BOTELHO, Thaïs Helena Falcão; NÖTH, Winfried. Deepfake: Inteligência Artificial para discriminação e geração de conteúdos. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 69-78.

Recebido em: 19 fev. 2021
Aprovado em: 13 mar. 2021

dx.doi.org/10.23925/1984-3585.2021i23p69-78

Deepfake:

Inteligência Artificial para discriminação e geração de conteúdos

Thaïs Helena Falcão Botelho[1]

Winfried Nöth[2]

Resumo: O termo "fake news" começou a povoar as mídias sociais, principalmente a partir de 2016, em função das eleições à presidência dos Estados Unidos. Estudos apontam que as notícias falsas acabam tendo um número maior de compartilhamento do que publicações de sites idôneos, podendo inclusive influenciar no resultado da eleição. É possível observar que a produção de notícias falsas se utiliza de recursos tecnológicos advindos do universo da mídia impressa como: fotos, textos e diagramação. Atualmente, grande parte das sociedades se utiliza, cada vez mais, de mídias digitais. Tais mídias viabilizam, além das linguagens do mundo impresso, outros tipos de linguagens, como conteúdos audiovisuais. Nesse ambiente virtual, novas tecnologias de Inteligência Artificial estão sendo desenvolvidas, como é o caso da deepfake, que também pode ser utilizada para a criação de conteúdo, inclusive para veiculação de notícias falsas. Tais veiculações podem ameaçar a confiança nas instituições e na democracia. Um dos caminhos propostos para combater as deepfakes é um trabalhado educativo, como a alfabetização midiática.

Palavras-chave: Deepfake. Fake news. Inteligência Artificial. GAN. Educação. Alfabetização midiática.

1 Doutoranda e Mestre em Tecnologias da Inteligência e do Design Digital, PUC – SP. Integrante do grupo Sociotramas, PUC–SP. Editora e pesquisadora de imagens para materiais educacionais. CV Lattes: lattes.cnpq.br/0035882440807212. E-mail: olhodofalcao.imagem@gmail.com.

2 Professor do Programa de Pós-Graduação em Tecnologias da Inteligência e Design Digital (TIDD/PUC-SP). ORCID: orcid.org/0000-0002-2518-9773. CV Lattes: lattes.cnpq.br/7221866306191176. E-mail: wnoth@pucsp.br.

Deepfake: IA for discrimination and generation of digital content

Abstract: The term "fake news" has become popular in the social media in 2016 during the elections for the presidency of the United States. Studies have shown that false news may have a greater number of likes than messages on reputable sites. They may even influence the outcome of elections. The production of false news uses technological resources from the world of print media, photos, texts, and layout. Currently, the digital media predominate. In addition to the printed text, they use audiovisual content. In this virtual environment, new artificial intelligence technologies are being developed, as is the case of deepfake, which can also be used for the production of content for disseminating false news. Such placements can threaten trust in institutions and democracy. One of the paths proposed to combat deepfakes is educational work through media literacy.

Keywords: Deepfake. Fake news. Artificial Intelligence. GANs. Education. Media literacy.

Em 2016 o termo fake news começou a circular por diversas mídias, principalmente como consequência do grande volume de conteúdos falsos que estavam sendo propagados pelas redes sociais, devido às eleições americanas à presidência dos Estados Unidos com os candidatos Donald Trump e Hillary Clinton. Uma das fake news desse período foi a publicada pelo site *WTOE 5 News*, afirmava que o Papa Francisco endossava Donald Trump como candidato (Figura 1).

Figura 1. *Fake post* de 17 de nov. de 2016 com a "notícia" que o papa apoia a candidatura de Donald Trump. Fonte: Press (2016).

De acordo com Gunther, Nisbeth e Beck (2018), "cerca de 10% de nossa amostra nacional e 8% dos partidários de Obama pensaram que essa declaração fosse verdadeira" (THE CONVERSATION, 2018).

Durante o período das eleições americanas, conforme levantamento feito pelo site *BuzzFeed.News*, enquanto 20 notícias falsas levaram a 8, 711 milhões de reações no *Facebook*, as 20 melhores matérias, produzidas por mídias tais como *New York Times*, *Washington Post* e a *NBC News*, renderam 7,367 milhões de reações no *Facebook* (SILVERMAN, 2016).

Mídia impressa – composição da fake news

Para a produção desse tipo de fake news são utilizados softwares para processos de diagramação, no caso não muito complexos, que possibilitaram a justaposição de duas fotos. São utilizados softwares que integram linguagens, como a escrita, verbal, design gráfico, fotografias, para criar um documento falso. Essa montagem, além de utilizar conhecimentos básicos de diagramação, lançou mão de recursos visuais como fotografias, ícones e estruturas, similares de sites de notícias idôneas, tais como abas e palavras como *home, US election*, dentre outros recursos.

Atualmente, há uma série de imagens, áudios, vídeo e textos sob a licença *Creative Commons* (CC). Conforme a licença CC aplicada num conteúdo, ele pode ser utilizado por qualquer pessoa, sem que ela tenha de solicitar a autorização do seu produtor e nem se preocupar com a remuneração de direitos autorais.

Para uma publicação dessas ter um efeito assim tão explosivo nas redes sociais, além da absoluta falta de ética, alguns dados são importantes. O Papa Francisco e o Donald Trump, em 2016, estavam na lista da *Time* das 100 pessoas mais influentes do planeta (TIME 100, 2016). É normal que personalidades desse calibre tenham seu dia a dia vastamente documentado e praticamente impossível que informações desse tipo se mantenham em sigilo. Além disso, é muito rara a possibilidade de uma mídia tão pouco conhecida conseguir apurar uma informação desse calibre antes das mídias mais tradicionais e profissionalizadas.

No entanto, é importante notar que o leitor que acreditou nessa mensagem nem parou para refletir que um evento dessa monta contaria com a presença massiva de mídias de todo o mundo. No mínimo, um simples aperto de mão dispararia centenas de flashes e uma vasta produção de imagens inundariam celulares, canais de TV, rádios e publicações impressas. Se por acaso, um encontro entre tais personalidades públicas não pudesse se efetuar presencialmente, com certeza as grandes mídias não deixariam um evento político dessa monta passar por uma varredura de diversos tipos de documentos que demonstrassem a legitimidade de tal endosso. Porém, bastou juntar as imagens, em um momento extremamente sensível das eleições americanas, dessas duas personalidades, global e diariamente documentadas, amarradas por um texto curto e falso, para tornar-se como uma onda que parece ter avançado com intuito de arrastar os eleitores para Trump.

As imagens têm consistências, são retratos de pessoas públicas. Foram capturadas em uma situação real por fotógrafos, em locais e momentos distintos. Ao serem colocadas juntas em uma mesma página, essas duas personalidades foram amarradas com um texto que condizia com uma realidade que nunca ocorreu. Observa-se que, pelo menos esses 8% dos eleitores de Obama, que acreditaram nessa publicação, possivelmente, não pararam para refletir alguns segundos, ou mesmo minutos, provavelmente convencidos pelo fato de que a fotografia é vista como um comprovante da afirmação do que o texto faz.

Há décadas produtores midiáticos fazem uso desse tipo de integração de linguagens para que possam publicar suas matérias. As tecnologias utilizadas para isso, existem há mais de cem anos, passou por processos extremamente manuais até chegar a uma construção plenamente digital, pela qual, atualmente, permite que grande parte da população possa integrar imagem com texto através de uma plataforma de rede social. No entanto, a partir da digitalização das linguagens, outras mídias e outras linguagens vêm concorrendo e, em termos numéricos, suplantando a lógica da mídia impressa. Em suma, a troca de informação, de bens simbólicos, vem passando por mudanças devido às mídias digitais disponíveis.

Constata-se que as mídias que se utilizam de signos sonoros, como *podcast*, rádio, ou as que integram som com imagem em movimento, como o caso do vídeo, já vinham ganhando um campo massivo nas trocas simbólicas já no século passado. Atualmente, acentuou-se ainda mais o consumo de conteúdo audiovisual devido a pandemia do COVID-19 e passam, a ser praticamente, a forma prioritária de apreender algo da realidade.

De acordo com um o estudo feito pela *Global Web Index*, de 2020, publicado pela *Visual Captalist*, com o objetivo de observar como a COVID-19 tem impactado o consumo de mídia, por geração, ele concluiu que

> mais de 80% dos consumidores nos EUA e no Reino Unido afirmam consumir mais mídias desde o surto, tais como conteúdos transmitidos pela TV e vídeos online (*YouTube, TikTok*), sendo esses os principais meios em todas as gerações e gêneros." (JONES, 2020, tradução nossa)

Nesse estudo verificou-se que as mídias mais consumidas são os vídeos, tanto para uma geração mais velha pesquisada, de 57 a 64 anos, como para a geração mais nova, de 16 a 23 anos. A maior diferença é que para a geração mais nova, a busca por vídeos se dá de forma online, enquanto, para a geração mais velha, se dá por transmissão, como são os casos das TVs abertas e a cabo.

O Brasil também segue essa tendência, segundo o levantamento *TIC domicilio 2019*, em termos de atividades culturais, "assistir a vídeos e ouvir música são as atividades culturais mais comuns entre usuários de Internet" (COMITÊ GESTOR, 2020, p. 24). Além de que, uma das atividades mais comuns dos usuários da internet é a comunicação pessoal e apontam para um "crescimento de chamadas por voz ou vídeo (73%)" (ibid. p. 15).

Mídias digitais e Inteligência Artificial – deepfake

A produção e consumo de linguagens, como o caso da audiovisual, ocorrem de forma integrada com o desenvolvimento de suas tecnologias. Com a digitalização, tornou-se possível a rastreabilidade das linguagens. O *Youtube* é um caso de integração de produção, rastreabilidade e consumo. "O *YouTube* é o segundo maior mecanismo de busca do mundo e o segundo site com mais tráfego, atrás apenas do *Google*" (KINAST, 2019). Sem uma tecnologia de rastreabilidade, não seria possível o *Youtube* atender aos seus usuários, que "assistem a mais de 180 milhões de horas de conteúdo nas telas de *smart* TVs todos os dias" (ibid.). Além disso, esse ambiente digital vem cada vez mais sofisticando os algoritmos de Inteligência Artificial (IA) na sua plataforma. Em março de 2020, durante a pandemia, anunciou que

> a companhia afirma que sua varredura dependerá mais do aprendizado de máquina e menos de revisores humanos. Normalmente, os algoritmos detectam a postagem potencialmente perigosa e a envia para avaliação humana. Como a força de trabalho da empresa também sofre redução devido ao isolamento dos colaboradores, seu sistema automatizado será utilizado de forma ampliada. (YUGE, 2020)

A IA vem cada vez mais se integrando nessa estrutura de rastreamento das linguagens, fazendo com que nela aumente sua própria aprendizagem. Tal aumento da capacidade computacional da IA foi "permitindo que até hoje os sucessos mais marcantes na aprendizagem profunda tenham envolvido modelos discriminativos" (GOODFELLOW *et al.*, 2014, p. 11, tradução nossa).

Em suma, a IA vinha trilhando seu caminho nos ambientes digitais para ações discriminativas, isto é, as que vinham, por exemplo, com intenção de categorizar conteúdos, tais como imagens e vídeos. Porém, em 2014, Ian Goodfellow, junto com outros pesquisadores, apresentaram as *Redes Adversárias Generativas* (GANs) que "são arquiteturas de redes neurais profundas compostas por duas redes colocadas uma contra a outra (daí o nome 'adversárias')" (DATA SCIENCE ACADEMY, 2021).

Isso quer dizer que há dois tipos de redes neurais que ao serem programadas para serem adversárias acabam criando um ambiente de aprendizagem profundo. A discriminadora analisa grandes conjuntos de dados e sua ação é etiquetá-los, marcá-los, se são falsos ou verdadeiros. Essa rede, então, age em cima dos dados para "reconhecer se são autênticos" (ibid.). A rede neural geradora trabalha para criar imagens sintéticas com o objetivo de receber a etiqueta de autêntica da discriminadora. Como a geradora recebe o *feedback* da discriminadora, ela aprimora a produção de dados, como conteúdos visuais, até conseguir a etiqueta de autenticidade. Por exemplo, para que uma imagem falsa possa chegar a ser considerada verdadeira, as GANs funcionam basicamente da seguinte forma:

> O gerador está criando novas imagens sintéticas que são transmitidas ao discriminador. O gerador gera as imagens *fake* na esperança de que elas também sejam consideradas autênticas, mesmo sendo falsas. O objetivo do gerador é gerar dígitos manuscritos cada vez melhores. O objetivo do discriminador é identificar imagens falsas do gerador. Ou seja, são duas redes adversárias, uma discriminativa (padrão que já estudamos até aqui no livro) e uma generativa que, em termos gerais, faz o oposto das redes. (Ibid.)

Essa tecnologia de IA aplicada para alterar conteúdos originais de vídeos e áudios, com propósito de que pareçam autênticas é nomeada de deepfake. "No decorrer do treinamento, o gerador aprende as mais sofisticadas técnicas sintéticas e o discriminador se transforma em um avaliador dos mais precisos" (PARK; HUH; KIM, 2020, p. 2, tradução nossa)

Como já visto, uma simples matéria, que se utilizou da junção de duas imagens disponíveis na rede feita por uma mídia irrelevante, levou a 8% dos eleitores do ex-presidente Barack Obama a acreditarem que o Papa Francisco realmente iria apoiar o candidato Donald Trump nas eleições norte-americanas. Nessa conjuntura, o lastro com a realidade foi criado através da documentação fotográfica das duas personalidades. Realmente Jeffrey Bruno e Gage Skidmore, fotojornalistas profissionais, fotografaram, respectivamente, as duas personalidades, mas em locais e datas distintas, provavelmente para a cobertura de alguma matéria que não tinha qualquer correlação com um possível apoio político às eleições de 2016. Tais imagens continuam tendo lastro, são registros de fatos que realmente ocorreram. No caso das deepfakes, a imagem ou o som da voz, que seria utilizada como um documento comprovatório de algum evento, acaba se tornando um simulacro da realidade, pois o registro visual e sonoro não precisa estar mais estaratrelado a um fato ocorrido. O trabalho para se produzir uma deepfake se dá na manipulação de dados digitais,

ela não precisa mais dos fatos em si para produzir notícias, bem como filmes. Os conteúdos são sintetizados através da nanotecnologia e algoritmos. Então, diante de uma sociedade que toma conhecimento da realidade através de vídeos e áudios, ignorar totalmente tais tecnologias pode, inclusive, ser mortal, como por exemplo, no caso da manipulação de um vídeo médico na indicação de um tratamento para doenças, que possam vir a ser fatais, tais como o câncer ou hipertensão.

Apesar disso, de acordo com uma revisão sobre o surgimento da tecnologia deepfake, feita por Mika Westerlund, essa tecnologia pode ter uma série de usos positivos, como a que permite "automação de dublagem realista de voz para filmes em qualquer idioma" (WESTERLUND, 2019, p. 41, tradução nossa), "detectar anormalidades em raios-X" (ibid.), tem "potencial para criar moléculas químicas virtuais para acelerar ciência dos materiais e descobertas médicas" (ibid.), dentre uma série de outras possibilidades. No entanto esse mesmo pesquisador alerta que as

> deepfakes são uma grande ameaça para sociedade, para os sistemas políticos e para as empresas, porque eles pressionam os jornalistas que lutam para filtrar o real a partir de notícias falsas, elas podem ameaçar a segurança nacional por disseminar uma propaganda que interfere nas eleições, dificultam a confiança dos cidadãos nas informações das autoridades e levantam questões de cibersegurança junto às pessoas e às organizações. (Ibid., p. 47)

Nesse mesmo estudo, Westerlund apresenta alguns caminhos para o combate a desinformação das deepfakes, como: "1) legislação e regulamentação, 2) políticas corporativas e ação voluntária, 3) educação e treinamento, 4) tecnologia anti-deepfake" (ibid.). Tal tecnologia, nos dias de hoje, é acessível e ainda não encontra barreiras plenamente efetivas para coibir a propagação de conteúdos falsos pela web. Mesmo que se tenha uma legislação, políticas e tecnologias anti-deepfake, a propagação de notícias falsas, por pessoas com interesses escusos, pode continuar a encontrar soluções convincentes, através do desenvolvimento da IA que cada vez mais tem o poder de: rastrear, entrelaçar e produzir conteúdo. Até mesmo porque, nesse caso das GANs, quanto mais se desenvolver a IA para ser discriminativa, mais ela aprenderá ser gerativa. Em suma, a essas redes neurais

> refletem como o cérebro humano funciona. Quanto mais o cérebro humano é exposto a exemplos de algo, como arremessar uma bola de basquete ou a letra de alguma música nova, mais rápido e mais preciso o cérebro pode reproduzi-lo. As redes neurais usam esse mesmo conceito; quanto mais exemplos são inseridos na rede, mais precisamente ela pode criar um novo exemplo do zero. (DACK, 2019, tradução nossa)

Então, pode-se compreender que o exemplo do cérebro humano foi aplicado na aprendizagem de máquinas, para que sejam mais "inteligentes". Tal exemplo pode também voltar-se para aprendizagem do próprio ser humano diante das mídias, possivelmente por meio de uma educação que expanda o repertório para a reflexão de seus cidadãos. Uma formação que aumente sua capacidade discriminatória e, ao mesmo tempo, o instrumente para geração de conteúdo, como por meio da alfabetização midiática e informacional. Uma educação que

> vise melhorar a alfabetização em mídia digital, aprimorar o comportamento online e o pensamento crítico, para possibilitar processos cognitivos e proteções concretas mais eficientes em direção a consumo e uso indevido de conteúdos digitais. (WESTERLUND, 2019, p. 47, tradução nossa)

Referências

COMITÊ GESTOR DA INTERNET NO BRASIL. TIC domicílios 2019: principais resultados. São Paulo: CGI, 2020. Disponível em: cetic.br/media/analises/tic_domicilios_2019_coletiva_imprensa.pdf. Acesso em: 10 abr. 2021.

DACK, Sean. Deep fakes, fake news, and what comes next. *The Henry M. Jackson School of International Studies*, University of Washington. Washington, 20 mar. 2019. Disponível em: jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next/. Acesso em: dez. 2020.

DATA SCIENCE ACADEMY. *Deep learning book*. São Paulo: Data Science Academy, 2021. Disponível em: deeplearningbook.com.br. Acesso em: abr. 2021.

GOODFELLOW, Ian J. *et al*. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. Disponível em: arxiv.org/abs/1406.2661. Acesso em: abr. 2021

GUNTHER, Richard; NISBET, Erik C.; BECK, Paul. Trump may owe his 2016 victory to "fake news", suggests a new study. *The Conversation*, 15/02/2018. Disponível em: theconversation.com/trump-may-owe-his-2016-victory-to-fake-news-new-study-suggests-91538. Acesso em: 1 abr. 2021.

JONES, Katie. How COVID-19 has impacted media consumption, by generation. *Visual Capitalist*, Vancouver, 7 abr. 2020. Disponível em: visualcapitalist.com/media-consumption-covid-19. Acesso em: abr. 2021.

KINAST, Priscilla. Os incríveis números do *Youtube* em 2019: quantos vídeos tem no *Youtube*? Qual vídeo mais assistido? Quantas pessoas usam o *Youtube*? Quais são os maiores canais? Quantas horas são assistidas a cada minuto no *Youtube*? Essas e outras respostas aqui. *Oficina da Net*, 7 ago. 2019. Disponível em: oficinadanet.com.br/tecnologia/26607-os-incriveis-numeros-do-youtube-em-2019. Acesso em: 30 mar. 2021

PARK, Sung-Wook; HUH, Jun-Ho; KIM, Jong-Chan. BEGAN v3: avoiding mode collapse in GANs using variational inference. *Electronics*, 9(4): 688, 2020. Disponível em: doi.org/10.3390/electronics9040688. Acesso em: 30 mar. 2021.

PRESS, Larry. A real-names domain registration policy would discourage political lying. *CircleID*, 17, nov. 2016. Disponível em: circleid.com/posts/20161117_real_names_domain_registration_policy_discourage_political_lying. Acesso em: 30 mar. 2021.

SILVERMAN, Craig. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed.News*. 16 nov. 2016. Disponível em: buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook#.uc9gevywE. Acesso em: 1 abr. 2021.

TIME 100: The 100 most influential people. *Time*, 21 abril, 2016. Disponível em: time.com/collection/2016-time-100/leaders/. Acesso em: 1 abr. 2021.

WESTERLUND, Mika. The emergence of deepfake technology: a review. *Technology Innovation Management Review*, v. 9, n. 11, 2019. Disponível em: timreview.ca/article/1282. Acesso em: 30 mar. 2021.

YUGE, Cláudio. YouTube vai usar mais IA e menos revisão humana em conteúdos sobre coronavírus. *Canaltech*, 16 mar. 2020. Disponível em: canaltech.com.br/internet/youtube-vai-usar-mais-sua-ia-e-pode-remover-mais-conteudo-sobre-coronavirus-161927/. Acesso em: 1 abr. 2021.