, tradução nossa).

Esses algoritmos podem gerar fala sintética que soa como

> o falante alvo com base no texto ou declarações do falante alvo, com resultados altamente convincentes (Arik *et al.*, 2018; Lorenzo-Trueba *et al.*, 2018). A voz sintética é amplamente adaptada para o desenvolvimento de diferentes aplicações, como dublagem automatizada para TV e cinema, *chatbots*, assistentes de IA, leitores de texto e vozes sintéticas personalizadas para pessoas com deficiência vocal. (MASOOD *et al.*, 2021, p. 15, tradução nossa)

Além disso, vozes sintéticas/falsas, alertam os autores, "tornaram-se uma ameaça crescente aos sistemas biométricos de voz e estão sendo usados para fins maliciosos, como ganhos de políticos, notícias falsas e golpes fraudulentos etc. Uma síntese de áudio mais complexa poderia combinar o poder da IA e edição manual." (MASOOD *et al.*, 2021, p. 15).

Modelos de síntese de voz alimentados por rede neural, como, por exemplo, *Tacotron*, do Google (um modelo de síntese de fala de ponta a ponta), *Wavenet*[4] ou *AdobeVoco*[5], podem gerar vozes sintética e falsas, mas

4 "*WaveNet*, desenvolvido pela *DeepMind* [adquirida pelo Google em 2014], em 2016, utiliza formas de onda de áudio brutas usando recursos acústicos, ou seja, espectrogramas, por meio de uma estrutura generativa que é treinada na fala gravada real. *WaveNet* é um modelo autorregressivo probabilístico que funciona determinando a distribuição de probabilidade do sinal acústico atual usando as probabilidades de amostras geradas" (MASOOD *et al.*, 2021, p. 15, tradução nossa).

5 Ver em: Jin et al. (2017).

com sons realistas convincentes, que se assemelham à voz da vítima, "a partir da entrada de texto para fornecem uma experiência de interação aprimorada entre humanos e máquinas, como a primeira etapa. Mais tarde, um software de edição de áudio, por exemplo *Audacity*, pode ser usado para combinar as diferentes peças de áudios originais e sintetizados para criar áudios mais poderosos" (MASOOD *et al.*, 2021, p. 15, tradução nossa).

Masood *et al*. (2021, p. 15, tradução nossa) continuam a explicar o processo: "Os modelos paramétricos enfatizam a extração de recursos acústicos a partir das entradas de texto fornecidas e convertendo-as em um sinal de áudio usando os *vocoders*." São resultados interessantes de texto para fala e são paramétricos, "devido ao desempenho aprimorado de parametrização de fala, modelagem do trato vocal e a implementação de redes neurais profundas evidentemente mostram o futuro da produção de fala artificial."

Uma abordagem promissora para melhorar as habilidades da IA, diz Hao (2021, tradução nossa), é expandir seus sentidos: "atualmente, IA com visão computacional ou reconhecimento de áudio pode sentir coisas, mas não pode 'falar' sobre o que vê e ouve usando algoritmos de linguagem natural. Mas e se você combinasse essas habilidades em um único sistema de IA? Poderiam esses sistemas começar a ganhar inteligência semelhante à humana?" Afinal, "um robô que pode ver, sentir, ouvir e se comunicar pode ser um assistente humano mais produtivo?" Hao justifica que as IAs "com múltiplos sentidos ganharão uma maior compreensão do mundo ao seu redor, alcançando uma inteligência muito mais flexível."

"Deepfakes de áudio são uma nova forma de ataque cibernético, com o potencial de causar graves danos a indivíduos devido a técnicas de síntese de voz altamente sofisticadas. [...] Golpes financeiros falsos assistidos por áudio aumentaram significativamente em 2019 devido à progressão em tecnologia de síntese de voz." (MASOOD *et al.*, 2021, p. 7, tradução nossa).

Técnica de observação em caso de deepfake audio

Entre os casos de fraude em áudio já registrados, vale recorrer a um para exemplificar como ocorrem. Em agosto de 2019, o CEO de uma empresa europeia, enganado por um áudio deepfake, fez uma transferência bancária de 243 mil dólares (HARWELL, 2019). "Um software de IA de imitação de voz foi usado para clonar a voz padrões da vítima treinando

algoritmos de ML usando gravações de áudio obtidas na internet. Se tais técnicas podem ser usadas para imitar a voz de um alto funcionário do governo ou um líder militar e aplicado em escala, poderia ter sérias implicações para a segurança nacional" (ARIK *et al.* apud MASOOD *et al.*, 2021, p. 6, tradução nossa).

O áudio DF foi demonstrado "em algumas demos de tecnologia chamativas. Mas a tecnologia também está começando a ser usada no mundo criminal. Lorenzo Franceschi-Bicchierai (2020, tradução nossa) relata que, em junho de 2020, um funcionário de uma empresa de tecnologia recebeu uma mensagem de voz "estranha e suspeita, em uma tentativa de fazer o funcionário enviar dinheiro para criminosos." A voz era "de uma pessoa que se identificou como CEO, pedindo 'assistência imediata para finalizar um negócio urgente'. Acontece que, apesar de parecer quase como o CEO, o correio de voz foi realmente criado com software de computador. Foi um deepfake de áudio, de acordo com uma empresa de segurança que investigou o incidente." A NISOS, uma empresa de consultoria de segurança com sede em Alexandria, Virgínia, "analisou o correio de voz e determinou que era falso, um áudio sintético projetado para enganar o receptor."

> O funcionário que recebeu o correio de voz, no entanto, não o aceitou e sinalizou para a empresa, que chamou a NISOS para investigar. Os pesquisadores da NISOS analisaram o áudio com uma ferramenta de espectrograma chamada *Spectrum3d*, na tentativa de detectar qualquer anomalia. "Você poderia dizer que havia algo errado no áudio", disse Dev Badlu, pesquisador da NISOS, à Motherboard. "Parece que eles basicamente pegaram cada palavra, cortaram e colaram novamente". (FRANCESCHI-BICCHIERAI, 2020, tradução nossa)

A questão torna-se crucial a ser debatida haja vista que sua ação foi danosa e, portanto, foi o que caracterizou a escolha do objeto deste estudo. "Badlu disse que sabia que era falso, porque havia muitos picos e vales no áudio, o que não é normal em conversas regulares. Além disso, ele acrescentou que, quando reduziu o volume do suposto CEO, os antecedentes eram 'absolutamente silenciosos', não havia nenhum ruído de fundo, o que era um claro sinal de falsificação" (FRANCESCHI-BICCHIERAI, 2020, tradução nossa).

Rob Volkert, outro pesquisador da NISOS, acredita que "os criminosos estavam testando a tecnologia para ver se os alvos os ligariam de volta". Em outras palavras, ele disse, "este foi apenas o primeiro passo de uma operação presumivelmente mais complexa que estava relativamente

perto de ter sucesso. 'Definitivamente parece humano. Eles marcaram essa caixa na medida em que: soa mais robótico ou mais humano? Eu diria mais humano'. Mas não parece suficiente o CEO" (FRANCESCHI-BICCHIERAI, 2020, tradução nossa).

"A capacidade de gerar áudio sintético estende o kit de ferramentas de um criminoso eletrônico, e o criminoso ainda precisa usar efetivamente as táticas de engenharia social para induzir alguém a agir". O relatório da NISOS aponta: "Criminosos e atores estatais potencialmente mais amplos também aprendem uns com os outros, de modo que esses casos de alto nível ganham mais notoriedade e sucesso, prevemos que mais atores ilícitos os tentem e aprendam com outros que abriram o caminho" (FRANCESCHI-BICCHIERAI, 2020, tradução nossa). Até agora, entrementes, "esses discursos sintetizados carecem de alguns aspectos da qualidade da voz, como expressividade, aspereza, soprosidade, estresse e emoção etc. específicos para uma identidade de destino", de acordo com Masood *et al.* (2021, p. 15, tradução nossa).

Ética colocada à prova

É urgente olhar além dos produtos fake, em si, para buscar maneiras de resguardar a ética na cacofonia informacional. Neste ínterim, Yuezun Li, Ming-Ching Chang e Siwei Lyu (2018 *apud* WARDLE, 2018, tradução nossa) traçam uma breve explicação: "Ao sintetizar diferentes elementos de arquivos de vídeo ou áudio existentes, a IA permite métodos relativamente fáceis para a criação de 'novos' conteúdos, nos quais os indivíduos parecem falar palavras e realizar ações que não são baseadas na realidade." Wardle (2018), por sua vez, norteia o que vem acontecendo com a prática *fake*, alertando ser provável que vejamos esses tipos de mídia sintética utilizados com maior frequência em campanhas de desinformação, à medida que tais técnicas se tornem mais rebuscadas.

A ênfase na "mídia sintética, coloquialmente conhecida como deepfakes, está em ascensão, com avanços na geração de texto, imagens e vídeo sintéticos, demonstrando o progresso da IA, mas também destacando o potencial para uso antiético ou perigoso", aponta o *Artificial Intelligence Index Report 2021* (HAI, 2021, p. 128, tradução nossa), relatório sobre 2020. Não à toa, os desafios éticos dos envolvidos em aplicações em IA se tornaram um ponto central, haja vista o crescimento de artigos que mencionam "ética" e palavras-chave relacionadas entre 2015 e 2020, embora o número médio de títulos de artigos da mesma correspondência à ética nas principais conferências de IA ainda permaneça baixo ao longo dos anos.

Karen Yeung, em sua contribuição ao relatório sobre IA e seu impacto nos padrões públicos, do *Committee on Standards in PuHAIblic Life*, é taxativa ao dizer que não é adequado empregar "argumentos jurídicos/técnicos para 'remendar' uma base legal 'implícita', dado que o poder, a escala e a intromissão dessas tecnologias criam sérias ameaças aos direitos e liberdades de indivíduos e para as bases coletivas de nossas liberdades democráticas" (LEADING..., 2020, tradução nossa).

Apesar do perigo que as regulações possam impedir o avanço da tecnologia em questão, é preciso debate-las a miúde. "A falta de parâmetros legais deixa em aberto uma lacuna jurídica,[6] regulatória e ética, com as más consequências que o uso de sistemas de IA sem governança pode trazer", aponta relatório da Transparência Brasil (2020, p. 5) intitulado *Recomendações de governança: uso de inteligência artificial pelo poder público*. Ao discutir e propor recomendações de governança para o uso de algoritmos de IA, o documento destaca que "é importante considerar a avaliação de riscos envolvendo ameaças reais e potenciais a direitos e ao espaço cívico, buscando alinhar promoção de inovação e tecnologia com responsabilidade pública e transparência" (TRANSPARÊNCIA BRASIL, 2020, p. 10).

Considerações

Como vimos com os autores que se debruçam sobre o tema, que nos despertam à crítica, as deepfakes, tanto de áudio quanto de vídeo, configuram-se como as mais nocivas peças de desinformação, pois enganam mais facilmente os crédulos, e até mesmo quem é esperto e sabe que as redes e os sites falsos estão repletos delas acaba caindo no "conto do vigário." De início, quem prestava mais atenção, podia ver sinais da manipulação, como os lábios levemente borrados ao proferir inverdades nos vídeos ou sinais de fala mal cortada, falta de nexo entre frases, excesso de cortes, de pausas etc. Com o passar do tempo, as produções foram se tornando mais precisas, e estão cada vez mais parecidas com suas vítimas. Por outro lado, as ferramentas também vão melhorando em usabilidade, como é praxe, ficando mais fáceis de serem manuseadas, dando a oportunidade de mais pessoas aproveitá-las.

6 Vale mencionar que o *Centre for Data Ethics and Innovation*, órgão criado pelo governo britânico em 2018 para assessorar na regulação do uso de Inteligência Artificial no Reino Unido, divulgou um relatório alertando para a necessidade de regulamentar a maneira como as redes sociais direcionam vídeos, anúncios e posts para seus usuários (UNITED KINGDOM, 2020).

Não se pode ignorar, assim, a necessidade de: (1) na educação, estimular jovens e adultos, por meio de alfabetização midiática, a adquirirem consciência, visão e ouvido crítico perante às deepfakes e a todo tipo de fake news que assolam o ciberespaço; (2) nas agências de checagem, desmascarar a desinformação, com trabalho árduo que parece não ter fim, porque, enquanto se desmascara uma FN, outras chegam no lugar; (3) na academia e nas empresas, pesquisar e realizar experimentos que procurem outras formas de armazenar conteúdo, a fim de que não seja adulterado, como, por exemplo, em plataformas *blockchain*; (4) dedicar esforços a fazer com que a sociedade possa entender o material que lhes chega e obter proximidade para participar da elaboração de leis de regulamentação, para que não esbarremos em censura e em ameaças à liberdade de expressão. Portanto, registrar o perigo das deepfakes serve, ao menos, como precaução para que continuemos a pensar maneiras de ações contra elas, de impedi-las e, assim, nos esforçar em aliviar o estrago que fazem na cultura democrática.

Referências

BOTS. *The Media Manipulation Casebook*. Disponível em: mediamanipulation.org/definitions/bots. Acesso em: 12 jul. 2021.

BUCCI, Eugênio. *Existe democracia sem verdade factual?* São Paulo: Estação das Letras e Cores, 2019.

DEEP LEARNING é tecnologia de aprendizado de máquina que mais cresce em todo o mundo. 2 out. 2017. Disponível em: unicamp.br/unicamp/noticias/2017/10/02/deep-learning-e-tecnologia-de-aprendizado-de-maquina-que-mais-cresce-em-todo-o. Acesso em: 17 jul. 2021.

FRANCESCHI-BICCHIERAI, Lorenzo. 2020. Listen to This Deepfake Audio Impersonating a CEO in Brazen Fraud Attempt. 23 jul. 2020. *Motherboard/Vice*, 23 jul. 2020. Disponível em: vice.com/en/article/pkyqvb/deepfake-audio-impersonating-ceo-fraud-attempt. Acesso em 17 jul. 2021.

GOGONI, Ronaldo. O que é deep fake e porque você deveria se preocupar com isso. *Tecnoblog*, 18 out. 2018. Disponível em: tecnoblog.net/264153/o-que-e-deep-fake-e-porque-voce-deveria-se-preocupar-com-isso. Acesso em: 13 jul. 2021.

HAI – HUMAN-CENTERED ARTIFICIAL INTELLIGENCE, STANFORD UNIVERSITY. *Artificial Intelligence Index Report 2021*. Palo Alto, CA, 2021. Disponível em: aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf. Acesso em: 15 jul. 2021.

HAMELEERS, Michael *et al*. A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated. *Social Media, Political Communication*, v. 37, n. 2, p. 281-301, 2020.

HAO, Karen. How Facebook got addicted to spreading misinformation. *MIT Technology Review*, 11 mar. 2021. Disponível em: technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation. Acesso em: 12 jul. 2021.

HARWELL, Drew. An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft. *The Washington Post*, 4 set. 2019. Disponível em: washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft. Acesso em: 14 jul. 2021.

JIN, Zeyu *et al*. Voco: text-based insertion and replacement in audio narration. *ACM Transactions on Graphics*, v. 36, n. 4, p. 1-13, jul. 2017.

JONES, M. Tim. *Um guia para iniciantes sobre inteligência artificial, aprendizado de máquina e computação cognitiva*. 1 jun. 2017. Disponível em: ibm.com/developerworks/br/library/guia-iniciantes-ia-maquina-computacao-cognitiva/index.html. Acesso em: 12 jul. 2021.

LEADING Birmingham expert contributes to review on Artificial Intelligence. 13 fev. 2020. Disponível em: birmingham.ac.uk/university/colleges/eps/news/2020/2/leading-birmingham-expert-contributes-to-review-on-artificial-intelligence.aspx. Acesso em: 18 jul. 2021.

LEAL, Luziane de Figueiredo Simão; MORAES FILHO, José Filomeno de. Inteligência artificial e democracia: os algoritmos podem influenciar uma campanha eleitoral? Uma análise do julgamento sobre o impulsionamento de propaganda eleitoral na internet do Tribunal Superior Eleitoral. *Direitos Fundamentais & Justiça*, Belo Horizonte, ano 13, n. 41, p. 343-356, jul./dez. 2019.

LEE, Kai-Fu. *AI Superpowers*: China, Silicon Valley, and the New World Order. Boston, MA: Mariner Books, 2018.

LI, Yuezun; CHANG, Ming-Ching; LYU, Siwei. Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *ArXiv*, vol. abs/1806.02877, 2018. Disponível em: arxiv.org/pdf/1806.02877.pdf. Acesso em: 04 ago. 2021.

MANOVICH, Lev. *Can We Think Without Categories?* Disponível em: manovich.net/content/04-projects/105-can-we-think-without-categories/manovich_can_we_think_without_categories_09_14_2018.pdf. Acesso em: 14 set. 2018.

MASOOD, Momina *et al*. Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. *arXiv.org*, 25 fev. 2021. Disponível em: arxiv.org/abs/2103.00484. Acesso em: 17 jul. 2021.

MESSARIS, P.; ABRAHAM, L. The role of images in framing news stories. In: REESE, Stephen D.; GANDY, Oscar H.; GRANT, August E. Grant (Eds.). *Framing public life*, Mahwah, NJ: Erlbaum, 2001, p. 215–226.

RECONTEXTUALIZED media. *The Media Manipulation Casebook*. Disponível em: mediamanipulation.org/definitions/recontextualized-media. Acesso em: 10 jul. 2021.

PRADO, Magaly. Inteligência artificial e algoritmos de enganação. In: SANTAELLA, Lucia (org.). *Inteligência artificial & redes sociais*. São Paulo: Educ, 2019. p. 57-72.

RAMONET, Ignácio. A opinião pública não quer a verdade, quer confirmar crenças. Entrevista a Cíntia Alves. *GGN*, 25 dez. 2018. Disponível em: jornalggn.com.br/midia/ignacio-ramonet-a-opiniao-publica-nao-quer-a-verdade-quer-informacoes-que-confirmam-suas-crencas/amp. Acesso em: 17 jul. 2021.

SHNURENKO, Igor; MUROVANA, Tatiana; KUSHCHU, Ibrahim. *Artificial Intelligence*: Media and Information Literacy, Human Rights and Freedom of Expression. Moscow, Hove: UNESCO Institute for Information Technologies in Education, TheNextMinds, 2020.

SPENCER, Michael K. Deep Fake, a mais recente ameaça distópica. *Outras Palavras*, 30 maio 2019. Disponível em: outraspalavras.net/tecnologiaemdisputa/deep-fake-a-ultima-distopia. Acesso em: 13 jul. 2021.

TANDOC, JR., Edson C.; WEI LIM, Zheng; LING, Richard. Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, v. 6, n. 2, p. 137-153, 2017.

TRANSPARÊNCIA BRASIL. *Recomendações de governança*: uso de inteligência artificial pelo poder público. São Paulo, 2020. Disponível em: bit.ly/3xfOXYo. Acesso em: 17 jul. 2021.

UNITED KINGDOM. Centre for Data Ethics and Innovation. *Online targeting*: Final report and recommendations. London, 2020. Disponível em: gov.uk/government/publications/cdei-review-of-online-targeting/online-targeting-final-report-and-recommendations. Acesso em: 17 jul. 2021.

VERIFICATION Tools. *Datajournalism.com*. Disponível em: datajournalism.com/read/handbook/verification-1/verification-tools/10-verification-tools. Acesso em: 14 jul. 2021.

WARDLE, Claire. Fake news: It's complicated. *First Draft*, 16 fev. 2017. Disponível em: firstdraftnews.org/articles/fake-news-complicated. Acesso em: 12 jul. 2021.

_____. *Information disorder*: the essential glossary. 2018. Disponível em: firstdraftnews.org/wp-content/uploads/2018/07/infoDisorder_glossary.pdf. Acesso em: 12 jul. 2021.