TEC COGS

23
JAN./JUN.
2021

REVISTA DIGITAL DE TECNOLOGIAS COGNITIVAS

ISSN 1984-3585

Deepfake

Programa de Pós-Graduação em
Tecnologias da Inteligência e Design Digital

Pontifícia Universidade Católica de São Paulo

Expediente

TECCOGS – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, ISSN: 1984-3585
Programa de Pós-graduação em Tecnologias da Inteligência e Design Digital (TIDD),
Pontifícia Universidade Católica de São Paulo (PUC-SP).

Esta edição foi viabilizada por verba do Plano de Incentivo à Pesquisa (PIPEq) da PUC-SP.

n. 23, jan./jun. 2021

Sumário

RESENHAS

SANTAELLA, Lucia. Editorial. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 6-8.

dx.doi.org/
10.23925/1984-3585.2021i23p6-8

Editorial

Por Lucia Santaella[1]

Fake news, desinformação e pós-verdade tornaram-se palavras de ordem pelas preocupações que causam e pelos sistemas de alarme que fazem soar em função do fato de que não são meras palavras, mas designações que se reportam aos efeitos nocivos que provocam no comportamento humano e na sociedade como um todo. É enorme a profusão de textos que vêm sendo produzidos a respeito dos riscos desse fenômeno, uma profusão que tende a crescer, na medida mesma em que ainda não foram encontrados antídotos eficazes para se deter a febre da enganação e da mentira cujas consequências interferem em processos decisórios.

Evidentemente, enganar e mentir não são ações novas. Em suas versões nefastas, elas fazem parte do arsenal humano para fazer o mal. Entretanto, essa tendência, antes mais rarefeita, encontra-se agora acionada pelos meios facilitadores que a internet colocou à mão, transformando a disseminação de boatos, falsidades e mentiras em uma verdadeira indústria a que não têm faltado adeptos. "O que fazer?" é uma pergunta que tem perseguido todos aqueles que eticamente prezam o bem comum.

É nesse contexto que o número 23 da TECCOGS busca interferir ao trazer à discussão o mais recente desdobramento das fake news nas formas videográficas das deepfakes que, justamente devido à sua natureza visual, auxiliada por tecnologias sofisticadas, muito provavelmente prometem trazer mais munição à indústria em curso e causar ainda mais danos do que as fake news.

---

1 Lucia Santaella é pesquisadora 1A do CNPq, professora titular da PUC-SP. Publicou 51 livros e organizou 24, além da publicação de mais de 400 artigos no Brasil e no exterior. Recebeu os prêmios Jabuti (2002, 2009, 2011 e 2014), o prêmio Sergio Motta (2005) e o prêmio Luiz Beltrão (2010). ORCID: orcid.org/0000-0002-0681-6073. CV Lattes: lattes.cnpq.br/7427854657719431. E-mail: lbraga@pucsp.br.

A entrevista para o número foi concedida por Demi Getschko, professor da PUC-SP e Diretor presidente do Núcleo de Informação e Coordenação do Ponto BR (NIC.br). Levando em conta a grande especialidade do entrevistado nas questões da segurança na internet, as perguntas buscaram suas abalizadas posições acerca dos pontos cruciais em que as legislações, no que concerne especialmente ao funcionamento das redes sociais, não firam a liberdade de expressão. De fato, a dificuldade da questão relativa ao ponto de equilíbrio exato entre a lei e a liberdade encontra-se intensificada nas redes. As posições de Demi Getschko sobre isso são sábias, demonstrando que o conhecimento alimentado pela atuação concreta nesse campo é a via régia para se compreender quais são os caminhos que os dilemas devem seguir para que posições lúcidas preponderem.

O número funciona, portanto, como um sinal de alerta, tocado em várias modulações. No artigo da seção "Dossiê", delineio e apresento a repercussão do tema das deepfakes em publicações selecionadas, de modo a contextualizar a leitura dos artigos deste número.

Começamos a seção "Artigos" com o texto de Souza e Santaella que visa evidenciar o caráter semiótico tanto das fake news quanto das deepfakes. Embora seus efeitos caminhem para um mesmo alvo, ou seja, enganar, as deepfakes têm mais poder de provocar credulidade, devido à restrição perceptiva que o ser humano tem de duvidar daquilo que vê.

Assinado por Prado, o artigo sobre deepfake de áudio explora os recursos de técnicas de Inteligência Artificial (IA) e o elenco de ferramentas para fabricar esse tipo de deepfake, de modo a levantar os efeitos produzidos por sua disseminação descontrolada como uma afronta à ética da informação. Botelho e Nöth também chamam atenção para as técnicas de IA que protagonizam a produção de deepfake, apontando para a necessidade de formação educacional como um tipo de freio capaz de estancar a enxurrada dos enganos.

As modulações de Fabio de Paula e Hessel soam em uma outra direção, a saber, a dos aspectos criativos que também existem quando o potencial das deepfakes são desviados para a sua exploração no mundo da cultura e do entretenimento, sobretudo quando a própria criação em diferentes formatos faz um uso planejado dessa tecnologia para fins de viés afirmativo como a criação de novas linguagens e obras.

Tanto Salgado e Santaella quanto Fanaya conduzem a discussão das deepfakes para um contexto mais amplo. No primeiro caso, o contexto é aquele da crise de confiança na sociedade a qual é nitidamente impulsio-

nada pelas deepfakes assim como também o é paradoxalmente, em parte, por avanços como o *blockchain*. Para Fanaya, a guerra da desinformação será intensificada pelas deepfakes. Elas tornaram essa guerra mais complexa e perigosa, na medida em que aquilo que ainda restava de réstea de dúvida nas fake news, tenderá a se dissipar, colapsando qualquer distinção entre verdadeiro e falso ou real e fictício. Contextualizando as deepfakes no problema da realidade sintetizada/simulada, o artigo as relaciona com o problema da verdade. Por fim, o artigo de Coelho e Hildebrand busca compreender como as estratégias discursivas são produzidas nas deepfakes, realizando, para isso, uma análise do percurso gerativo de sentido tripartido nos níveis narrativo, discursivo e fundamental.

A seção "Extra dossiê" apresenta "How can we change habits?", o terceiro da série de quatro diálogos sobre semiótica cognitiva entre Vincent Colapietro e Winfried Nöth. No texto, eles comentam a natureza lógica da geração e mudança de hábitos, explorando algumas de suas múltiplas causas e efeitos nos contextos sociais e midiáticos contemporâneos.

Na seção "Resenhas", Kaufman aborda "Ethics of Artificial Intelligence", coletânea de artigos inéditos organizada por Metthew Liao em 2020. Os textos, em sua multiplicidade, versam sobre a construção da ética na e por meio da IA no presente e em futuros próximos e distantes.

Ao fim e ao cabo, não obstante a crise de confiança e as distorções que a disseminação da mentira tem provocado na sociedade, continuamos a acreditar que um dos antídotos possíveis contra esse tipo de insanidade encontra-se no respeito pela informação responsável e no dever ético de difundi-la em prol do conhecimento que, cedo ou tarde, é aquele que ocupará a posição de vencedor. Este número da revista TECCOGS está alimentado por esses princípios.

ENTREVISTA

GETSCHKO, Demi. Entrevista concedida a Lucia Santaella. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 10-13.

Recebido em: 1 abr. 2021
Aprovado em: 11 abr. 2021

dx.doi.org/10.23925/1984-3585.2021i23p10-13

Entrevista com Demi Getschko

Por Lucia Santaella[1]

Antes de tudo, aqui vêm os agradecimentos da Revista TECCOGS por você ter gentilmente aceitado responder a esta entrevista, aliás, esta é a revista do PEPG de Tecnologias da Inteligência e Design Digital, que tem a honra de contar com sua presença no corpo docente. Demi Getschko é também diretor presidente do Núcleo de Informação e Coordenação do Ponto BR (NIC.br).

Lucia Santaella (L.S.): *Tenho seguido suas colunas no Estado de S. Paulo e tenho apreço por elas principalmente em função da polilateralidade com que você trata os desafios atuais da internet, em especial no que concerne aos dois temas que pretendo tratar nesta entrevista: a segurança e a privacidade, pois, tanto quanto posso ver, são esses temas que tocam mais de perto as consequências que o dilúvio de fake news e sua extensão nas deepfakes podem trazer à sociedade. Você concorda com esses dois temas ou encontra algum outro que seja de igual ou maior importância?*

Demi Getschko (D.G.): Meu foco nas colunas é tecnologia e Internet mas, de fato, acabo tendo uns arroubos em áreas relacionadas. Hoje em dia é difícil encontrar algum tema que esteja "descolado" da Internet. Assim, sempre há como traçar paralelos entre o que encontramos hoje em dia e a revolução que a rede trouxe. Quero me concentrar também em Internet da Coisas e Inteligência Artificial por tudo o que elas trazem, tanto em termos de progresso e conforto, como em riscos de segurança e alterações do comportamento social. E, claro, ligando-as naturalmente à Internet.

---

1 Lucia Santaella é pesquisadora 1A do CNPq, professora titular da PUC-SP. Publicou 51 livros e organizou 24, além da publicação de mais de 400 artigos no Brasil e no exterior. Recebeu os prêmios Jabuti (2002, 2009, 2011 e 2014), o prêmio Sergio Motta (2005) e o prêmio Luiz Beltrão (2010). ORCID: orcid.org/0000-0002-0681-6073. CV Lattes: lattes.cnpq.br/7427854657719431. E-mail: lbraga@pucsp.br.

L.S.: *Existe um certo consenso de que não pode haver outro meio de estancar a proliferação de fake news sem que haja leis que regulamentem o funcionamento das plataformas de redes sociais com rigor. Entretanto, isso não impede o risco de soluções pouco democráticas de projetos de lei que podem impor algum tipo de censura. Como você vê essa questão?*

D.G.: Concordo e tenho muito medo de legislação açodada, que cause mais danos que benefícios... Como "linha de apoio" para essa posição, sirvo-me do sétimo mandamento do decálogo do Comitê Gestor da Internet (CGI), que preconiza que os "responsáveis finais são os que devem ser encontrados". Penso ser perigoso confiar esse poder de "controle" às plataformas, que poder já o tem, e em excesso... Notícias falsas sempre houve, mas a Internet potencializa tremendamente o fenômeno, ao oferecer facilidades aos geradores das falsidades e ao proporcionar a eles um vasto conjunto de "vítimas" facilmente encontráveis. O único caminho realmente promissor seria educar os receptores para que sempre lessem tudo o que recebem "com um grão de sal", e com mentalidade crítica. Não é porque, de alguma forma automática e misteriosa, vêm ao nosso encontro matérias que defendem pontos de vista de que compartilhamos, que eles deveriam ser consideradas, a priori, verídicas...

L.S.: *Censura é sempre uma ameaça a ser expelida, especialmente, dizem os especialistas, em um país como o Brasil no qual o pensamento nacional sobre a liberdade de expressão encontra-se ainda pouco consolidado. Uma vez que você tem uma visão internacional sobre essa questão, como percebe a situação brasileira em um panorama geopolítico?*

D.G.: O Brasil tem uma imagem muito boa em temas técnicos da rede e em legislação. O modelo e o trabalho que o CGI faz, por exemplo, é citado amiúde como exemplo a ser seguido. O conjunto Marco Civil mais Lei Geral de Proteção de Dados (LGPD) mostra a maturidade que a Internet brasileira tem. Acho que "avanços" nessa área devem ser vistos com muita "prudência". E exemplos de avanços atabalhoados que podem tolher a liberdade de expressão, mesmo entre nações muito avançadas e liberais, existem e em quantidade...

L.S.: *Mesmo que haja regulamentação adequada contra a proliferação aparentemente incontrolável dos discursos de ódio e das fake news nas redes sociais, como enfrentar soluções adaptadas à instantaneidade da era digital, sem cair no descompasso de meras transposições de técnicas de aplicação do Direito que eram próprias da era analógica?*

D.G.: Como disse, penso que nada deve ser combatido aprioristicamente. Medidas "a priori" normalmente caem no espectro da "censura"... Penso que os internautas devem gozar de total liberdade em escrever o que quiserem e, até devido a essa liberdade, serem responsabilizados pelos danos ou impropriedades que cometerem. A máxima liberdade vem junto com a máxima responsabilidade. Não devemos cair na tentação de, por uma pretensa classificação entre "verdadeiro" ou "falso", suprimir a liberdade de expressão... Aliás, como Nietzsche teria dito, "as convicções, mais que as mentiras, são inimigas poderosas da verdade".

L.S.: *A história humana está farta de exemplos de que nunca os fins podem justificar os meios. Sua coluna a respeito disso é exemplar, especialmente neste momento em que assistimos a uma verdadeira corrida pela regulamentação do uso das redes, inclusive impondo fronteiras apressadas ao desenvolvimento da Inteligência Artificial, muitas vezes sem conhecimento de causa. Diante disso, quais os caminhos para evitar os efeitos colaterais dessa corrida?*

D.G.: A angústia em resolver algo que nos aflige não justifica, por exemplo, inverter o axioma de que todos devem ser considerados inocentes até prova em contrário. Partir-se da hipótese de que "é melhor espionar o que todos fazem para prevenir malfeitos e riscos" é promulgar o vigilantismo, e isso deveria ser combatido. A LGPD pode ajudar nesse campo, espero...

L.S.: *Sua ideia de que a privacidade é contextual funciona como um achado diante do costumeiro tratamento da privacidade como um conceito redondo, fechado e, consequentemente, mal compreendido. As palavras circulam de boca em boca sem que as pessoas se deem ao trabalho de buscar fontes confiáveis para entender seus verdadeiros efeitos. Creio que a palavra privacidade está sendo vítima disso e precisa ser colocada em seu devido lugar. Gostaria de saber o que você pensa sobre isso.*

D.G.: O que penso é que estamos, de novo, num fenômeno pendular. Se víamos nossos dados, sentimentos e emoções serem devassados sob diferentes pretextos, com a LGPD o pêndulo pode ir ao outro extremo e virmos a perder transparências, que também são fundamentais. A arte aqui consiste em equilibrar nosso direito à privacidade com os direitos da comunidade em identificar seus interlocutores. A internet fornece exemplos que poderiam jogar luz nisso: quando alguém registra um domínio, cria potencialmente um local na rede de onde informação poderá ser disseminada. Ora, seria ruim que um "perfil falso" assumisse a identidade de alguém para, por exemplo, atacar terceiros. Por isso na rede, a par que sempre aceita o que o registrante declara, exibido a informação para que, em caso de alguém se sentir prejudicado, não possa o malfeito ser atribuído a outra vítima, essa que teria sido personificada. Assim a própria comunidade, de alguma forma, colabora para a correta responsabilização eventual. Em resumo, há que se equilibrar privacidade com responsabilidade e transparência.

L.S.: *É notório o avanço que o Marco Civil da Internet provocou. Como você vê esse marco no contexto dos desafios atuais das fake news e das deepfakes que estão vindo por aí?*

D.G.: Creio que o Marco Civil ainda é (e assim continuará a ser por um bom tempo) o melhor apoio que temos para lutar por uma rede sã. É a forma correta garantir os direitos de todos na rede, ao mesmo tempo em que se apontam caminhos que levem à responsabilização dos verdadeiros agentes que causam os problemas... É uma lei que levou mais de seis anos em debate público, até a sua homologação, e não deveria ser afetada por emendas *ad hoc* ou precipitadas...

DOSSIÊ

SANTAELLA, Lucia. As irmãs siamesas fake news e pós-verdade expandidas nas deepfakes. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 15-24.

Recebido em: 11 abr. 2021
Aprovado em: 31 abr. 2021

dx.doi.org/10.23925/1984-3585.2021i23p15-24

As irmãs siamesas fake news e pós-verdade expandidas nas deepfakes

Lucia Santaella[1]

Resumo: A partir de 2016, a expressão fake news e sua parceira, pós-verdade, passaram crescentemente a tomar conta das mídias noticiosas e interpretativas, com muitas matérias publicadas sobre o tema, ganhando, inclusive, as discussões mais detalhadas e bem-informadas da pesquisa acadêmica. Nesse contexto este dossiê pretende apresentar esse tema nas repercussões que tem obtido em algumas publicações selecionadas com o objetivo de preparar o leitor com informações preliminares à leitura dos artigos sobre deepfake que se apresentam neste número da TECCOGS. Infelizmente ainda não contamos no Brasil com livros sobre deepfake, o mais jovem rebento das fake news. Vem daí a relevância deste número que toma a dianteira na busca de uma primeira sistematização sobre essa questão que promete trazer consequências ainda mais nefastas para o equilíbrio social do que as fake news.

Palavras-chave: Fake news. Pós-verdade. Deepfake.

1 Lucia Santaella é pesquisadora 1A do CNPq, professora titular da PUC-SP. Publicou 51 livros e organizou 24, além da publicação de mais de 400 artigos no Brasil e no exterior. Recebeu os prêmios Jabuti (2002, 2009, 2011 e 2014), o prêmio Sergio Motta (2005) e o prêmio Luiz Beltrão (2010). ORCID: orcid.org/0000-0002-0681-6073. CV Lattes: lattes.cnpq.br/7427854657719431. E-mail: lbraga@pucsp.br.

The siamese sisters fake news and post-truth expanded in deepfakes

Abstract: From 2016 on, the expression fake news and its partner, post-truth, have increasingly taken over the news and interpretive media, with many articles published on the subject, even gaining the more detailed and well-informed discussions in academic research. This dossier intends to present this theme in the repercussions it has obtained in some selected publications with the objective of preparing the reader with preliminary information for the articles about deepfake that are presented in this issue. Unfortunately, we still don't have books in Brazil about deepfake, the youngest offspring of fake news. Hence the relevance of this issue, which takes the lead in the search for a first systematization of this subject, which promises to bring even more disastrous consequences for social balance than fake news.

Keywords: Fake news. Post-truth. Deepfake.

As deepfakes são uma extensão, em peças de áudio e vídeo, das fake news, estas quase sempre verbais. Ambas estão associadas às multifacetadas questões relativas à pós-verdade. Esta expressão ganhou notoriedade a partir de 2016, quando se deram as surpreendentes vitórias de Trump nas eleições dos Estados Unidos e do plebiscito Brexit, no Reino Unido. Tanto uma quanto a outra foram consideradas batalhas vencedoras devido à proliferação de notícias falsas que correram pelas redes digitais e que impulsionaram o voto de eleitores mal-informados e, consequentemente, crédulos em relação às enxurradas de mensagens politicamente distorcidas que receberam.

Logo depois, em 2018, deu-se o escândalo do *Cambridge Analytica*, uma empresa inglesa de análise de dados. Um de seus pesquisadores havia desenvolvido um aplicativo de extração de dados pessoais e obtido autorização do Facebook para sua aplicação. Então, pesquisas revelaram que 50 milhões de dados de usuários foram vendidos pelo pesquisador à *Cambridge Analytica* que tinha Trump como cliente. A influência que isso teve nas eleições presidenciais foi desvendada por jornalistas investigadores, redundando, por fim, na falência da empresa e no questionamento do Facebook acerca da permissão de que dados privados de usuários sejam repassados a quaisquer interesses externos. Os fatos acionaram o alerta quanto ao poder político das notícias falsas e, desde 2018, o termo "fake news" passou a tomar conta das mídias noticiosas e interpretativas, com muitas matérias publicadas sobre o tema, ganhando, inclusive, as discussões mais detalhadas e bem-informadas da pesquisa acadêmica.

O objetivo deste dossiê é apresentar o tema nas repercussões que tem obtido em algumas publicações selecionadas, de modo a preparar o leitor com informações preliminares à leitura dos artigos que se apresentam neste número da TECCOGS. Já há um bom número de autores tanto brasileiros quanto estrangeiros, cujas obras obtiveram tradução em português, que se dedicaram ao assunto, quase sempre ligando as fake news à sua irmã siamesa, a pós-verdade. Infelizmente ainda não contamos no Brasil com livros sobre deepfake, o mais jovem rebento de ambas. Vem daí a relevância deste número que toma a dianteira na busca de uma primeira sistematização sobre essa questão que promete trazer consequências ainda mais nefastas para o equilíbrio social do que as fake news.

Para a consecução do objetivo pretendido, segue-se a breve discussão do estado da arte limitado a livros, no caso das fake news, com a finalidade de abrir caminho para os artigos sobre deepfake, presentes neste número, que irão tratar devidamente de algumas das principais preocupações relativas a essa nova forma de enganação. A discussão abaixo buscará tomar uma linha cronológica de modo a evidenciar a evolução das preocupações, especialmente depois que as fake news se tornaram no Brasil objetos de investigação judicial.

Não deixa de ser de interesse que Serva (2001) tenha publicado, *avant la lettre*, um livro sobre jornalismo e desinformação já que a desinformação se tornaria o tema-chave relativo a fake news a ponto de alguns preferirem a palavra desinformação em lugar de fake news. Mas foi a partir de 2016 que a "pós-verdade" começou a atrair os acadêmicos, por ter sido escolhida pelo Dicionário Oxford como a palavra do ano, cujo verbete lhe dava o seguinte significado: "relativo a ou que denota circunstâncias nas quais fatos objetivos são menos influenciadores na formação da opinião pública do que apelos à emoção ou à crença pessoal" (G1, 2016).

Assim sendo, já em 2017, apareceram duas publicações no Brasil com a questão da pós-verdade em seus títulos. Assinado por Dunker *et al.* (2017), o livro *Ética e pós-verdade*, por ter sido escrito antes que a questão das fake news tivesse explodido, está mais voltado para os jogos interpretativos da expressão "pós-verdade", quase sempre preocupados com o significado do prefixo "pós", inclusive nas correlações com a pós-modernidade (ibid. p. 8, 95). Trata-se de uma correlação infeliz, em primeiro lugar, porque perde o foco do sentido situado da pós-verdade no contexto das fake news. Em segundo lugar porque a relação com a pós-modernidade exigiria um estudo muito bem fundamentado da complexidade multifacetada desse conceito na sua evolução desde os anos 1980 até os nossos dias.

A outra publicação de 2017 sobre pós-verdade voltou-se para a sua relação com a educação (Chates, org., 2017). Os textos têm o mérito de colocar sob sua mira as transformações dos processos educacionais a partir da avalanche de informações veiculadas nas redes digitais e em que medida isso afeta a construção do conhecimento e da verdade e os procedimentos de formação. Embora apareça no título, o sentido situado de pós-verdade não é discutido no livro.

De fato, foi só a partir de 2018 que as publicações começaram a atacar as questões correlatas das formações de bolhas – também chamadas de câmaras de eco e vieses da confirmação – com as fake news e a pós-verdade, o que conduziu à constituição mais própria daquilo que chamo

de sentido situado desta última (Ferrari, 2018; Santaella, 2018). Publicado no mesmo ano, o texto de D´Ancone (2018) tem mais um caráter de denúncia, especialmente contra Trump, do que de pesquisa de escritos sobre o tema. Como parece ser de praxe, o autor retorna ao pretenso parentesco da pós-verdade com a pós-modernidade. Esta vê-se reduzida aos autores do pós-estruturalismo que, lamentavelmente, são pasteurizados sob o rótulo de construtivistas (ibid., p. 85). Além disso, o livro baseia-se em uma visão simplificada da verdade (p. 121). Se, de fato, é necessário mantê-la, também é preciso considerar que, embora as mentiras possam ser absolutas, não há verdade absoluta, uma questão que não necessita ser discutida estritamente no campo do relativismo. Ainda de 2018 é o livro de Keyes que se dedica à discussão prolongada da mentira, revelando uma preocupação com a oposição entre mentira e honestidade como se a honestidade tivesse, por si só, um poder de combate contra o enxame de fake news.

O ano de 2018 foi a data das eleições presidenciais no Brasil, momento em que as correlações entre os três fatores (bolhas, fake news e pós-verdade) esquentaram, devidamente acompanhadas pela batalha travada contra a mentira, quase sempre de teor político, pelas instituições de checagem dos fatos, uma batalha travada não apenas contra a mentira humana, mas também contra a sua extensão nos *bots*.

Os *bots*, abreviação de robôs, são programas de software que executam repetitivamente tarefas automatizadas, pré-definidas. São insidiosos porque imitam comportamentos humanos. Lima (2018) nos informa que, por serem programados para espalhar informações muito mais rápido do que seres humanos conseguem fazer, as redes sociais são bombardeadas com as fake news, "causando um efeito dominó: quanto mais pessoas reais têm contato com a notícia falsa, mais elas acreditam que a informação é verídica, e acabam por compartilhá-la."

Pior do que isso, a falsidade tem o poder de se difundir "mais longe, mais rápido, mais profundamente e mais amplamente do que a verdade em todas as categorias de informação". Enquanto uma informação falsa necessita de aproximadamente 10 horas para alcançar 1500 usuários no Twitter, uma informação verídica precisa de 60 horas. Disso se conclui "que o fator humano é mais importante na disseminação de notícias falsas que os bots em si" (ibid.).

A partir dessas comprovações, fica claro que não pode haver outro meio de estancar a proliferação de fake news sem que haja leis que regulamentem o funcionamento das plataformas de redes sociais com rigor. A par desse meio, alguns pesquisadores, entre os quais me coloco, defendem que legislações apenas não são suficientes, mas é preciso também desenvolver programas educativos para e nas redes.

A partir de 2019, não é preciso sair do território brasileiro para constatar que as publicações sobre as fake news começaram a se multiplicar. É de 2019 o livro organizado por Barbosa, contendo uma diversidade de subtemas tais como o da credibilidade e confiança que prescindem da verdade e que estiveram nas bases da eleição de Bolsonaro (Bruno e Roque, 2019). Ou então, a crítica de Bucci (2019), em defesa do jornalismo ético, o qual implica constatar que "news não são fake – e fake news não são news". Ou ainda, a complexa discussão sobre a "liberdade de expressão ou dever de falar a verdade" (Macedo Júnior, 2019). Depois de discutir os problemas envolvidos nesse dilema, Macedo chama atenção para os perigos das soluções pouco democráticas de projetos de lei que podem impor algum tipo de censura, o que pode também "representar uma nova ameaça ao ainda pouco consolidado pensamento nacional sobre a liberdade de expressão" (ibid. p. 83).

Se, de um lado, as redes sociais adicionam um ganho para a pluralidade informacional da esfera pública, de outro lado, elas produzem efeitos colaterais nefastos que desembocam em patologias sociais. "A propagação viral de discursos de ódio, de conteúdo racista e xenófobo, atinge não apenas os grupos ou indivíduos diretamente atacados, mas todos aqueles que nas sociedades abertas defendem a liberdade" (ABBOUD *et al.*, 2018, p. 167). Entram, assim, em pauta os dilemas relativos a regular o ecossistema digital. Para começar, segundo os autores, "o grande desafio trazido pela proliferação aparentemente incontrolável dos discursos de ódio e das fake news nas redes sociais é encontrar soluções adequadas à instantaneidade da era digital, e não meras transposições de técnicas de aplicação do Direito que têm origem na era analógica e não oferecem soluções inteiramente satisfatórias" (ibid.).

No caso do Brasil, enquanto se dava o comprovado avanço promovido pelo marco civil da internet relativo à segurança e privacidade, a enxurrada de fake news acabou por se converter em questão judicial. A fonte geradora da perversão tinha sua proveniência no gabinete do ódio que se instalou no centro do poder. Vítimas desse ódio foram notoriamente mulheres jornalistas, em especial Patrícia Campos Mello, repórter especial da *Folha de S. Paulo* que se especializou no fenômeno da manipulação de narrativas em campanhas eleitorais, nos Estados Unidos, na Índia e no Brasil. Durante as eleições de 2018 no Brasil, Mello publicou uma série de reportagens sobre o financiamento do disparo em massa de notícias falsas em benefício do então candidato Jair Bolsonaro. A reação imediata foi vitimizar a jornalista em uma violenta campanha de difamação e in-

timidação, tornando-se ela mesma objeto de fake news. Sua coragem no enfrentamento dessa batalha e em defesa da liberdade de expressão lhe valeu prêmios, inclusive, recentemente, na França. Em seu livro sobre "A máquina do ódio", Mello (2020) fez o relato dos bastidores desses eventos, funcionando como um manifesto que advoga a favor da liberdade da informação.

O ano de 2020, avançando para 2021, foi pródigo em publicações que exploram a questão das fake news, fornecendo-nos um panorama multifacetado para compreender esse fenômeno e encontrar os caminhos mais eficazes para enfrentá-lo. Por exemplo, Cursino *et al.* (2021), no livro "Discurso e (pós) verdade", exploram a verdade e a pós-verdade nas suas relações com o discurso. Chegam, assim, bem perto da natureza semiótica das fake news. Faustino (2020), por sua vez, discute a liberdade de expressão no contexto das fake news, aliás, um limiar bastante difícil de ser estabelecido para a legislação das fakes news que não fira princípios de liberdade de expressão.

Era de se esperar que a área de Direito se visse diretamente interpelada pelos dilemas das fake news e pós-verdade. Não é casual que publicações sobre isso não cessem de aparecer. Rais (org., 2018) apresenta um panorama interdisciplinar sob o carro-chefe do direito. Menezes (2020) enriquece o pensamento sobre regulação ao colocá-lo, com cuidados metodológicos, no contexto da modernidade. Na sua preocupação com a preservação dos valores democráticos, sugere a necessidade de um conceito original de "consenso paradigmático" nas sociedades pluralistas que, para isso, dependem de instituições sociais fortes que possam fazer frente às contingências das instabilidades informacionais. Por fim, Jorge *et al.* (2021), para criar uma espécie de anteparo às eleições que virão em 2022, no Brasil, publicaram uma espécie de tratado de enfrentamento sob o título de "Fake news e eleições. O guia definitivo", com a apresentação auxiliar de ferramentas técnicas e jurídicas.

Longe de ser exaustivo, o panorama acima pretende funcionar, de um lado, como um mostruário da atenção que as fake news estão despertando e, de outro lado, como um indicador de que os problemas e dilemas que as fake news apresentam só tendem a se intensificar com as *deepfakes*, daí a importância de se tratar sem demora deste tema que está emergindo e que promete trazer danos para o futuro. Variadas definições de deepfakes serão encontradas nos artigos presentes neste número da TECCOGS, dedicado justamente a essa questão. O dossiê não tem a finalidade de antecipar o que os artigos irão discutir, mas apenas levantar o que parece ser a distinção fundamental entre fake news e deepfake e as consequências, ao mesmo tempo, similares, mas, sobretudo, distintas que estes últimos podem trazer.

Enquanto as fake news têm, na sua grande maioria, uma natureza semiótica verbal, as deepfakes fogem desse domínio para penetrar no reino da visualidade e sonoridade como base para o verbal audível. O primeiro problema que surge diz respeito ao fato de que, até certo ponto, as fake news seguem um padrão composicional que pessoas relativamente bem informadas já conseguem detectar. Alguns dos elementos desse padrão encontram-se, por exemplo, no design pouco sofisticado, tendendo para um sensacionalismo indisfarçável. Enquanto para alguns, isso é gerador de suspeita, de outro lado, para outros, induz credulidade, conforme foi discutido por Beiguelman (2021, p. 276) sobre os designs dos vídeos caseiros que explodiu com o YouTube. Trata-se aí de uma estética que procura se contrapor "ao imaginário tecnicamente perfeito do padrão de qualidade hollywoodiano (ou da Rede Globo)". Dispensando mediações tecnológicas mais sofisticadas, as imagens funcionam "como se fossem decalques do real, sem nenhuma interferência dos meios que a produzem e de quem os instrumentaliza. É nessa idealizada contraposição que reside a eficácia da estética amadora" (ibid.). Levado para as deepfakes, esse é um fator que, longe de gerar suspeita, intensifica a credibilidade.

Em segundo lugar, a natureza assertiva das fake news, sua natureza sempre falseadora em relação àquilo a que se refere, ou seja, os fatos que ela noticia, permite que essa referência seja testada, como o é, pelas organizações de checagem de fatos. Mesmo quando fake news assumem o tom genérico de uma voz proveniente de fontes desconhecidas, como, por exemplo, "celulares dão câncer", embora esse registro de linguagem pareça vir de um oráculo, existem fontes de informação confiáveis em que a afirmação pode ser testada. Diante disso, aonde a distinção em relação às deepfakes aparece?

De acordo com a semiótica, a ação interpretativa de um discurso verbal é muito distinta daquela de um signo visual, em especial quando se trata de um vídeo. Mas, antes do vídeo, retomemos as questões levantadas pela evolução da fotografia. Durante algum tempo, o caráter documental da fotografia foi inquestionável, pois a imagem fotográfica funciona, pretensamente, como um duplo confiável daquilo que ela captura. Trata-se, portanto, do registro de um fragmento de fato existente na realidade. Vem daí o emprego da fotografia no jornalismo, assim como o apreço dos historiadores pelos documentos fotográficos.

Contudo, a partir das manipulações das fotografias pelos programas computacionais, o que se deu a partir dos anos 1980, a crença até então relativamente inabalável de que as fotos não poderiam mentir foi colocada em questão (Santaella e Nöth, 2012). A descrença só tendeu a se intensificar com a introdução de programas cada vez mais sofisticados de manipulação da imagem, extremamente fáceis de colocar em uso.

Não obstante a suspeita relativa ao caráter documental da fotografia haver sido, até certo ponto, incorporada pela cultura, existe uma questão de base concernente aos processos perceptivos humanos que precisa ser considerada na medida em que afeta justamente a problemática das deepfakes.

A sutilíssima teoria lógica da percepção desenvolvida por C. S. Peirce revela que os julgamentos perceptivos instantâneos, que resultam dos processos perceptivos humanos, são indubitáveis (Santaella, em progresso). Não somos capazes e, portanto, não podemos duvidar daquilo que vemos. Os julgamentos perceptivos só podem mudar quando, por algum motivo, eles são colocados em dúvida, de modo que o processo perceptivo é reencenado para sanar a dúvida. Isso significa que, em primeira instância, não somos capazes de duvidar das deepfakes. Enquanto os signos fotográficos, por serem estáticos, já demonstram o hiato espaço-temporal entre o registro e a realidade que flui e se transforma para além desse registro, no caso dos vídeos, cria-se, a par do indubitável do que os olhos veem, o pacto narrativo que impede a suspensão da crença. Tudo isso funciona como um indicador de que muita campanha educativa terá que ser desenvolvida para que as deepfakes sejam colocadas sob suspeita como necessariamente devem ser.

Referências

ABBOUD, Georges; NERY JÚNIOR; Nelson; CAMPOS, Ricardo. *Fake news e regulação*. São Paulo: Thompson Reuters, Revista dos Tribunais, 2018.

BARBOSA, Mariana (org.). *Pós-verdade e fake news*: reflexões sobre a guerra de narrativas. Rio de Janeiro: Cobogó, 2019.

BEIGUELMAN, Giselle. *Políticas da imagem*: vigilância e resistência na dadosfera. São Paulo: Ubu, 2021.

BRUNO, Fernanda; ROQUE, Tatiana. A ponta de um iceberg de desconfiança. *In*: BARBOSA, Mariana (org.). *Pós-verdade e fake news*: reflexões sobre a guerra de narrativas. Rio de Janeiro: Cobogó, 2019, p. 13-24.

BUCCI, Eugênio. News não são fake – e fake news não são news. *In*: BARBOSA, Mariana (org.). *Pós-verdade e fake news*: reflexões sobre a guerra de narrativas. Rio de Janeiro: Cobogó, 2019, p. 37-48.

CHATES, Tatiana de Jesus (org.). *Perspectivas educacionais em tempos de pós-verdade*. Jundiaí: Paco, 2017.

CURSINO, Carlos; SARGENTINI, Luzmara; PIOVEZANI, Vanice. *Discurso e (pós)verdade*. São Paulo: Parábola, 2021.

D'ANCONE, Matthew. *Pós-verdade*: a nova guerra contra os fatos em tempos de fake news. Tradução Carlos Szlak. Barueri: Faro, 2018.

DUNKER, Christian; TEZZA, Cristovão; FUKS, Julián; TIBURI, Marcia; SAFATLE, Vladimir. *Ética e pós-verdade*. Porto Alegre: Dublinense, 2017.

FAUSTINO, André. *Fake news*: a liberdade de expressão nas redes sociais na sociedade da informação. São Paulo: Lura, 2020.

FERRARI, Pollyana. *Como sair das bolhas*. São Paulo: Educ, 2021.

G1. Pós-verdade é eleita a palavra do ano pelo Dicionário Oxford. *Globo Notícias*, 16/11/2016. Disponível em: g1.globo.com/educacao/noticia/pos-verdade-e-eleita-a-palavra-do-ano-pelo-dicionario-oxford.ghtml. Acesso em: 12 ago. 2021.

JORGE, Higor Vinicius N.; JORGE JÚNIOR, Hélio Molina; NOVAIS, Kayki; FONSECA, Ricardo Magno T. *Fake news e eleições*: o guia definitivo. Salvador: JusPODIVM, 2021.

KEYES, Ralph. *A era da pós-verdade*: desonestidade e enganação na vida contemporânea. Tradução Fábio Creder. Petrópolis: Vozes, 2018.

LIMA, Ramalho. Estudo revela que *bots* espalham fake news massivamente em poucos segundos, 2018. *Tecmundo*. Disponível em: tecmundo.com.br/internet/136479-estudo-revela-bots-espalham-fake-news-massivamente-segundos.htm. Acesso em: 12 ago. 2021.

MACEDO JÚNIOR, Ronaldo P. Liberdade de expressão ou dever de falar a verdade. In: BARBOSA, Mariana (org.). *Pós-verdade e fake news*: reflexões sobre a guerra de narrativas. Rio de Janeiro: Cobogó, 2019, p. 79-86.

MELLO, Patrícia Campos. *A máquina do ódio*: notas de uma repórter sobre fake news e violência digital. São Paulo: Companhia das Letras, 2020.

MENEZES, Paulo Brasil. *Fake news*: modernidade, metodologia e regulação. Salvador: JusPODIVUM, 2020.

RAIS, Diogo (org.). *Fake news*: a conexão entre a desinformação e o direito. São Paulo: Revista dos Tribunais, 2018.

SANTAELLA, Lucia. *A pós-verdade é verdadeira ou falsa?* São Paulo: Estação das Letras e Cores, 2018.

_____. *Não somos capazes de duvidar do que vemos*. Em progresso.

SERVA, Leão. *Jornalismo e desinformação*. São Paulo: Senac, 2001.

ARTIGOS

SOUZA, Carlos Eduardo; SANTAELLA, Lucia. Deepfakes na perspectiva da semiótica. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 26-44.

Recebido em: 10 fev. 2021
Aprovado em: 1 mar. 2021

dx.doi.org/10.23925/1984-3585.2021i23p26-44

Deepfakes na perspectiva da semiótica

Carlos Eduardo de Souza[1]

Lucia Santaella[2]

Resumo: As notícias falsas são uma arma de desinformação conhecida, capaz de ameaçar o estado democrático. Em um momento em que as mídias tradicionais são constantemente atacadas e acusadas de fazerem parte de uma grande conspiração para manter o poder das classes dominantes, as pessoas transformaram as redes sociais em sua fonte primária de informação, pois nelas existe menos controle sobre o que circula e, pretensamente, trazem menor risco de manipulação pelas mídias tradicionais. Na intersecção de fatores como a democratização na produção de conteúdo sem qualquer supervisão, a personalização das mensagens que não confrontam o usuário e o estímulo ao compartilhamento, as fake news encontram um campo para florescer e se propagar. Em 2017 elas evoluíram, com o nome de deepfake, deixaram de ser apenas mensagens no formato de texto, para contar com a manipulação de imagens, áudios e vídeos. Sua capacidade de forjar a realidade de maneira praticamente imperceptível, até mesmo para especialistas, chamou a atenção das mídias e da academia. Para discutir os efeitos e as ameaças dessa tecnologia, e como combatê-las, esse artigo, baseado na semiótica desenvolvida por Peirce, aponta para o esforço feito pelos produtores do vídeo *In Event of Moon Disaster* em criar um conteúdo educativo para alertar as pessoas sobre as consequências das deepfakes.

Palavras-chave: Deepfake. Fake News. Semiótica. Peirce. Redes Sociais.

1 Carlos Eduardo de Souza é bacharel em Administração pela Universidade Presbiteriana Mackenzie e mestrando em Tecnologias da Inteligência e Design Digital, PUC-SP. CV Lattes: lattes.cnpq.br/4424563031670368. E-mail: cadu.souza81@gmail.com.

2 Lucia Santaella é pesquisadora 1A do CNPq, professora titular da PUC-SP. Publicou 51 livros e organizou 24, além da publicação de mais de 400 artigos no Brasil e no exterior. Recebeu os prêmios Jabuti (2002, 2009, 2011 e 2014), o prêmio Sergio Motta (2005) e o prêmio Luiz Beltrão (2010). ORCID: orcid.org/0000-0002-0681-6073. CV Lattes: lattes.cnpq.br/7427854657719431. E-mail: lbraga@pucsp.br.

Deepfakes in a semiotic perspective

Abstract: Fake news is a known weapon of disinformation capable of threatening democracy. At a time when traditional media are constantly attacked and accused of being part of a major conspiracy to maintain the power of the ruling classes, people have turned social networks into their primary source of information. As there is less control over what circulates there, it is supposed that they bring less risk of manipulation by traditional media. At the intersection of factors such as the democratization of content production without any supervision, the personalization of messages that do not confront the user and the encouragement of sharing, fake news finds a field to flourish and spread. In 2017, under the name of deepfake, they evolved from being just messages in text format, to relying on the manipulation of images, audio and videos. Its ability to forge reality practically imperceptibly, even for specialists, caught the attention of the media and academia. To discuss the effects and threats of this technology, and how to combat them, this article, based on the semiotics developed by Peirce, points to the effort made by the producers of the video "In Event of Moon Disaster" to create educational content to alert people about the consequences of deepfakes.

Key words: Deepfake. Fake News. Semiotics. Peirce. Social networks.

Fake news

No contexto da globalização, as corporações de mídias enfrentam desafios que atingem sua base estrutural, de formas e de conteúdo. Nesse cenário, acadêmicos e jornalistas, entre outros profissionais, são frequentemente confrontados com o fenômeno das notícias falsas (ou fake news) (BERDUYGINA; VLADIMIROVA; CHERNYAEVA, 2019). Em 2017, o termo "fake news" foi aclamado como termo do ano pelo dicionário Collins, seguindo o que o dicionário chamou de "presença onipresente" nos últimos doze meses então transcorridos. Segundo o monitoramento feito pelos lexicógrafos que trabalham para o Collins, a presença do termo cresceu 365% na comparação ano contra ano. Ainda de acordo com o trabalho do Collins, fake news, tem sido utilizado nos Estados Unidos para descrever "falsas, frequentemente sensacionalistas, informações disseminadas sob o pretexto de reportagem informativa" (BROWN, 2017; FLOOD, 2017, p. 1). De acordo com Santaella:

> Notícias falsas costumam ser definidas como notícias, estórias, boatos, fofocas ou rumores que são deliberadamente criados para ludibriar ou fornecer informações enganadoras. Elas visam influenciar as crenças das pessoas, manipulá-las politicamente ou causar confusões em prol de interesses escusos. (SANTAELLA, 2018, p. 263-265)

A definição apresentada pode abarcar uma grande variedade de formas de desinformação para fins comerciais e de publicidade, frequentemente com forte apelo visual (BERDUYGINA; VLADIMIROVA; CHERNYAEVA, 2019; SANTAELLA 2018). O contexto atual das fake news é substancialmente diferente daquele comumente encontrado no domínio dos meios tradicionais de comunicação, quando a produção de notícias era limitada e confiável, na medida em que seguia um conjunto de normas e princípios adotados pelos jornalistas. As novas formas de produzir e consumir informação e notícias não guardam padrões semelhantes.

Assim, o termo fake news passou a se referir a postagens virais baseadas em contas fictícias feitas para se parecerem com notícias. Vários tipos de notícias falsas estão identificados à medida que crescem os estudos sobre o tema (BOTHA; PIETERSE, 2020). Alguns dos tipos mais comuns são: (a) sátira ou paródia criadas para fins de entretenimento, sem

intenção de causar dano, mas potencialmente podendo enganar a audiência; (b) fabricação de notícias, podendo envolver manipulação de fotos ou vídeos, que são notícias não baseadas em fatos, porém publicadas como se fossem reportagens para transmitir legitimidade com o intuito de enganar; (c) manipulação de fotos, envolvendo a manipulação de imagens ou vídeos reais para estabelecer uma narrativa falsa; (d) conexão falsa, quando manchetes, imagens e/ou legendas incluem notícias ou artigos que contêm conteúdo genuíno e preciso, mas fazem uso de títulos enganosos ou sensacionalistas; (e) publicidade e relações públicas, em que a publicidade e os comunicados de imprensa são publicados como notícias, sendo muitas vezes um conteúdo patrocinado; (f) captura de cliques, notícias fabricadas propositalmente para ganhar mais visitantes em um site e aumentar a receita de publicidade; (g) conteúdo enganoso, transmitindo informações enganosas para enquadrar indivíduos ou questões (BORGES, GAMBARATO, 2019; BOTHA, PIETERSE, 2020).

Para os fins deste artigo, seguiremos com o conceito de fake news apresentado por Allcott e Gentzkow (2017, p. 2013), "artigos de notícias que são intencionalmente e comprovadamente falsos, e podem enganar os leitores". O uso de manchetes sensacionalistas para seduzir a audiência é uma prática de longa data. Entretanto, dentro das redes sociais, tanto o alcance, quanto o efeito da disseminação de conteúdo ocorrem em uma escala muito mais rápida, de modo que essa informação distorcida, imprecisa ou falsa adquire um enorme potencial para causar impactos reais, em poucos minutos (FIGUEIRA; OLIVEIRA, 2017).

De fato, a informação falsa espalha-se rapidamente pelas redes sociais, podendo impactar milhões de usuários (ibid.), como sugerem dois estudos conduzidos por Allcott e Gentzkow. No primeiro, constatou-se que 38 milhões de compartilhamentos de notícias falsas que aconteceram nas redes sociais, resultaram em 760 milhões de ocorrências de um usuário clicando e lendo uma notícia falsa, ou cerca de três histórias lidas por adulto americano. No segundo, uma lista de sites de notícias falsas, na qual pouco mais da metade dos artigos parecem ser falsos, recebeu 159 milhões de visitas durante o mês da eleição, ou 0,64 por adulto nos EUA (ALLCOTT; GENTZKOW, 2017).

De acordo com Westerlund (2019), atualmente, 20% dos usuários da Internet obtêm suas notícias via YouTube, sendo essa porcentagem menor apenas que a de usuários que obtém informações pelo Facebook. A crescente popularidade do vídeo mostra a importância da criação de ferramentas que confirmem a autenticidade do conteúdo dessa mídia e notícias, pois à medida que as novas tecnologias permitem a alteração convincente do conteúdo, torna-se mais fácil obter e divulgar informações incorretas através das mídias sociais.

À medida que o crescimento das redes sociais vem destruindo as barreiras de entrada que existiam para prevenir a disseminação das fakes news, esse fenômeno permite que qualquer pessoa possa criar e disseminar conteúdo (BERDUYGINA; VLADIMIROVA; CHERNYAEVA, 2019). Suportada pela lógica das redes sociais e dos buscadores de informação, de facilidade de publicação e compartilhamento de conteúdo sem qualquer avaliação de terceiros, sem checagem dos fatos ou critério editorial, qualquer conteúdo produzido pode atingir milhões de usuários (ALLCOTT; GENTZKOW, 2017). O ápice dessa lógica parece ter sido atingido pela popularização das mídias móveis, que tornaram qualquer lugar um ponto de produção e compartilhamento instantâneo de informação. Imagem, som e vídeo podem ser criados e disseminado por milhões de pessoas para milhões de pessoas em diversas plataformas, por usuários que muitas vezes desconhecem o funcionamento dos algoritmos que funcionam nessas redes (SANTAELLA, 2018).

A partir do momento em que cresce a descrença nas mídias tradicionais, fica aberto o espaço para que as mídias alternativas as desqualifiquem como não confiáveis. Ademais, à medida que as notícias falsas tornam-se cada vez mais sofisticadas, criam-se as condições para o império da "pós-verdade", que é caracterizado pela desinformação digital e guerra de informação liderada por atores malévolos executando campanhas de informações falsas para manipular a opinião pública e aumentar a polarização política, usando para isso canais de comunicação alternativos, independentes e descentralizados sem qualquer compromisso com a informação factual (WESTERLUND, 2019).

A impulsão dessas informações conta com diversas heurísticas cognitivas que compõem três fenômenos em particular – a dinâmica da "cascata de informações", a atração humana por informações negativas e novas, e as bolhas de isolamento especialmente úteis para explicar por que notícias se tornam virais (CHESNEY; CITRON, 2019). A cascata de informações está relacionada com a dinâmica que surge a partir do momento em que as pessoas não prestam atenção nas informações que estão compartilhando, pois presumem que outros determinaram de forma confiável a credibilidade da informação antes de transmiti-la (ibid.). Sendo levadas a questionar toda a informação que recebem das mídias tradicionais, as pessoas se protegem usando como fontes confiáveis suas redes sociais e buscam opiniões que apoiem suas ideias existentes. Na verdade, muitas pessoas estão abertas a qualquer coisa que confirme suas visões existentes, mesmo que suspeitem que seja falso (WESTERLUND, 2019).

A lógica dos algoritmos de recomendação tem um papel muito relevante no contexto das notícias falsas, possivelmente criando bolhas de isolamento, uma vez que o algoritmo assume como verdade a predileção do usuário por determinado conteúdo, assim como cria câmaras de eco, já que há um maior peso para conteúdo publicado e compartilhado por pessoas que declaram ter as mesmas orientações do usuário, gerando, assim, um isolamento contra opiniões contrárias (BORGES; GAMBARATO, 2019; CHESNEY; CITRON, 2019).

Pesquisas demonstram que as pessoas frequentemente ignoram informações que contradizem suas crenças e interpretam evidências ambíguas como consistentes desde que alinhadas com as suas crenças (CHESNEY; CITRON, 2019). A consequência desse fenômeno é a fragmentação da "esfera pública" em subesferas fortemente separadas, cada uma delas fluindo de acordo com sua própria interpretação do "mundo autocongruente, geralmente intolerante e agressiva para as demais subesferas, com as quais a possibilidade de o diálogo e a compreensão diminuiriam rapidamente" (CITTON, 2021, p. 49).

Outra característica das redes sociais é sua operação com base na quantidade de cliques e volume de tráfico que uma determinada publicação recebe, uma vez que não há juízo de valor, basta a menor interação com o conteúdo falso para que o algoritmo contabilize esse fato como interesse naquele conteúdo (SANTAELLA, 2018). A partir do momento em que essa interação é capturada, o processo de personalização do conteúdo se encarrega de perpetuar os ecos que passam através dos filtros da bolha criada (BORGES; GAMBARATO, 2019). Todo o processo está baseado em grandes bancos de dados, nos quais são armazenadas informações sobre quais sites o usuário acessou, com que perfis interagiu e de que forma, o que foi compartilhado; praticamente todas as interações feitas através do computador são registradas e utilizadas para selecionar as informações que o usuário receberá. Esse grau de personalização pode ter consequências na forma como o usuário avalia o mundo e toma decisões (GUARDA; OHLSON; ROMANINI, 2018).

Conforme praticamente todas as nossas interações na internet são coletadas, agrupadas e analisadas por algoritmos de Inteligência Artificial (IA), as redes sociais e buscadores de informação passam a controlar e moldar nosso consumo de informação. Assim, a IA torna-se extremamente capaz de prever o que nos seduz, vicia ou enfurece, para manter-nos clicando, lendo, assistindo e compartilhando. Ademais, essa mesma IA é capaz de forjar mídias sintéticas em tempo real calibrado para explorar esses vieses (FLETCHER, 2018). Nessa intersecção, reside o combustível para o pesadelo do fim das democracias.

Ao fim e ao cabo, o poder das notícias falsas provém do apelo ao emocional, que captura rapidamente a atenção do leitor, levando-o, além do clique, ao compartilhamento, apenas com base na manchete, sem qualquer filtro de criticidade ao conteúdo da postagem (SANTAELLA, 2018). No modelo de análise das redes sociais como fonte de informação desenvolvido por Allcott e Gentzkow (2017, p. 221), também encontramos elementos que demonstram a permissividade dessa plataformas, (a) baixos custos de entrada e produção de conteúdo, o que torna as estratégias de produção de notícias falsas bastante rentáveis; (b) o formato em que as postagens são exibidas torna difícil avaliar a veracidade de um artigo; (c) nas redes sociais as pessoas estão mais propensas a ler o que pessoas com a mesma orientação ideológica publicam e compartilham.

Santaella (ibid.), destaca três traços das notícias falsas que instigam sua propagação pela internet, "desinformação, desconfiança e manipulação". Considerando que atravessamos um momento de persistência da "pós verdade", termo que a literatura psicológica trata como viés de confirmação, ou seja, a tendência que os indivíduos têm em praticar uma escuta seletiva (ROMANINI; OHLSON, 2018), a relação dos usuários nas mídias sociais é mais influenciada pelas emoções do que pela criticidade quanto ao conteúdo adulterado, que ganha cada vez mais espaço para crescimento, uma vez que as emoções e crenças pessoais tornam-se mais importantes na formação da opinião pública do que os fatos verificados (GUARDA; OHLSON; ROMANINI, 2018).

Nesse contexto, as revelações das pesquisas em psicologia social e outros campos aumentam as preocupações quanto ao consumo de informação, pois ficam minadas as antigas teorias de humanos como agentes que avaliavam friamente as evidências frente aos seus valores e objetivos, para então modificar essas crenças em respostas às evidências. Tais pesquisas apontam o contrário, para o fato de que primeiro adquirimos crenças e apegos, e que posteriormente filtramos as evidências, muitas vezes inconscientemente, de forma a manter a coerência das nossas identidades (FLETCHER, 2018).

> Nós alinhamos nossas lentes de crença / interpretação com as de nosso grupo social (tanto online quanto off-line), explicando evidências e interpretações que nos alienariam do "nós" do nosso grupo. Da mesma forma, detectamos e evitamos evidências e interpretações de que se identifique com "eles". (FLETCHER, 2018, p. 466)

Os relatos das notícias falsas excluem a realidade externa ou a distorcem propositalmente. Separar notícias falsas de notícias verdadeiras baseia-se em conhecer a relação entre a reportagem, o fato externo ao qual ela se refere e qual a intenção por trás da reportagem. Evidentemen-

te a linha entre o uso de processos equivocados e ações deliberadamente enganosas não é clara, e a verdade não é intrínseca à notícia; a esta cabe o papel de narrar um acontecimento (BORGES; GAMBARATO, 2019). Sob esse aspecto, a semiótica tem uma contribuição a dar, conforme será discutido adiante. Por ora, é importante apontar para a intensificação dos problemas quando as fake news são convertidas em deepfakes.

Das fake news às deepfakes

O aspecto histórico das fake news não é algo novo e há razões para crer que o papel desse tipo de conteúdo continuará a crescer. Com menores barreiras de entrada, incentivos a monetização de conteúdo on-line, queda na confiança nos meios tradicionais de comunicação em massa, crescimento das mídias sociais e o aumento da polarização política (ALLCOTT; GENTZKOW, 2017), novas formas de fake news. Surgiram com potencial ainda maior de enganar a audiência, como são as deepfakes.

A primeira aparição da tecnologia deepfake foi na plataforma de mídia social *Reddit* publicada por um usuário anônimo em novembro de 2017 (BOTHA; PIETERSE, 2020). As deepfakes podem ser definidas como mídias sintéticas geradas com o uso de IA. É uma junção das expressões *deep learning* (aprendizado profundo) e *fake* (falso) (ASHISH, 2020; BATTAGLIA, 2020). Deepfakes também podem ser entendidas como o produto de aplicativos de IA que fundem, combinam, substituem e sobrepõem imagens e vídeo clipes para criar vídeos falsos que parecem autênticos (WESTERLUND, 2019). Para nossos propósitos, seguiremos com a definição apresentada por Chesney e Citron (2019, p. 1757-1758),

> Isso deu origem ao rótulo de "deepfakes" para essas personificações digitalizadas. Usamos esse rótulo aqui de forma mais ampla, como uma abreviatura para toda a gama de falsificações digitais hiper-realistas de imagens, vídeo e áudio. Esta gama completa implicará, mais cedo ou mais tarde, uma perturbadora gama de usos maliciosos. Não somos de forma alguma os primeiros a observar que *deep fakes* irão migrar muito além do contexto da pornografia, com grande potencial de danos.

Assim, como uma combinação de "*deep learning*" e *fake*, as deepfakes são mídias sintéticas geradas por IA, altamente realistas, de difícil detecção (CHESNEY; CITRON, 2019). A partir dessa tecnologia é possível produzir vídeos hiper-realistas manipulados digitalmente para representar pessoas dizendo e fazendo coisas que nunca realmente aconteceram. Contando com redes neurais que analisam grandes conjuntos de amostras de dados para aprender a imitar as expressões faciais, maneirismos, voz e inflexões

de uma pessoa, é possível colocar qualquer um em qualquer situação (WESTERLUND, 2019).

No passado, a manipulação de vídeos exigia grandes recursos e estava à disposição de poucas empresas. Atualmente, entretanto, computadores domésticos permitem "uma assustadoramente precisa troca de rosto em um único computador de jogo, possivelmente em menos de vinte e quatro horas. Sem equipe, sem recursos, sem dinheiro" (FLETCHER, 2018, p. 463). Essas tecnologias têm o potencial de criar falsificações mais reais e profundas, mais difíceis de detectar, com um potencial ainda maior de sabotagem à democracia, pois permitem a produção de vídeos de notícias aparentemente legítimas que põem em cheque a reputação de jornalistas e da mídia, colocam pessoas em lugares e situações nas quais elas nunca estiveram com o interesse em distorcer a opinião pública (GUARDA; OHLSON; ROMANINI, 2018; WESTERLUND, 2019). Sua eficácia e suas ameaças não estão apenas em sua capacidade de forjar realidades, "mas sim em sua capacidade de ressoar dentro do atual estado afetivo das multidões" (CITTON, 2021, p. 50).

A disponibilidade dos algoritmos utilizados para a criação de mídias sintéticas a partir de IA permitiu a rápida automação do processo de deepfake. Para criar um vídeo com conteúdo falso é necessário apenas a seleção de imagens da face da pessoa que será substituída e da face da pessoa que será sobreposta (BOTHA; PIETERSE, 2020). Com tal facilidade disponível torna-se cada mais difícil saber em que confiar, o que resulta em prejuízos para a tomada de decisão, entre outras coisas (WESTERLUND, 2019).

O ponto de inflexão das deepfakes encontra-se no escopo, escala e na sofisticação da tecnologia envolvida, já que quase qualquer pessoa com um computador pode fabricar vídeos falsos que são praticamente indistinguíveis da mídia autêntica (FLETCHER, 2018). É provável que, no futuro, as deepfakes evoluam para pornografia de vingança, *cyberbullying*, desqualificação de provas em tribunais, sabotagem política, propaganda terrorista, chantagem, manipulação de mercado e notícias falsas (WESTERLUND, 2019).

Com os avanços tecnológico, é razoável apostar em um futuro no qual a guerra da desinformação estará mais bem estruturada e o uso da IA para criar e entregar vídeos falsos adaptados aos preconceitos específicos dos usuários de mídia social estará disseminado. As chamadas deepfakes serão a desinformação transformada em armas destinadas a interferir nas eleições e semear agitação civil (ibid.).

As redes sociais e buscadores de internet, ao proporcionarem o contato frequente com a desinformação, fazem com que as pessoas não se sintam confiantes em toda a informação disponível, resultando em um fenômeno denominado de "apocalipse da informação" ou "apatia da realidade". Além disso, as pessoas podem até descartar as filmagens genuínas como falsas, simplesmente porque se enraizaram na noção de que tudo em que não querem acreditar deve ser falso. Em outras palavras, a maior ameaça não é que as pessoas sejam enganadas, mas que passem a considerar tudo como engano (ibid.).

Em um mundo onde praticamente tudo pode ser manipulado artificialmente, a "apatia da realidade" pode representar uma grave ameaça à capacidade da sociedade compartilhar e cooperar com base em "percepções precisas até mesmo das mais básicas ou urgentes realidades, tornando as sociedades ainda mais vulneráveis ao governo autocrático" (FLETCHER, 2018, p. 465).

Nessa batalha cibernética, as técnicas computacionais para detecção das deepfakes estão focadas em identificar um artefato específico, mas não generalizam bem para detectar novas versões. O desempenho dos detectores de última geração está diminuindo rapidamente à medida que a qualidade das deepfakes melhora (MIRSKY; LEE, 2021; KORSHUNOV; MARCEL, 2018). A divulgação de novos avanços no combate às deepfakes leva a uma evolução da sua produção, criando uma corrida armamentista de IA, com o aperfeiçoamento das técnicas destinadas a gerar deepfakes cada vez mais sofisticadas e difíceis de detectar. Assim, trocamos um sabor de distopia de uma ficção científica por outra (FLETCHER, 2018), à medida que cada ciclo de novas tecnologias de combate à deepfake carrega a semente da evolução da próxima geração de deepfakes.

Tais condições só aumentam a importância da mídia confiável para a preservação da democracia. Essas preocupações são mais oportunas e relevantes do que nunca: vivemos em um contexto dominado pela "pós verdade", em uma era de crescente populismo autoritário, acompanhada de repressões em veículos de jornalismo legítimo e demandas de base por justiça ambiental e racial. Urge que os cidadãos obtenham informações confiáveis e um melhor entendimento de como se envolver com este ambiente de mídia fragmentado (MOONDISASTER, 2021).

Infelizmente, a alfabetização midiática não pode fornecer um antídoto mágico para o tremendo aumento da desinformação. Como estratégia pedagógica central, no entanto, pode nos ajudar a cultivar a crítica interior, transformando-nos de consumidores passivos da mídia em um público criterioso. Ensinar sobre deepfakes significa alertar para a ameaça perniciosa que representam, para as várias abordagens que visam combatê-las e para os usos alternativos de mídia sintética (ibid.).

Assim, abordagens como a educação e o treinamento são cruciais nesse combate. É preciso aumentar a conscientização pública tanto entre nativos digitais quanto pessoas mais velhas, para compreender que um vídeo, ao contrário do que parece, pode não fornecer uma informação precisa do que aconteceu, e encontrar quais pistas perceptivas podem ajudar a identificar e combater o efeito danoso das deepfakes (WESTERLUND, 2019 MIRSKY; LEE, 2021). Um caminho auxiliar que se apresenta para isso é também o da semiótica, conforme será discutido abaixo a partir de um caso exemplar.

In event of moon disaster

O caso exemplar, a ser comentado semioticamente, é a obra "In event of moon disaster". Em uma intersecção do campo da arte e política, os pesquisadores do MIT, Francesca Panetta e Halsey Burgund, propuseram uma ideia provocativa usando a tecnologia das deepfakes, através da qual procuram mostrar o potencial da tecnologia para educar as pessoas sobre seu uso. Eles escolheram recriar uma versão alternativa da viagem à Lua, realizada pela nave Apollo em 1969 (HAO, 2020; MOONDISASTER, 2021). Antes da missão, os redatores do presidente Richard Nixon elaboraram dois discursos, um para o cenário de sucesso e outro, designado "In event of moon disaster", para caso as coisas não saíssem de acordo com o planejado. O verdadeiro Nixon, nunca precisou proferir o segundo, porém, a partir da sobreposição de imagens e áudio os pesquisadores foram capazes de representar o que seria o presidente Nixon proferindo o segundo discurso (HAO, 2020). Questionados se eles não estariam disseminando informações falsas os autores da produção foram enfáticos em negar tal pergunta:

> Como artistas multimídia e jornalistas que trabalharam por uma década em um cenário de mídia em constante mudança, acreditamos que as informações apresentadas como falsas em um contexto artístico e educacional não são desinformações. Na verdade, pode ser fortalecedor: experimentar um uso poderoso de novas tecnologias de forma transparente que tem o potencial de ficar com os espectadores e torná-los mais cautelosos sobre o que verão no futuro. Usando as técnicas mais avançadas disponíveis e insistindo na criação de um vídeo usando visuais sintéticos e áudio sintético (um "deepfake completo"), pretendemos mostrar para onde essa tecnologia está caminhando - e quais podem ser algumas das principais consequências (MOONDISASTER, 2021, p. 1).

In event of moon disaster ilustra as possibilidades de tecnologias deepfake ao recriar uma versão alternativa da missão da Apollo 11. O ponto de partida é: o que teria acontecido caso algo de errado ocorresse e os astronautas não pudessem voltar para casa? Um discurso de contingência

para essa possibilidade foi preparado, mas nunca feito pelo presidente Nixon – até agora. Nesse projeto de arte foi criada uma história alternativa, pedindo a todos nós que consideremos como as novas tecnologias podem dobrar, redirecionar e ofuscar a verdade ao nosso redor (MOONDISASTER, 2021).

Para construir essa versão alternativa da história, várias técnicas de desinformação foram usadas – desde a edição enganosa simples até tecnologias deepfakes mais complexas. Para recriar o discurso de contingência, a peça usou técnicas de *deep learning* para criar uma voz sintética de Nixon e técnicas de substituição de diálogo para replicar o movimento da boca e dos lábios de Nixon. Ao criar essa história alternativa, o projeto explora a influência e a difusão da desinformação e das tecnologias das deepfakes em nossa sociedade contemporânea (MOONDISASTER, 2021). Esse trabalho funciona como um exemplar a ser utilizado para finalidades educativas sob o olhar da semiótica.

A contribuição da semiótica

A semiótica oferece uma série de elementos para a análise da interpretação "dos signos produzidos pelas novas tecnologias, assim como do seu papel potencial em ambientes com grandes concentrações de notícias falsas, como as redes sociais" (FERRAREZI; BORGES, 2020, p. 60). Assim, a semiótica, na vertente desenvolvida por C. S. Peirce, pode servir de bússola ao longo do debate sobre o tema,

> Pois, como ciência da significação, da denotação e da interpretação dos processos de linguagem e de comunicação, essa ciência pode nos oferecer conceitos fundamentais, capazes de nos guiar na tarefa de perscrutar os modos de produção, interpretação e disseminação das Fake News. [...] Uma ciência de base filosófica e, como ciência, cria conceitos com a finalidade de nos ajudar a pensar. Portanto, para apreender esses conceitos é preciso exercitar a paciência teórica. Só isso pode trazer compensações consequentes para os modos como interpretamos os problemas relativos às Fake News. (SANTAELLA, 2020, p. 13-14).

A semiótica tem no signo seu elemento central. Sob certo aspecto ou capacidade, o signo representa algo para um intérprete. Todas as formas de comunicação podem ser consideradas signos (SANTAELLA, 2020; BORGES; GAMBARATO, 2019). O signo é composto por três elementos, o signo, o objeto e o interpretante. Sua função representativa é aquela de mediar entre o objeto representado e o efeito que produz na mente do intérprete, efeito chamado de interpretante (BORGES; GAM-

BARATO, 2019; SANTAELLA, 2020; FERRAREZI; BORGES, 2020). À medida que os signos representam alguma coisa, ou seja, seu objeto em alguns de seus aspectos, com referência a algum um tipo de ideia, eles são, portanto, mais ou menos precisos e também podem ser usados para enganar, quando existe uma colisão em vez de uma correspondência entre o signo e aquilo que ele professa representar (BORGES; GAMBARATO, 2019).

Entretanto, o intérprete não está limitado à representação que um determinado signo faz do seu objeto. Sendo os signos por natureza incompletos, isso permite ao intérprete reportar-se ao contexto do objeto do signo por meio da experiência colateral que teve, tem ou poderá vir a ter com ele" (SANTAELLA, 2020, p. 16). Segundo esta autora (ibid., p. 18):

> Se o signo é parte de um contexto existencial, factual, maior do que ele, sua verdade ou falsidade pode ser averiguada por experiência colateral com o objeto do signo, quer dizer, o campo de referências do signo. Isso é justamente aquilo a que Hanna Arendt (1972) deu o nome de verdade factual, que é, efetivamente, a única classe de signo que, pelo fato de funcionar como um indicador, um índice de seu objeto de referência, a saber, o acontecimento, o fato ocorrido, pode ser interpretado como verdadeiro ou falso, por meio do rastreamento desse objeto de referência. Essa distinção sígnica precisa ser feita para se evitar que tudo, indiscriminadamente, entre no saco de gatos das Fake News.

Assim, a semiótica pode ser empregada como modelo de análise das deepfakes, e, em especial, do experimento conduzido por Francesca Panetta e Halsey Burgund, acima descrito, na medida em que esse caso pode contribuir para criar processos educativos que auxiliem as pessoas a repensarem suas visões do mundo, do outro e de si mesmos. Uma vez que fake news e deepfakes são inegavelmente signos, do tipo factuais, ou seja, representam fatos, elas implicam as noções de realidade (a realidade dos fatos) e de verdade (dentro de suas capacidades e limites o signo pode apresentar fidelidade aos fatos ou, então, mentir em relação a eles). Como se pode ver, os conceitos de realidade e verdade estão implícitos e decorrem da noção de signos (ROMANINI; OHLSON, 2018). Segundo NÖTH (2016, p. 137), o real se define como

> Independente dos meus e dos seus caprichos (CP 5.311). Em 1903, ele postula que o real é como é, independentemente de como imaginamos que ele seja (1903, CP 7.659). Também em 1877, Peirce define o real como independente de qualquer conhecedor, [...] o real Peirciano não é apenas um *ens*, um modo de ser; ele age sobre nossos sentidos. Existem coisas Reais cujas características são inteiramente independentes de nossas opiniões sobre elas.

Embora a realidade nela mesma seja algo que independe do que possamos pensar sobre ela, a realidade nos é acessível pela mediação do signo, podendo, assim, afetar nossos pensamentos, produzindo como efeito uma ideia, ou melhor, um interpretante coincidente ou não com a realidade. Assim, a contribuição da semiótica para refletir sobre o combate às fake news e deepfakes, reside na conexão necessária entre o pensamento e a realidade (BORGES; GAMBARATO, 2019).

Ora, signos são por natureza sociais. Portanto, na perspectiva de Peirce, os acontecimentos sociais estão conectados à experiência humana, e, logo, devem ser analisados em um processo lógico de investigação comunitário, interessado na verdade dos fatos. É no confronto dos nossos julgamentos isolados com o julgamento da comunidade que surge a verdade factual. Contudo, jamais estamos de pose da última verdade, no máximo nos esforçamos para atingir um estágio de crença, a partir do uso de recursos materiais e tempo, no qual nos sentimos confortáveis e não encontramos benefícios práticos em prosseguir para um grau maior de precisão, com mais investigação (ROMANINI; OHLSON, 2018).

De acordo com Peirce, há quatro métodos através dos quais atinge-se a fixação de uma ideia e se estabelece uma crença. Três oferecem o conforto da crença fácil, ao passo que limitam a busca da verdade: o método da tenacidade, no qual por afinidade o indivíduo se apega a uma crença e nega qualquer evidência que a confronte, permanecendo em seu estado de conforto; o método dogmático, quando uma instituição passa a ter o poder de determinar o que é verdade e justificar a crença; e o método a priori, quando o indivíduo assume como verdadeiro um sistema de proposições universais e passa a aceitar apenas os fatos que confirmam essas proposições. Esses métodos, combinados ou não, são a base das estratégias de produção e distribuição de fake news e deepfakes (ROMANINI; OHLSON, 2018; GUARDA; OHLSON; ROMANINI, 2018; BORGES; GAMBARATO, 2019). O último método é chamado científico e, segundo Peirce, nele nossas crenças são formadas pela aceitação de algo externo, não sendo influenciadas por nossas próprias fantasias, mas sim por eventos externos mediados por signos confiáveis. Uma crença comum baseada na força de eventos externos e compartilhada por muitos pode ser chamada de chamada de conclusão (BORGES; GAMBARATO, 2019).

É sob esse aspecto que a produção *In event of moon disaster* pode ser utilizada como um caso exemplar para o desenvolvimento de contra crenças no combate às deepfakes, em razão do didatismo com que nos apresenta a facção de uma deepfake e dos efeitos de credulidade que elas estão aptas a produzir. Cabe muito bem aqui a sugestão de McLuhan (1964, p. 66 *apud* BIGGIO; BUSTAMANTE, 2021, p. 161):

> A capacidade do artista de se desviar do golpe violento da nova tecnologia em qualquer época, e impedir tal violência com plena consciência, é antiquíssimo [...]. O artista pode corrigir as relações dos sentidos antes que o golpe da nova tecnologia tenha entorpecido os procedimentos conscientes. Ele pode corrigi-los antes que o entorpecimento e a tentativa subliminar e a reação comecem.

Ao alertar sobre os perigos das novas tecnologias usadas na produção de mídias sintéticas, os autores contribuem para o "enriquecimento dos interpretantes que podem ser gerados a partir do cotejo cuidadoso das relações entre o signo e aquilo a que ele se refere" (SANTAELLA, 2020, p. 22).

Considerações finais

O uso de notícias falsas como fonte de desinformação não é um fenômeno inédito na história. O que está em rápida transformação são os usos sem precedentes de dados, algoritmos e uma infraestrutura de comunicação global com capacidade transformar qualquer pessoa em produtor de conteúdo com potencial de atingir milhões de pessoas em pouco tempo.

A democratização dos meios para a produção de conteúdo tirou dos antigos conglomerados de mídia de massa a primazia de fonte de informação e a transferiu para as redes sociais. Longe dos métodos tradicionais de certificação de produção de conteúdo, impulsionados pelos baixos custos de entrada e possibilidade rápida de rentabilizar o conteúdo por meio da receita de propaganda, esses espaços tornaram-se prolíficos para a proliferação de conteúdo falso.

Impulsionados pela lógica do compartilhamento, dos bancos de dados com registros de toda interação on-line dos usuários e algoritmos capazes de entregar informação extremamente personalizada aos usuários, que cada vez mais tem nas redes e buscadores sua fonte principal de informação, as notícias falsas se alastraram e mostraram-se capazes de decidir os destinos das sociedades.

Um passo significativo foi dado em 2017 com o surgimento das deepfakes, que ao se apoiarem na imensa quantidade de imagens e vídeos disponíveis na rede e usando o compartilhamento dos programas de código aberto, colocaram a poucos cliques de distância de qualquer pessoa com um computador a possibilidade da criação de vídeos, que vão desde a sátira e paródia com a finalidade de rir, passando pelo pornô de vingança, *cyberbullying* e manipulação da opinião pública.

No entanto a maior ameaça das deepfakes está, além da sua capacidade de colocar pessoas em lugares e situações nas quais elas nunca estiveram ou em dizer coisas que elas nunca disseram ou ainda sua difícil detecção, a maior ameaça reside na possibilidade de levar as pessoas a simplesmente duvidar de tudo o que veem, a constante exposição a conteúdos falsos pode tornar as sociedades incapazes de compartilhar e cooperar com base em observações básicas da realidade. O combate a essa ameaça é urgente e, até o momento, o que os estudiosos têm encontrado é a possibilidade de entramos em uma guerra sem fim, pois, a cada nova geração de ferramentas para combater as deepfakes, as bases para que elas sigam evoluindo já estariam lançadas.

A educação dos cidadãos sobre essa ameaça e o desenvolvimento de um senso crítico sobre as notícias que circulam na mídia podem funcionar como antídotos. A produção "*In the event of moon disaster*", ao recriar um hipotético discurso proferido pelo ex-presidente americano Richard Nixon, procura alertar os incautos para as possibilidades dessa tecnologia. Sendo uma questão que remete necessariamente ao que deveríamos entender por realidade e verdade, encontramos na semiótica de Peirce um arcabouço capaz de acompanhar a leitura dos impactos desse fenômeno. Ainda que este artigo não tenha investigado o contexto cultural, econômico e político, a saber, o contexto dos objetos dos signos deepfakes, como elementos que afetam a fixação das crenças, o pragmatismo de Peirce nos aponta um caminho promissor para conhecer as estratégias que os produtores desse tipo de conteúdo empregam e, consequentemente, formas de combatê-lo, especialmente por meio de alertas importantes para os impactos causados pelo consumo e compartilhamento de notícias e vídeos sem a análise crítica do que está por trás deles. E o que está por trás é sempre mais alarmante do que podemos imaginar. Há, portanto, que enfrentá-lo, cara a cara.

Referências

ALLCOTT, Hunt; GENTZKOW, Matthew. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, v. 31, n. 2, p. 211-236, 2017.

ASHISH, Jaiman. AI generated synthetic media, aka deepfakes. *Towards Data Science*, 9 ago., 2020. Disponível em: towardsdatascience.com/ai-generated-synthetic-media-aka-*deepfakes*-7c021dea40e1. Acesso em: 13 jul. 2021.

_____. Positive use cases of deepfakes. *Towards Data Science*, 14 ago. 2020. Disponível em: towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387. Acesso em: 13 jul. 2021.

BATTAGLIA, Rafael. Afinal, o que são deepfakes? *Super Interessante*, 29 out. 2020. Disponível em: super.abril.com.br/tecnologia/afinal-o-que-sao-*deepfakes*. Acesso em: 15 jul. 2021.

BERDUYGINA, Oksana N.; VLADIMIROVA, Tatyana N.; CHERNYAEVA, Elena V. Trends in the spread of fake news in mass media. *Media Watch*, v. 10, n. 1, p. 122-132, 2019.

BIGGIO, Federico; BUSTAMANTE, Victoria Vanessa dos Santos. Elusive masks: A semiotic approach of contemporary acts of masking. *Lexia: Revista de Semiótica*, v. 37-38, p. 141-164, 2021.

BORGES, Priscila Monteiro; GAMBARATO, Renira Rampazzo. The role of beliefs and behavior on Facebook: A semiotic approach to algorithms, fake news, and transmedia journalism. *International Journal of Communication*, v. 13, p. 603-618, 2019.

BOTHA, Johnny; PIETERSE, Heloise. Fake news and deepfakes: A dangerous threat for 21st century information security. *Anais* ICCWS 2020 15th International Conference on Cyber Warfare and Security. Academic Conferences and publishing limited, 2020, p. 57.

BROWN, Mike. Why 'fake news' won Collins dictionary's word of the year. *Inverse*, 11 fev. 2017. Disponível em: inverse.com/article/38041-donald-trump-*fake-news*-word-of-the-year. Acesso em: 15 jul. 2021.

CHESNEY, Bobby; CITRON, Danielle. Deep *fakes*: A looming challenge for privacy, democracy, and national security. *California Law Review*, v. 107, n. 6, p. 1753-1820, 2019.

CITTON, Yves. Could deep *fake*s uncover the deeper truth of an ontology of the networked images? *The Nordic Journal of Aesthetics*, v. 30, n. 61-62, p. 46-64, 2021.

DAVE, Johnson. What is a deep*fake*? Everything you need to know about the AI-powered *fake* media. *Insider*, 22 jan. 2021. Disponível em: businessinsider.com/what-is-deep*fake*. Acesso em: 13 jul. 2021.

FERRAREZI, Fernanda; BORGES, Priscila. O que é e o que parece ser: Imagens criadas por inteligência artificial como elementos atuantes na pós-verdade. *Anais*... 43º Congresso Brasileiro de Ciências da Comunicação Virtual. INTERCOM – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação e Universidade Federal da Bahia, 1-10 dez. 2020, p. 388-416.

FIGUEIRA, Álvaro; OLIVEIRA, Luciana. The current state of fake news: Challenges and opportunities. *Procedia Computer Science*, v. 121, p. 817-825, 2017.

FLETCHER, John. Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance. *Theatre Journal*, v. 70, n. 4, p. 455-471, 2018.

FLOOD, Alison. Fake news is 'very real' word of the year for 2017. *The Guadian*, 02 nov. 2017. Disponível em: theguardian.com/books/2017/nov/02/*fake-news*-is-very-real-word-of-the-year-for-2017. Acesso em: 15 jul. 2021.

GUARDA, Rebeka F.; OHLSON, Marcia P.; ROMANINI, Anderson V. Disinformation, dystopia and post-reality in social media: A semiotic-cognitive perspective. *Education for Information*, v. 34, n. 3, p. 185-197, 2018.

HAO, Karen. Inside the strange new world of being a deepfake actor. *MIT Technology Review*, 09 out. 2020. Disponível em: technologyreview.com/2020/10/09/1009850/ai-deepfake-acting/. Acesso em: 15 jul. 2021.

KORSHUNOV, Pavel; MARCEL, Sébastien. Deepfakes: A new threat to face recognition? Assessment and detection. *Idiap (=Istituto Dalle Molle di Intelligenza Artificiale Percettiva) Research Report* 18-2018, Martigny, 2018. Disponível em: publications.idiap.ch/downloads/reports/2018/Korshunov_Idiap-RR-18-2018.pdf. Acesso em: 15 jul. 2021.

MIRSKY, Yisroel; LEE, Wenke. The creation and detection of deepfakes. *ACM Computing Surveys*, v. 54, n. 1, p.1-41, 2021.

MOONDISASTER. Resources. Disponível em: moondisaster.org/about. Acesso em: 15 jul. 2021.

NÖTH, Winfried. Reconstruções semióticas da realidade: Reflexões sobre a realidade puramente objetiva de John Deely. *TECCOGS: Revista de Tecnologias Cognitivas*, n. 13, p. 132-140, 2016.

PEIRCE, Charles S. Fraser's "The Works of George Berkeley". *North American Review*, v. 113, p. 449-472, out. 1871.

ROMANINI, Anderson Vinicius; OHLSON, Márcia Pinheiro. De elos bem fechados: O pragmatismo e a semiótica peirceana como fundamentos para a tecnologia blockchain utilizada no combate às fake news. *Revista Communicare*, São Paulo, v. 18, n. 2, p. 60-73, 2018.

SANTAELLA, Lucia. *A pós-verdade é verdadeira ou falsa?* (Coleção Interrogações). Barueri: Estação das Letras e Cores, 2018.

_____. A semiótica das fake news. *Verbum – Cadernos de pós-graduação*, v. 9, n. 2, p. 9-25, 2020.

WAAL, Cornelis de. *Peirce*: A Guide for the Perplexed. London: Bloomsbury, 2013.

WESTERLUND, Mika. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, v. 9, n. 11, p. 39-52, 2019.

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.

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.

DE PAULA ASSIS JR., Fabio; HESSEL, Ana Maria Di Grado. Entre ver e crer: deepfake e criação para arte e enetretenimento. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 79-89.

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

dx.doi.org/10.23925/1984-3585.2021i23p79-89

Entre ver e crer:

deepfakes e criação para arte e entretenimento

Fabio de Paula Assis Junior[1]

Ana Maria Di Grado Hessel[2]

Resumo: Ao se analisar a deepfake sob a ótica histórica da produção artística, encontram-se paralelos no passado que podem ajudar a solucionar alguns dos desafios incitados pela nova tecnologia. Afinal, esta não é a primeira vez em que a sociedade se questiona se ver é, de fato, acreditar. A nascente experiência de uso de obras consagradas pela deepfake suscita o potencial de sua exploração no mundo da cultura e do entretenimento, sobretudo quando a própria criação em diferentes formatos fizer um uso planejado dessa tecnologia para fins de viés afirmativo como a criação de novas linguagens e obras. Em uma era marcada pela desinformação, o desafio é garantir a literacia do público acerca da veracidade daquilo que lê, vê e consome. Em vez de se combater o uso das deepfakes, o esforço deve caminhar no sentido de se promover seu enorme potencial positivo sobretudo no mundo das artes e da cultura.

Palavras-chave: Deepfake. Arte. Cultura. Entretenimento. Web.

1 Fabio de Paula é arquiteto, jornalista e professor da pós-graduação em neurobusiness na Fundação Getúlio Vargas, bem como das graduações em jornalismo e publicidade e propaganda na Faculdade Cásper Líbero, São Paulo. É mestre e doutorando em Tecnologias da Inteligência e Design Digital pela PUC-SP, graduado em jornalismo pela Cásper Líbero, e em arquitetura e urbanismo pela FAU/USP. ORCID: orcid.org/0000-0001-6090-7091. CV-Lattes: lattes.cnpq.br/6630094676180781. E-mail: fabiodepaula77@gmail.com.

2 Doutora e Mestre em Educação. Graduada em Pedagogia na PUC-SP, com especialização em Informática pela UFPA. É Professora do Departamento de Educação: Formação Docente, Gestão e Tecnologias; é pesquisadora e professora credenciada no Programa de Estudos Pós-Graduados em Tecnologias da Inteligência e Design Digital – TIDD/PUC-SP. É pesquisadora do GEPI, GEPEC e GPTED. ORCID: orcid.org/0000-0003-4776-7754. CV Lattes: lattes.cnpq.br/2150241303883701. E-mail: digrado@uol.com.br.

Between seeing and believing: deepfakes and creation for art and entertainment

Abstract: When analyzing deepfake from the historical perspective of artistic production, it is possible to find parallels that help to understand some of the challenges prompted by this new technology. This is not the first time the society has questioned whether seeing is, in fact, believing. The nascent experience of using deepfake raises the potential for its exploitation in culture and entertainment, especially when the creation itself in different formats makes a planned use of this technology for affirmative purposes such as creation of new languages. In an era marked by misinformation, the challenge is to ensure the public's literacy about the veracity of what they read, see and consume. Instead of fighting the use of deepfakes, the effort must move towards promoting their enormous positive potential, especially in arts and culture.

Keywords: Deepfake. Art. Culture. Entertainment. Web.

A expressão deepfake entrou há alguns anos no repertório da web e na cultura popular, suscitando preocupação e provocando as empresas digitais a se prepararem para um futuro com áudios e vídeos forjados ao sabor da criatividade humana e da capacidade do aprendizado de máquina. Mas, enquanto essa tecnologia gera apreensão por conta da imediata associação com o universo contemporâneo da desinformação, dado que qualquer conteúdo viraliza-se na Internet de modo instantâneo e caótico, a deepfake também pode gerar impactos afirmativos e impulsionar experimentos positivos, especialmente no mercado da arte e da cultura.

Trata-se de uma tecnologia popularizada recentemente que depende de codificadores automáticos e outras técnicas de aprendizado de máquina para que uma mídia de imagem, de vídeo ou de áudio preexistente seja manipulada, usando redes neurais artificiais, de modo a substituir, via semelhança, a pessoa, o objeto ou a voz.

Apesar de soarem como reais, deepfakes são criadas a partir de tecnologias de *deep learning*, que realizam a edição de pixels ou sons e, então, forjam até mesmo a criação de rostos em filmes e de vozes em gravações preexistentes. A alteração realística e em tempo real pode criar vídeos fantásticos a um baixo custo, mas também provocar danos incríveis.

Trata-se de uma expressão majoritariamente usada para descrever qualquer conteúdo de vídeo que aparenta ser realista, e, na verdade, é falsificado. Mesmo sendo recente, seu uso não está mais restrito ao universo dos cientistas da computação, uma vez que rapidamente surgiram centenas de aplicativos que empregam Inteligência Artificial (IA) para tornar o uso da deepfake acessível para o consumidor. Alguns desses *apps*, inclusive, são usados como ferramentas de captura de dados dos usuários para sua venda não autorizada.

Não tardou, portanto, que a tecnologia fosse descrita por uma infinidade de notícias que apontam suas implicações mais nefastas, como o potencial de uso na produção de propaganda falsa, tanto para a criação de perfis falsos em redes sociais e até em campanhas de difamação. Apesar de a maioria das notícias concentrar-se em elencar os aspectos negativos da deepfake, como toda tecnologia, ela também traz benefícios. No campo do entretenimento e da cultura, por exemplo, o método pode ser usado para produzir arte, engajar o público e promover experiências que nunca haviam sido realizadas anteriormente.

Tecnologias frequentemente desempenham papeis positivos na sociedade, mesmo que cada geração demore a reconhecer como ela transforma a cultura. A fotografia, por exemplo, alterou o modo como entendemos a pintura e as artes plásticas como um todo. O mesmo vale para as deepfakes, que agora alteram a forma como entendemos e usamos vídeos e áudios. Ao se analisar a deepfake sob a ótica da produção artística e histórica, encontram-se paralelos no passado que podem ajudar a solucionar alguns dos desafios incitados pela nova tecnologia. Afinal, esta não é a primeira vez na História em que a sociedade se questiona se ver é, de fato, acreditar.

O crítico cultural alemão Walter Benjamin, por exemplo, escreveu sobre as transformações da arte e da cultura causadas pelos avanços tecnológicos de sua época, no início do século XX. No texto "A obra de arte na era de sua reprodutibilidade técnica", de 1935, Benjamin afirma que a arte é sempre um constructo entre a sociedade e as tecnologias, e que a influência dos avanços tecnológicos na produção artística gera complexidades que a sociedade tem a missão de compreender e aceitar. Para o crítico alemão, estudiosos e curadores de arte desempenham um papel fundamental nesse processo de aceitação, porque são eles que contextualizam as novas tecnologias dentro da ética artística. Dessa forma, sem a confiança nas instituições governamentais e privadas que fazem a gestão da cultura, esse processo não será bem-sucedido.

Se a lógica de Benjamin for aplicada ao debate sobre as deepfakes, pode-se abrir grupos de possibilidades criativas que funcionam como salvaguarda à tecnologia, cuja reação tem sido negativa principalmente por causa de seu potencial de colaborar para a desinformação midiática. Mas a inovação das deepfakes é inegável, não somente para replicar obras de arte ou ressuscitar artistas mortos, mas também para revolucionar a produção em áreas da arte que não somente a pintura ou a escultura.

No cinema, há exemplos consagrados. No final de 2020, o canal do YouTube chamado Shamook usou a deepfake em cenas envolvendo personagens famosos da saga cinematográfica de ficção científica Star Wars. O vídeo "Deepfaking Tarkin & Leia in Rogue One: A Star Wars Story" (DEEPFAKING, 2020) apresenta uma cena ultrarrealista e aparentemente atual da personagem Princesa Leia (a atriz Carrie Fisher que a interpreta, porém, faleceu em 2016) interagindo com o Governador Tarkin (o ator Peter Cushing, que deu vida em 1977 ao personagem no filme Guerra nas Estrelas, faleceu em 1994).

Já o longa-metragem de fantasia sobre super-heróis "Homem-Aranha: Longe de Casa", de 2019, embora não tenha sido concebido com deepfakes, aborda diretamente o assunto em seu enredo e teve um popular trailer realizado com a tecnologia. Mas, mesmo os roteiristas do filme alertam para o fato de que a tecnologia não pode ser tratada sem seriedade ou como algo exclusivo do mundo da ficção. Um dos roteiristas do filme, Erik Sommers, fez o seguinte alerta em entrevista de 2019 para o jornal Los Angeles Daily News: "Muitos problemas surgem e são encenados no entretenimento pop, mas continuam sendo problemas reais. Somente porque são algo divertido quanto um filme do Homem-Aranha, não significa que não sejam um problema real. Conforme a tecnologia e seu uso se tornam mais prevalentes, aumenta também sua capacidade de manipular e enganar as pessoas" (STAUSS, 2019).

Para o coautor do filme, Chris McKenna, que também falou ao jornal californiano, "a manipulação da realidade ocorre do começo ao fim do filme. Discutimos deepfakes em um determinado ponto, discutimos como agora é fácil alterar digitalmente a percepção das pessoas sobre a realidade". Ele complementa a entrevista com as seguintes questões: "o que é real, o que é falso, o que alimenta seu viés de confirmação, o que alimenta qual narrativa você quer ou precisa contar?"

Para além da questão do simulacro, a deepfake está sendo utilizada para subverter – no sentido mais amplo dessa palavra – obras de arte seminais. Não se trata de profanação ou adulteração, mas da busca e da construção de novos significados para trabalhos artísticos cujo valor já é consolidado.

A "Mona Lisa" (c. 1503-1506), por exemplo, foi objeto de um estudo realizado pelo Samsung AI Center de Moscou e do Instituto de Ciência e Tecnologia de Skolkovo, e publicado no artigo "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models" (ZAKHAROV; SHYSHEYA; BURKOV; LEMPITSKY, 2019) em que os pesquisadores tiraram fotos da famosa pintura de Leonardo da Vinci e as manipularam com fala e movimento, como se ela fosse uma pessoa real. Além da animação em si, o processo envolveu o uso de um algoritmo capaz de identificar os elementos faciais da obra para dar vida à mulher ali retratada. A tecnologia também foi utilizada para animar outras obras famosas como "Garota com brinco de pérola" (c. 1665), de Johannes Vermeer, "Retrato de mulher desconhecida" (1883), de Ivan Kramskoy, e o "Retrato de Johanna Staude" (1917-1918), de Gustav Klimt.

A nascente experiência de uso de obras consagradas pela deepfake suscita o potencial de sua exploração no mundo da cultura e do entretenimento, sobretudo quando a própria criação das obras de arte em diferentes formatos fizer um uso planejado e consciente dessa tecnologia. Aliás, a competição pelo engajamento e atenção do consumidor – e, portanto, do dinheiro - não é algo restrito ao mundo das notícias, dos shopping centers ou dos supermercados. No cosmo dos museus e centros culturais, a concorrência também está presente. Nesse sentido, o uso da deepfake também encontra espaço para utilização, uma vez que pode ser usada como tecnologia para criar experiências inesperadas e, ao mesmo tempo, customizadas para os usuários, como ter uma conversa com um artista ou ser saudado por uma figura ilustre do passado ao adentrar em uma galeria.

No Dalí Museum de Saint Petersburg, Estados Unidos, os visitantes puderam conferir em 2019, na exposição "Dalí lives (via artificial intelligence)" (DALÍ MUSEUM, 2019) uma experiência conduzida por IA em que o artista espanhol foi recriado em tamanho real. A partir daí, recebia os visitantes desde sua chegada e contava sobre sua vida, bem como das obras ali expostas. A experiência talvez tenha sido a primeira a envolver o uso da deepfake com um propósito artístico, institucional e educativo, mas certamente não será a última, e também revela o enorme potencial dessa tecnologia para atrair público e engajar seu interesse. Os visitantes puderam ainda ter a experiência de tirar *selfies* com uma representação em tamanho real, feita via deepfake, do próprio artista. A técnica foi capaz graças ao uso de algoritmos de IA, com milhares de fotos do rosto de Dali, que permitiram a criação do simulacro do mestre da pintura surrealista.

As deepfakes também permitem que artistas e criativos produzam novas formas de expressão. Nos últimos anos, por exemplo, a tecnologia não apenas se tornou uma ferramenta de envolvimento com a arte e os artistas do passado, mas também para criar formas de expressão artística, desde o uso em perfis de redes sociais dedicados ao entretenimento, como o do jornalista e "deepfaker" Brunno Sartori, com 474 mil seguidores no Instagram (BRUNNOSARTTORI), até obras criadas por artistas internacionais consagrados como Hao Li, que conduz uma startup chamada *Pinscreen* cujo foco são os negócios de entretenimento e arte baseados na deepfake (PINCSREEN, 2020).

Enquanto Li aposta na tecnologia para promover negócios, Sartori une-se a um time cada vez mais numeroso em escala global que aposta na deepfake como ferramenta política de provocação social e transformação. Para isso, se no Brasil predominam os vídeos de Instagram que combi-

nam vídeos de humor com os rostos de políticos e outras figuras públicas, no exterior proliferam as exposições e obras de arte como a realizada em 2019 por Bill Posters e Daniel Howe que criaram, com deepfake, um vídeo falso e conceitual de Mark Zuckerberg (POSTERS, 2019) para a exposição "Alternate Realities" da Site Gallery X, em Sheffield (SITE GALLERY X, 2019). O vídeo também postado em 2019 no Instagram (rede que pertence a Zuckerberg) descreve o poder sinistro do Facebook sobre seus usuários. O discurso revela o modelo de negócios baseado em *big data* e na exploração dos dados de seus usuários, e parece ter mais impacto justamente por parecer vir diretamente de seu fundador.

Enquanto o Homem-Aranha luta contra a deepfake em seu novo filme e o público das redes sociais parece adorar os vídeos de YouTube e Instagram com o rosto de Jair Bolsonaro transfigurado para algum personagem em uma cena cômica de novela, o desafio das deepfakes continua sendo enorme. Afinal, a tecnologia pode colocar em vídeos o rosto de alguém no corpo de outras pessoas, manipular o áudio das palavras que uma pessoa realmente falou em algo que ela não disse, ou seja, a deepfake é perigosa, seja você uma celebridade, um político ou uma pessoa comum.

O risco é real, tanto assim que famosos como a atriz Scarlett Johansson, cansada de ver seu rosto substituir digitalmente os de artistas nus em filmes adultos, declarou em 2018 ao jornal *Washington Post* que, por enquanto, tentar lutar contra o fenômeno é inútil (HARWELL, 2018). Mas se os artistas do mercado de entretenimento talvez precisem aceitar a deepfake como um preço a se pagar para fazer negócios, uma pessoa que não é famosa pode sofrer consequências e ter sua vida arruinada pela mesma prática, por exemplo, com sua imagem sendo usada em um vídeo falso de *revenge porn*.

Na era da desinformação, o desafio é garantir a consciência dos consumidores acerca da veracidade daquilo que leem, veem e consomem. Não é diferente com a deepfake, cujo uso indevido merece – e já possui – um enorme esforço nas ciências da computação para ajudar os usuários na detecção de mídia falsa ou alterada. Em vez de se combater o uso das deepfakes, o esforço deve caminhar no sentido de se promover a literacia a seu respeito, bem como provocar seu enorme potencial positivo sobretudo no mundo das artes e da cultura.

Há esforços significativos para proteger a sociedade contra usos negativos e abusivos da deepfake. Ao passo em que os algoritmos de detecção sofisticam, os mecanismos de produção também evoluem. Quando se trata de arte e cultura, a deepfake pode ser usada por artistas, museus

e organizações para atrair o público, expandindo as possibilidades de sua utilização para além de exposições, permitindo até mesmo que as pinturas e retratos em galerias museus ganhem vida ou que a tecnologia, combinada com *chatbots* alimentados por IA, forneçam possíveis interações – na ficção ou na realidade – com atores e até mesmo figuras influentes que já morreram.

Não faltam exemplos de usos benéficos da deepfake: a tecnologia pode ser usada para arrecadação de fundos, por exemplo, com o envio de vídeos personalizados com o simulacro de um artista solicitando contribuições a um museu. As deepfakes, portanto, emergem como uma ferramenta essencial de marketing, seja artístico e cultural, mas também em qualquer ação voltada ao aprimoramento da sociedade, desde que acompanhado de ações educativas e afirmativas voltadas à comunidade e às empresas, mas também à formulação de legislação específica.

O maior desafio ético na aceitação das deepfakes é, assim, de autoridade, direito que necessariamente inclui debates mediados pelo Estado e que envolve reconhecer e lidar com o fato de que há uma a conexão entre a emergente tecnologia com a desinformação em curso. Dessa forma, o primeiro passo para catalisar o potencial positivo das deepfakes e da web em si é adotar salvaguardas para que se restaurem a segurança e a confiança das informações disponíveis na Internet.

Mesmo que o Estado deva participar desse processo, trata-se de uma confiança que deve ser conquistada pelo indivíduo e pela coletividade, e, nesse sentido, o artigo de Walter Benjamin reforça que não se deve temer novas formas de arte aumentadas pela tecnologia, desde que haja um entendimento comum de que prevalecerá a boa governança. No caso da arte e do entretenimento, essa autoridade reside sobretudo nos museus, produtoras, historiadores, curadores e artistas.

Foi assim com as primeiras gravações de áudio e com o surgimento da fotografia. O mesmo aconteceu com a popularização da tecnologia computacional a partir da década de 1980, que permitiu o surgimento e uso massivo de softwares de retoques de sons e imagens antes possível de serem feitos somente em grandes agências especializadas. À medida que todas essas tecnologias foram se consagrando no mercado, os consumidores passaram a aceitar que nem tudo que ouvem ou veem é necessariamente a realidade. Com as deepfakes não será diferente: conforme as ferramentas de manipulação de vídeo e áudio forem incorporadas em um uso generalizado, elas não serão entendidas como um *bug* ou como uma ameaça, mas como um recurso. Mais ainda, os próprios fabricantes

de hardware devem estar ansiosos para criar essa demanda e investir em pesquisas que tanto promova seu uso quanto afirme a autoridade sobre ela, afinal as deepfakes têm grande potencial para vender muitos computadores e aplicativos.

A tecnologia, tampouco, é imparável. Há, inclusive, ferramentas também baseadas em IA que conseguem identificar e sinalizar vídeos deepfake, mas elas não são amplamente conhecidas e utilizadas. A Microsoft já desenvolveu uma ferramenta de IA para identificar deepfakes, o software *Video Authenticator*, porém ele não é suficiente para prevenir todos os possíveis riscos causados pela deepfake. Há outras ferramentas computacionais disponíveis, mas muito difíceis de serem usadas pela maioria das pessoas, que, por enquanto, vão continuar dependendo de evidências de falsidade nos vídeos a que assistem na web, sobretudo em um mundo onde cada vez mais sites questionáveis propagam conteúdo falso como se fossem notícias autênticas, muitas vezes sendo financiados por empresas, instituições e até governos.

Na política, o temor é que, em uma mídia inundada pela desinformação, a deepfake seja usada para se criar vídeos falsos de candidatos dizendo ou fazendo algo que não fizeram e divulgá-los no decorrer do ciclo eleitoral. A legislação sobre o assunto, porém, ainda é nascente. No estado norte-americano da Virgínia, o legislativo incluiu em 2019 a proibição de uso de imagens falsas em qualquer publicação digital à emenda que rege o *revenge porn* – que pode incluir ou utilizar deepfakes, ainda que não se restrinja a eles.

O desafio, porém, reside no fato de que, uma vez que algo é publicado na web, torna-se quase impossível de ser removido. Na Califórnia, por sua vez, já há legislação, o California Senate Bill 564, que procura evitar abusos com o uso de deepfake que podem impactar as carreiras de atores, mas também de apresentadores e repórteres de televisão que também podem ter seus rostos expostos em vídeos públicos. Na prática, a legislação permitirá que pessoas famosas possam negociar, em seus contratos, uma proteção contra o uso de deepfake com a sua imagem.

Essa lei, no entanto, além de restrita a um estado norte-americano, não se estende a pessoas comuns cuja imagem também pode ser registrada em fotos e vídeos anônimos. Portanto, além do nível contratual, é importante que a questão seja debatida na esfera educacional, mas sobretudo ampliada em termos de legislação. A solução do problema exige desde o desenvolvimento contínuo de melhores tecnologias de detecção de deepfake até de políticas nas grandes plataformas de mídia social, uma

legislação específica sobre o assunto, mas sobretudo um público mais bem informado. Cada um de nós precisa estar cada vez mais ciente de nossos próprios vieses de confirmação, e lembrar em qualquer experiência de consumo de arte e entretenimento que, há mais de um século, sabemos que imagens filmadas podem ser manipuladas. E a deepfake é apenas mais uma das tecnologias que, nesse intervalo de tempo, convencionou-se chamar de efeitos especiais.

Referências

BENJAMIN, Walter. *A obra de arte na era de sua reprodutibilidade técnica*. Tradução: Gabriel Valladão Silva, 12. ed. Porto Alegre: L&PM, 2018.

BRUNNOSARTTORI (Site). Disponível em: instagram.com/brunnosarttori e brunnosarttori.contactin.bio. Acesso em: 30 jun. 2021.

DALÍ MUSEUM. Behind the scenes: Dali Lives, St. Petersburg, FL, 11/05/2019. Disponível em: thedali.org/exhibit/dali-lives. Acesso em: 30 jun. 2021.

DEEPFAKING Tarkin & Leia in Rogue One: A Star Wars Story. *Paperspace* 08/12/2020. Disponível em: youtube.com/watch?v=_CXMb_MO3aw&ab_channel=Shamook. Acesso em: 30 jun. 2021.

HARWELL, Drew. Scarlett Johansson on fake AI-generated sex videos: "Nothing can stop someone from cutting and pasting my image". *Washington Post*, 31/12/2018. Disponível em: washingtonpost.com/technology/2018/12/31/scarlett-johansson-fake-ai-generated-sex-videos-nothing-can-stop-someone-cutting-pasting-my-image. Acesso em: 30 jun. 2021.

PINSCREEN, Los Angeles, 2020 (Site). Disponível em: pinscreen.com. Acesso em: 30 de jun. 2021.

POSTERS, Bill. Zuckerberg: We're increasing transparency on ads. 07/06/2019. Disponível em: instagram.com/p/ByaVigGFP2U e instagram.com/bill_posters_uk. Acesso em: 30 jun. 2021.

SITE GALLERY X Sheffield Doc/Fest: Alternate Realities – Subconscious Sensibilities. Sheffield, 2019. Disponível em: sitegallery.org/exhibition/site-gallery-x-sheffield-doc-fest-alternative-realities. Acesso em: 30 jun. 2021.

STRAUSS, Bob. From Hollywood to Washington to 'Far from home,' deepfakes are a growing concern. *Los Angeles Daily News*, 22/07/2019. Disponível em: dailynews.com/2019/07/22/from-hollywood-to-washington-to-far-from-home-deepfakes-are-a-growing-concern. Acesso em: 30 jun. 2021.

ZAKHAROV, Egor; SHYSHEYA, Aliaksandra; BURKOV, Egor; LEMPITSKY, Victor. *Few-shot adversarial learning of realistic neural talking head models*. Cornell University, 2019. Disponível em: arxiv.org/abs/1905.08233. Acesso em: 30 jun. 2021.

SANTAELLA, Lucia; SALGADO, Marcelo de Mattos. Deepfake e as consequências sociais da mecanização da desconfiança. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 90-103.

Recebido em: 25 jan. 2021
Aprovado em: 22 fev. 2021

dx.doi.org/10.23925/1984-3585.2021i23p90-103

Deepfake e as consequências sociais da mecanização da desconfiança

Lucia Santaella[1]

Marcelo de Mattos Salgado[2]

Resumo: O artigo tem origem em trabalhos anteriores que analisaram a crise de confiança na sociedade e como esta seria paradoxalmente impulsionada, em parte, por avanços como o *blockchain*, tecnologia de segurança baseada em algoritmos de Inteligência Artificial. Ao mecanizar a confiança, o *blockchain* paradoxalmente reduz sua importância ou até torna a mesma desnecessária como elemento dos laços entre duas partes humanas. Tal inovação contribui para mudar gradualmente a dinâmica do sistema social da confiança interpessoal para uma sociedade do controle, ao substituir a confiança entre indivíduos por tecnologias de segurança e concentração de dados nas mãos de poucos – em particular, as big techs e burocracias estatais. Já a deepfake é uma manipulação audiovisual que também se baseia na contemporânea tecnologia algorítmica do *deep learning* e do *machine learning*. No entanto, o objetivo da deepfake é criar representações e percepções alternativas – e propositadamente falsas – da realidade, efetivamente mecanizando a desconfiança entre seres humanos. Este texto procura avaliar, desde o trabalho anterior, as consequências do deepfake para as relações humanas e a sociedade no contexto da crise de confiança.

Palavras-chave: Deepfake. Inteligência Artificial. Confiança. Controle. Redes digitais.

1 Lucia Santaella é pesquisadora 1A do CNPq, professora titular da PUC-SP. Publicou 51 livros e organizou 24, além da publicação de mais de 400 artigos no Brasil e no exterior. Recebeu os prêmios Jabuti (2002, 2009, 2011 e 2014), o prêmio Sergio Motta (2005) e o prêmio Luiz Beltrão (2010). ORCID: orcid.org/0000-0002-0681-6073. CV Lattes: lattes.cnpq.br/7427854657719431. E-mail: lbraga@pucsp.br.

2 Marcelo de Mattos Salgado é doutorando na PUC-SP (TIDD), jornalista e professor. Estuda o aumento da polarização nas redes digitais desde 2016. ORCID: orcid.org/0000-0001-8243-4977. CV Lattes: lattes.cnpq.br/9529682415224917. E-mail: msalgadosp@gmail.com.

Deepfake and the social consequences of the mechanization of distrust

Abstract: The article originates from previous works that analyzed the crisis of trust in society and how this would be driven, in part, by advances such as blockchain, a security technology based on Artificial Intelligence algorithms. By mechanizing trust, blockchain paradoxically reduces its importance or even makes it unnecessary as an element of the bonds between two human parts. Such innovation contributes to gradually change the dynamics of the social system from interpersonal trust to a society of control, replacing trust between individuals with security technologies and data concentration in the hands of a few – in particular, big techs and state bureaucracies. Deepfake is an audiovisual manipulation that is also based on the algorithmic technology of deep learning and machine learning. However, the purpose of deepfake is to create alternative – and purposefully false – representations and perceptions of reality, effectively mechanizing the distrust between human beings. This text seeks to assess, from the previous work, the consequences of deepfake for human relations and society in the context of the crisis of trust.

Keywords: Deepfake. Artificial Intelligence. Trust. Control. Digital networks.

Introdução

A crise de confiança, tanto de pessoas em instituições (governos, empresas, mídia e ONGs) quanto dos cidadãos entre si ainda vivida por países ocidentais, em geral, foi estabelecida e analisada em trabalho anterior (SALGADO, 2020). Este contexto será brevemente atualizado e estendido à China, sobretudo por conta da pandemia de Covid-19, com início em torno de fevereiro de 2020. A nova praga e as medidas adotadas contra ela continuam a arrasar o planeta de muitas formas – desde as mais evidentes milhões de mortes e questões de saúde física e mental, até questões políticas, econômicas e de controle social. Este quadro produz, também, mudanças severas na confiabilidade entre indivíduos e grupos mundo afora.

Segundo pesquisa da consultoria Edelman, cujo *trust index* (índice de confiança) calcula a média da confiabilidade dos cidadãos em governos, empresas, mídia e ONGs, a variação entre novembro de 2019 e novembro de 2020 na China foi de -10 pontos percentuais – de 82% para 72% – a maior queda do mundo, no período (EDELMAN, 2021, p. 9). O interessante é que, historicamente, a população chinesa tende a confiar (ou a responder que confia) muitíssimo em suas instituições, o que torna o registro desta queda abrupta ainda mais notável. Para uma comparação quanto ao referido índice de confiança em novembro de 2020: os EUA estavam com 48%; Brasil, 51%; Alemanha, 53%; e África do Sul, 48%. A mesma pesquisa indica a gravidade do impacto que a pandemia de Covid-19 teve sobre a fieza das pessoas a respeito de, particularmente, as burocracias governamentais (ibid., p. 5 e 11), com quedas importantes em países europeus e na China. Por outro lado, as empresas surgem como único ator – entre governos, empresas, mídia e ONGs – a reunir, ao mesmo tempo, competência e ética aos olhos das populações analisadas (ibid., p. 6-7).

O interregno da pandemia

A queda da fidúcia na mídia e no jornalismo profissionais mundo afora, que já acontece há vários anos (EDELMAN, 2021; BRENAN, 2020; NÓBREGA, 2020; TWENGE, CAMPBELL; CARTER, 2014), viveu momentos de certa calmaria, com perdas menores e até ganhos de confiança, sobretudo na primeira metade de 2020 (NEWMAN, 2020, p. 10; NEWMAN, 2021, p. 18). Os possíveis motivos: audiência cativa de pessoas com medo, vulneráveis e presas em suas casas por conta de *lockdowns* e medidas similares, a partir da pandemia; e/ou seu retorno voluntário a fontes tradicionais. No entanto, os dados mais tardios de 2020 já trazem, de novo, quedas marcantes de confiabilidade das mídias profissionais (BRENAN, 2020; NOBREGA, 2020). A Edelman também registra esse desgaste continuado: de acordo com sua pesquisa mais recente, em médias globais até os primeiros meses de 2021, 59% dos entrevistados acreditam que "jornalistas e repórteres estão intencionalmente tentando enganar as pessoas dizendo coisas que eles sabem ser falsas ou exageros grosseiros"; 59% pensam que "a maioria das organizações noticiosas estão mais preocupadas com apoiar uma ideologia ou posição política do que informar o público". Por fim, 53% das pessoas dizem confiar em mídias tradicionais em 2021, queda de 8% em relação a 2020, quando eram 61%; e um declínio de 12% em relação a 2019, quando 65% diziam confiar na imprensa (EDELMAN, 2021, p. 24 e 25). A Edelman considera "confiável" um resultado entre 60% e 100%.

Detalhe curioso do mesmo estudo da Edelman: a derrocada da credibilidade em informações obtidas nos mecanismos de pesquisa, tais como Google, responsável por 92,2% (STATCOUNTER, 2021) das buscas feitas em todo o mundo. Em 2017 e 2019, tais instrumentos teriam alcançado seu ponto mais alto de fidúcia: 65%. Agora, em 2021, esta fieza no Google e em sistemas afins despencou para 53%, sua maior queda histórica. Algo mais particular – em vez de informações, notícias – foi obtido pela Reuters em 2019 (NEWMAN, 2019, p. 21), que analisou a baixa confiança em notícias visualizadas a partir de mecanismos de busca (33%); e em redes digitais (23%). A exceção novamente foi, de forma ainda mais específica, a cobertura da pandemia de Covid-19, sobretudo na primeira parte de 2020, quando notícias obtidas em mecanismos de busca obtiveram 45% de confiabilidade e aquelas encontradas em redes digitais, 26% – números que, mesmo assim, são bastante modestos.

Ainda no sentido de compreender a situação de crescente descrença popular em dados obtidos em mecanismos de pesquisa e grandes redes digitais, vale mencionar o livro "Covid-19: o grande *reset*", escrito pelo presidente do Fórum Econômico Mundial, Klaus Schwab, com Thierry

Malleret. O grande *reset* – grosso modo, mudanças sugeridas na gestão global, ou seja, de todos os países, em política, economia e meio ambiente – é proposto por Schwab como um caminho para o mundo seguir desde supostas lições com a pandemia de Covid-19 e em um ambiente internacional de queda generalizada da confiança no contrato social (SCHWAB; MALLERET, 2020, p. 73–74). Em certo momento, os autores refletem sobre a colaboração entre Apple e Google desde abril de 2020 para criar um aplicativo que permita a oficiais de saúde fazer a engenharia reversa dos passos e conexões de pessoas infectadas pelo vírus, rastreando suas vidas – o que provoca em parte considerável do público inquietação e até medo sobre crescente vigilância digital e controle estatal e empresarial (sobretudo, por parte das companhias *big tech*). A partir daí, passa a existir um problema básico que restringe a eficiência de aplicativos como o criado por Apple e Google:

> A pessoa que carrega o aparelho móvel teria que voluntariamente baixar o aplicativo e concordar em compartilhar seus dados, e as duas companhias deixaram claro que sua tecnologia não seria fornecida a agências de saúde pública que não cumpram com suas diretrizes de privacidade. (SCHWAB; MALLERET, 2020, p. 124)

De outro modo: mesmo com a garantia das gigantes digitais – e dos governos que decidem adotar tais aplicativos – que regras de proteção à privacidade (criadas por aqueles atores) seriam seguidas, a adesão do público em geral é ainda muito baixa, o que prejudica a criação de uma base de dados e a pretensa eficiência em controlar a propagação do vírus. A principal razão por trás do receio de cidadãos comuns por todo o mundo em ceder seus dados, certa ou equivocada, é evidente. A rigor, é um dilema já clássico, mas ainda mais perceptível em sua versão atualizada: deve-se confiar em burocracias e empresas e renunciar à própria privacidade para, em teoria, estar mais seguro e protegido de um vírus perigoso?

> Nos próximos meses e anos, o equilíbrio entre benefícios de saúde pública e perda de privacidade será avaliado cuidadosamente, tornando-se o assunto de muitas conversas animadas e debates acalorados. A maioria das pessoas, temerosa do perigo representado pela Covid-19, perguntará: não é tolice não aproveitar o poder da tecnologia para vir em nosso resgate quando somos vítimas de um surto e enfrentamos uma situação de vida ou morte? Eles então estarão dispostos a renunciar a muita privacidade e concordarão que, em tais circunstâncias, o poder público pode legitimamente anular os direitos individuais. Então, quando a crise passar, alguns podem perceber que seu país se transformou repentinamente em um lugar onde não desejam mais viver. (SCHWAB; MALLERET, 2020, p. 127)

Schwab e Malleret citam mesmo os alertas de Yuval Harari e Evgeny Morozov – este último está convencido dos riscos de que a pandemia de Covid-19 pode levar a uma distopia, "um futuro negro de um estado de vigilância tecno-totalitário" (ibid., p. 129). Já Harari acredita ser falsa a escolha entre privacidade e saúde: "Podemos e devemos ter privacidade e saúde. Podemos escolher proteger nossa saúde e interromper a epidemia de coronavírus, não instituindo regimes de vigilância totalitários, mas sim empoderando os cidadãos" (HARARI, 2020). O autor acredita que isto pode ser feito a partir dos exemplos de Taiwan, Coreia do Sul e Cingapura, que "se apoiaram muito mais em testagem extensa, relatos honestos e na cooperação voluntária de um público bem-informado" (ibid.). Caso contrário, Harari sinaliza para o risco, com precedentes históricos, de medidas de controle social durarem mais do que o período mais grave da pandemia.

> Você poderia, é claro, argumentar em favor da vigilância biométrica como uma medida temporária tomada durante um estado de emergência. Ela acabaria uma vez que a emergência acabasse. Mas medidas temporárias têm o péssimo hábito de sobreviver às emergências, especialmente porque sempre há uma nova emergência espreitando no horizonte (HARARI, 2020).

Em outro documento de julho de 2020, o Fórum Econômico Mundial, em parceria com a McGill University, também aborda a ascensão da Internet dos Corpos (IoB), desdobramento da Internet das Coisas (IoT) que implica coletar nossos dados físicos por meio de dispositivos que podem ser implantados, engolidos ou simplesmente usados, gerando enormes quantidades de informações relacionadas à saúde. Algumas dessas soluções, como rastreadores de fitness, são uma extensão da Internet das Coisas, mas agora a Internet dos Corpos passa a ter controle sobre funções vitais do corpo convertendo-o em fonte geradora de dados pessoais. Com isso, levanta-se um conjunto específico de oportunidades e desafios, desde questões de privacidade até questões legais e éticas.

Segundo este *briefing*, "a Internet dos Corpos transformou o corpo humano em uma plataforma de tecnologia, visto que gera enormes quantidades de dados biométricos e comportamentais humanos" (WEF; MCGILL, 2020, p. 7). Essa penetração ainda mais íntima na pessoalidade e na privacidade, entretanto, não evoca confiabilidade por parte de muitos cidadãos comuns – e com algum fundamento.

> Há uma consciência cada vez maior da vulnerabilidade de acessórios vestíveis [*wearables*] e dispositivos médicos ligados à Internet das Coisas a hackers e ataques cibernéticos, que expõem vidas humanas a potenciais danos físicos e riscos de privacidade. [...] Pesquisadores encontraram sérias falhas de segurança em *smartwatches* infantis, que os hackers podem usar para rastrear crianças, obter acesso a áudio e fazer ligações telefônicas para elas. A privacidade é um fator importante que afeta a confiança dos consumidores e a adoção de dispositivos ligados à Internet dos Corpos. (WEF; MCGILL, 2020, p. 10)

Some a este cenário de grave crise da confiança – com especial desconforto a respeito de dados obtidos em meios digitais e concedidos a governos e *big techs* – a inserção de tecnologias com base em algoritmos de Inteligência Artificial (IA). Esta realidade se faz cada vez mais vívida no dia a dia de bilhões de pessoas – seja em uma simples pesquisa no Google, a encomenda de uma refeição no aplicativo iFood ou a busca por um namorado no OkCupid – e acrescenta elementos ainda mais complexos à discussão sobre confiança. Em trabalhos anteriores (SANTAELLA, 2020; SALGADO, 2020), abordamos o blockchain, software de criptografia assimétrica sob a forma de cadeias de código (WILLIAMS, 2019, p. 9). Seu objetivo inicial era garantir segurança nas transações com a moeda digital Bitcoin – no entanto, o blockchain vem se espalhando como medida de segurança muito mais abrangente.

Conforme Salgado (2020), o modo como o blockchain se relaciona à confiança, resumidamente, é o seguinte: a tecnologia criptográfica permite, em uma troca social e/ou econômica qualquer, a substituição da confiabilidade – elemento historicamente constitutivo de laços humanos – por mecanismos de segurança. O blockchain seria, assim, *trustless* ou "aconfiável" (WILLIAMS, p. 15) precisamente porque consiste na mecanização da confiança, o que tornaria a mesma confiabilidade irrelevante em uma troca administrada via *blockchain*: a segurança criptográfica é o que importa.

O deepfake no contexto da crise de confiabilidade

A partir desta prévia análise acerca do contexto de crise generalizada da confiança, este texto presente busca considerar como a tecnologia algorítmica emergente do deepfake se insere no mesmo cenário. Paris e Donovan (2019, p. 5) indicam que os primeiros exemplos mais amplamente reconhecidos de tecnologia deepfake datam de novembro de 2017, quando um usuário do fórum Reddit enviou uma série de vídeos com os rostos de atrizes famosas, como Gal Gadot e Scarlet Johansson, enxertados nos corpos de outros atores em situações pornográficas:

> Desde então, a mídia noticiosa e, portanto, o público em geral, começaram a usar o termo '*deepfake*' para se referir a este gênero de vídeos que usam alguma forma de aprendizado 'profundo' ou de máquina para hibridizar ou gerar corpos e rostos humanos. (PARIS; DONOVAN, 2019, p. 5)

Em uma definição muito sucinta, o deepfake consiste, sobretudo, em manipulação audiovisual (ibid.). O deepfake é produto de aplicações de IA que fundem, combinam, substituem e sobrepõem imagens e clipes de vídeos para criar vídeos falsos, que parecem autênticos, de pessoas dizendo qualquer coisa em situações de caráter humorístico, pornográfico e/ou político – sem seu consentimento (WESTERLUND, 2019, p. 39). Isto significa que o deepfake, seja de forma intencional ou acidental, comumente distorce a percepção de terceiros a respeito de um indivíduo, associando seu nome e identidade a ideias e/ou atos que, por definição, não exprimem a realidade ou verdades sobre aquela pessoa. Karnouskos (2020, p. 1) considera os riscos e caos potencial, por exemplo, no simples cenário de uma teleconferência de uma empresa que não tem como avaliar precisamente quem está, de fato, presente: quais imagens são verdadeiras, de pessoas realmente conectadas, e quais são deepfakes.

O autor também traz um exemplo do que talvez possamos chamar de *metadeepfake*: o caso de um vídeo deepfake de 2018 com Barack Obama falando sobre os riscos dos deepfakes, algo que nunca acontecera. Por fim, Karnouskos registra a certa facilidade com que um usuário não avançado, com um computador doméstico e aplicativos disponíveis gratuitamente na Web, como o *DeepFaceLa*b (ibid.), já pode criar deepfakes de níveis simples a intermediário – portanto, capazes de ludibriar os sentidos de muitos. Este fenômeno dos deepfakes feitos por indivíduos comuns, com tecnologia ordinária, é denominado por outros autores de *cheap fakes* (em tradução literal, "falsos baratos"). Os *cheap fakes* também são deepfakes, ou seja, tratam de manipulação audiovisual – apenas de forma menos sofisticada. Paris e Donovan caracterizam com clareza as diferenças entre deepfake e *cheap fake* e já apontam as complexas consequências para sociedade, cultura, política e, até mesmo, para o conceito de verdade – certamente próximo da confiança – que, neste contexto tecnológico, ganharia um caráter mais nitidamente relativista e socioconstrutivista.

> O processo deepfake é o meio mais dependente do computador e também o menos publicamente acessível de criar mídia enganosa. Outras formas de manipulação audiovisual – "cheap fakes" – dependem de software barato e acessível, ou de nenhum software. Ambos deepfakes e cheap fakes são capazes de borrar a linha entre expressão e evidência. Ambos podem ser usados para influenciar a política das evidências: como as evidências mudam e são alteradas por sua existência em estruturas culturais, sociais e políticas. (PARIS; DONOVAN, 2019, p. 2–3)

> A cobertura de notícias afirma que os deepfakes estão prestes a destruir a reivindicação da verdade por meio dos vídeos, borrando permanentemente a linha entre o vídeo comprobatório e o expressivo. Mas o que a cobertura desse fenômeno deepfake muitas vezes ignora é que a "verdade" do conteúdo audiovisual nunca foi estável – a verdade é social, política e culturalmente determinada. E as pessoas são capazes de manipular a verdade com deepfakes e cheap fakes. (Ibid., p. 6)

Por tudo que vimos até agora, não é difícil perceber a conexão multifacetada e profunda entre questões que giram em torno da confiabilidade e do deepfake. Por exemplo: distância entre público e privado, privacidade, anonimato e o que é ou não associado – inclusive falsamente – a nosso nome e identidade. Afinal, "[o] respeito está ligado aos nomes" e "[p]ode-se definir confiança como crença nos nomes" (HAN, 2018, p. 14–15). Becker (1996, p. 44) lembra que a confiança é um elemento das relações interpessoais que envolve nossas expectativas a respeito dos outros – por exemplo, de alguma ação (HAWLEY, 2014, p. 7) e/ou quanto a manterem sua palavra. A confiabilidade trata também, pois, da esperança que depositamos em indivíduos específicos e de nossa percepção e crença na reciprocidade de atos, como sugere Adam Smith (1984, p. 337): "Nós confiamos no homem que parece disposto a confiar em nós".

No entanto, como criar expectativas sobre os outros – e, em tempo, laços sólidos de fieza – no atual contexto já estabelecido de crise de confiança, então salpicado por tecnologias, que, por um lado, substituem a fidúcia por segurança criptográfica, caso do *blockchain*; e, por outro, recursos que permitem a criação de manipulações audiovisuais cada vez mais convincentes, como os deepfakes?

A crença de correspondência a fatos, que o deepfake cria e da qual é quase impossível se livrar, pode ser explicada pelo funcionamento do sistema perceptivo humano, a saber, o ser humano não está mentalmente equipado para duvidar daquilo que seus olhos veem. Isso não significa que a percepção não erra, mas sim, que, para corrigir o erro, um dado perceptivo precisa ser comparado a outro. É apenas em função dessa comparação que é possível constatar um erro ou equívoco perceptivo (SANTAELLA, 2012, p. 89-138).

Além disso, para complementar as consequências nefastas do deepfake, há o elemento complicador das redes digitais, que permitem uma propagação muitas vezes viral de seu conteúdo (KARNOUSKOS, 2020, p. 2). Entre outras questões, as redes digitais contribuem para extinguir a distância e a diferença entre o público e privado: "A falta de distância leva a que o privado e o público se misturem. A comunicação digital for-

nece essa exposição pornográfica da intimidade e da esfera privada. Também as redes sociais se mostram como espaços da exposição do privado" (HAN, 2014, p. 13). Uma vez que uma das tônicas do deepfake consiste em explorar situações ligadas à moralidade, especialmente em seus aspectos risíveis, é importante verificar o tipo de abrigo específico que as redes dão ao deepfake e em qual medida intensificam a crise de confiança generalizada que estamos vivendo.

A exibição do espaço privado nas redes

A desconfiança foi e é estudada por muitos autores em áreas como filosofia, psicologia e sociologia – nem por isso, ela faz por merecer uma atenção suficientemente distinta. Jason D'Cruz indicou a importância desta diferenciação: "A confiança é comumente descrita usando a metáfora de uma 'cola social' invisível que só chama a atenção quando está ausente. É inicialmente tentador pensar na desconfiança como a ausência da 'cola' da confiança. Mas essa maneira de pensar simplifica demais a relação conceitual entre confiança e desconfiança" (D'CRUZ, 2020, p. 41).

A partir de outras referências, o autor delineia que a desconfiança geralmente vem acompanhada de sentimento de insegurança, cinismo, desprezo e medo, ou seja, trata de uma postura ativamente negativa, em vez da certa neutralidade de um modo agnóstico do tipo "aguarde e observe". É importante distinguir, ainda, a desconfiança de ordem moral da mera desconfiança normativa quanto a uma habilidade particular, por exemplo (ibid., p. 42). Desconfiar que uma pessoa é uma assassina é bem mais grave do que suspeitar de sua capacidade de cozinhar – ou de seu tato social. A desconfiança de teor moral é, certamente, sua versão crucial também em relação aos deepfakes e seus efeitos potencialmente mais devastadores entre pessoas, comunidades, confiança e tecido social.

Katherine Hawley, outra estudiosa sobre temas em torno de confiança e desconfiança, define ambas quanto a expectativas de que uma pessoa nos faça algo da seguinte forma: "Confiar em alguém para fazer algo é acreditar que ela tem o compromisso de fazê-lo e confiar nela para cumprir esse compromisso. Desconfiar de alguém para fazer algo é acreditar que ela tem o compromisso de fazê-lo, mas não confiar nela para cumprir esse compromisso" (HAWLEY, 2014, p. 10). A autora distingue "confiança" de "contar com algo ou alguém" (em inglês, *reliance*, frequentemente traduzido para a língua portuguesa como "confiança"). "Conto com a estante para que suporte o vaso [...] eu não confio na estante. Mas

também não desconfio dela" (ibid., p. 3). Deve-se notar que "contar com" também se aplica a pessoas: como diz Hawley, não contar com os outros não significa, necessariamente desconfiar deles. "Meus colegas nunca me compraram champanhe, então, em particular, não conto com eles para que me comprem champanhe na próxima sexta-feira" (ibid.).

No caso das deepfakes, as questões centrais parecem muito mais frequentemente – e mesmo, parcialmente, por definição – ocorrer a respeito de verdade ou mentira, com caráter pessoal e moral. Desta forma, a desconfiança – acompanhada comumente por insegurança, cinismo, desprezo e medo – seria uma consequência inescapável da manipulação audiovisual com objetivo de produzir conteúdo falso, algo que, a depender da qualidade do deepfake e da capacidade e atenção de seu intérprete, pode fabricar uma quantidade ilimitada de ruídos e leituras equivocadas sobre pessoas e, desta forma, sobre a própria realidade.

Finalmente, o efeito repetitivo das deepfakes nas redes e cumulativo ao longo de meses e anos somado a outros fatores discutidos em textos anteriores e à própria crise generalizada de confiança merece atenção. Se há algo bem estabelecido sobre a natureza humana é que nós aprendemos e adaptamos nosso comportamento: é bastante plausível propor que, se determinadas mídias, canais, grupos ou pessoas constantemente nos expõem a conteúdo falso, mentiroso ou distorcido – como deepfakes ou afins –, nosso grau de desconfiança em relação àqueles aumentará mais ainda. Ou seja, tanto confiança quanto desconfiança são modulares e podem variar em grau e magnitude ao longo do tempo, conforme as pesquisas relatadas no início deste artigo comprovam – não é uma questão binária, tal que há confiança ou não há; ocorre uma acumulação de evidências que empurra em uma ou na outra direção. Em outras palavras: quanto menor for a confiabilidade, por exemplo, de informações obtidas em redes digitais ou mecanismos de pesquisa *online*, mais grave fica a situação da estabilidade social e mais difícil será recuperar a confiança das pessoas.

Considerações finais

A partir deste artigo e de trabalhos anteriores, é possível destacar um dilema ou enigma que surge como consequência dos referidos estudos quanto ao uso de tecnologias digitais baseadas em algoritmos de IA – especificamente, os casos do *blockchain* e da deepfake – na conjuntura da grave crise de confiança mundial, com ênfase a ambientes digitais. Por

um lado, com o *blockchain*, temos a *mecanização da confiança* a partir da substituição de confiabilidade por segurança criptográfica, o que paradoxalmente elimina a confiança como laço humano historicamente experienciado da equação – ou seja, da troca social e/ou comercial.

Por outro lado, desde o sofisticado simulacro da deepfake, temos uma mais óbvia transgressão da verdade e o processo aqui denominado *mecanização da desconfiança*, que – diferente do blockchain – não traz consigo nenhum paradoxo ou complexidade. De outro modo: a deepfake realiza de fato o que se propõe a fazer e tende a propagar falsidades e informações erradas, muitas vezes desmoralizantes, a respeito de pessoas e grupos de tal modo a incitar reações equivocadas, ruídos comunicacionais e, no limite da lei, os crimes de difamação e calúnia. Desta forma, não é difícil concluir que, a exemplo do blockchain, mas de outra maneira, a deepfake também contribui, mesmo que de modo fragmentado e que, a rigor, deve ser analisado caso-a-caso, para o aprofundamento da crise de confiança que já abala o mundo.

O dilema e enigma, pois, dizem respeito exatamente à situação apresentada: se uma tecnologia de IA criada com a intenção de melhorar a segurança das trocas (*blockchain*) termina por substituir e eliminar a confiabilidade, reduzindo sua vivência como laço humano; e outra aplicação (deepfake) voltada para manipulações audiovisuais falsas efetivamente e ativamente aflige a confiança pessoal e social, o que pode ser feito? Haverá alguma tecnologia digital e, mais precisamente, algo baseado em algoritmos de IA que possa mitigar a severa crise de confiança generalizada que aflige o mundo – ou seria um padrão inescapável?

Referências

BECKER, Lawrence. Trust as noncognitive security about motives. *Ethics*, Chicago, v. 107, p. 43-61, 1996.

BRENAN, Megan. Americans remain distrustful of mass media. *Gallup*. *Politics*, September 30, 2020. Disponível em: news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx. 2020. Acesso em: 29 abr. 2021.

D'CRUZ, Jason. Trust and distrust. *In*: SIMON, Judith (ed.). *The Routledge handbook of trust and philosophy*. London: Routledge, 2020, p. 41-51.

EDELMAN Trust Barometer 2021. Disponível em: edelman.com/sites/g/files/aatuss191/files/2021-01/2021-edelman-trust-barometer.pdf. 2021. Acesso em: 29 abr. 2021.

HAN, Byung-Chul. *A sociedade da transparência*. Lisboa: Relógio D'Água, 2014.

_____. *No enxame*. São Paulo: Vozes, 2018.

HARARI, Yuval. The world after coronavirus. *Financial Times*, March 20, 2020. Disponível em: ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75. Acesso em: 29 jun. 2021.

HAWLEY, Katherine. Trust, distrust and commitment. *Noûs*, Hoboken, NJ, v. 48, n. 1, 2014, p. 1–20. Disponível em: onlinelibrary.wiley.com/doi/epdf/10.1111/nous.12000. 2014. Acesso em: 29 jun. 2021.

KARNOUSKOS, Stamatis. Artificial Intelligence in digital media: the era of deepfakes. *In: IEEE Transactions on Technology and Society*, v. 1, n. 3, Sept. 2020, p. 138-147. Disponível em: researchgate.net/publication/342795647_Artificial_Intelligence_in_Digital_Media_The_Era_of_Deepfakes. 2020. Acesso em: 28 jun. 2021.

LIU, Xiao et al. Shaping the future of the Internet of bodies: new challenges of technology governance (Briefing Paper, World Economic Forum, July 2020). Disponível em: weforum.org/docs/WEF_IoB_briefing_paper_2020.pdf. 2020. Acesso em: 10 jul. 2021.

NEWMAN, Nic, et al. *Reuters Institute Digital News Report 2019*. Oxford. Disponível em: reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_0.pdf. 2019. Acesso em: 9 jun. 2021.

NEWMAN, Nic, et al. *Reuters Institute Digital News Report 2020*. Oxford. Disponível em: reutersinstitute.politics.ox.ac.uk/risj-review/google-and-university-oxford-agree-extension-support-digital-news-project-august-2020. Acesso em: 9 jun. 2021.

NEWMAN, Nic, et al. *Reuters Institute Digital News Report 2021*. Oxford. Disponível em: reutersinstitute.politics.ox.ac.uk/digital-news-report/2021. Acesso em: 12 jul. 2021.

NOBREGA, Ighor. PoderData mostra queda de confiança dos brasileiros na imprensa. *Poder 360*, 29 dez. 2020. Disponível em: poder360.com.br/poderdata/poderdata-mostra-queda-de-confianca-dos-brasileiros-na-imprensa/. 2020. Acesso em: 8 jun. 2021.

PARIS, Britt; DONOVAN, Joan. *Deepfakes and cheap fakes*. Thousand Oaks: Sage (=Data & Society's Media Manipulation research initiative). Disponível em: datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1-1.pdf. 2019. Acesso em: 29 abr. 2021.

SALGADO, Marcelo. Blockchain e a crise de confiança na sociedade do controle. *In*: SANTAELLA, Lucia (org.). *A expansão social do blockchain*. São Paulo: EDUC, 2020. p. 25–39.

SANTAELLA, Lucia. *Percepção:* fenomenologia, ecologia, semiótica. São Paulo: Cengage Learning, 2012.

SANTAELLA, Lucia. Blockchain: de onde veio, onde está e para onde vai. *In*: SANTAELLA, Lucia (org.). *A expansão social do blockchain*. São Paulo: EDUC, 2020. p. 11–24.

SCHWAB, Klaus; MALLERET, Thierry. *Covid-19:* The great reset. La Vergne, TN: Lightning Source, 2020.

SMITH, Adam. *The theory of moral sentiments*. Indianapolis, IN: Liberty Fund, 1984.

STATCOUNTER GlobalStats. Browser market share worldwide, July 2020-July 2021. Disponível em: gs.statcounter.com/search-engine-market-share. 2021. Acesso em: 8 jun. 2021.

TWENGE, Jean; CAMPBELL, W. Keith; CARTER, Nathan. Declines in trust in others and confidence in institutions among American adults and late adolescents, 1972–2012. *Psychological Science*, v. 25, n.10, p. 1914-1923, 2014. Acesso em: 28 jun. 2021.

WESTERLUND, Mika. The emergence of deepfake technology: a review. *Technology Innovation Management Review*, v. 9, n.11, p. 39-52, Nov. 2019. Acesso em: 29 abr. 2021. p. 39–52.

WILLIAMS, Stephen. *Blockchain:* the next everything. New York, NY: Scribner, 2019.

FANAYA, Patrícia Fonseca. Deepfake e a realidade sintetizada. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 104-118.

Recebido em: 25 jan. 2021
Aprovado em: 22 fev. 2021

dx.doi.org/10.23925/1984-3585.2021i23p104-118

Deepfake e a realidade sintetizada

Patrícia Fonseca Fanaya[1]

Resumo: As tecnologias deepfake, nas quais os computadores são treinados para manipular áudio e imagens em um nível cada vez mais impressionante de perfeição, estão sendo desenvolvidas a uma velocidade vertiginosa, têm um potencial de popularização exponencial e carregam o poder de espalhar ainda mais desinformação e mentiras disfarçadas de realidade e verdade. Além disso, podem alimentar com ainda mais sofisticação as mais bizarras teorias da conspiração que têm o poder de ameaçar, a política, o direito e a ciência. Diante disso, as tecnologias deepfake tornarão a guerra da desinformação ainda mais complexa e perigosa, fazendo com que qualquer distinção entre verdadeiro e falso ou real e fictício entre em colapso total. Este artigo tem como objetivo contextualizar as deepfakes no problema da realidade sintetizada ou simulada e relacioná-lo com o problema da verdade.

Palavras-chave: Deepfake. Realidade Simulada. Realidade Sintetizada. Verdade.

1 Doutora em Comunicação e Semiótica PUC-SP; Mestre em Estudos da Tradução PGET/UFSC; Bacharel em Comunicação Social–Publicidade UnB-DF. Pesquisadora pós-doutorado PPGFIL/UFSC. ORCID: orcid.org/0000-0002-4842-9813. CV Lattes: lattes.cnpq.br/1786831156933368. E-mail: patriciafanaya@gmail.com.

Deepfake and synthesized reality

Abstract: Deepfake technologies, by which computers are trained to manipulate audio and images to an increasingly impressive level of perfection, are being developed rapidly. They have a great potential of popularization as well as the power of spreading disinformation and lies disguised as reality and truth. Besides, they can feed with even more sophistication the most bizarre conspiracy theories, which have the power to threaten policies, the law, and science. Hence, deepfake technologies make the war of disinformation even more complex and dangerous, causing any distinction between true and false or real and fictitious to break down. This article aims at contextualizing the phenomenon of deepfakes within the context of synthesized or simulated reality, relating it to the problem of truth.

Keywords: Deepfake. Simulated Reality. Synthesized Reality. Truth.

Tempos de guerra da desinformação

O termo deepfake refere-se às mídias sintéticas nas quais imagens ou sons capturados de determinadas pessoas são substituídos pelos de outras por meio de técnicas avançadas de aprendizagem de máquina e Inteligência Artificial (IA), com a finalidade de manipular conteúdos visuais e/ou sonoros, com enorme potencial de falseamento da realidade. Celebridades e políticos têm sido os alvos preferenciais dessas manipulações, que têm se popularizado exponencialmente até mesmo por meio de aplicativos gratuitos de celular – o que tem colocado em risco a reputação de muita gente, servido às vinganças pessoais mais comezinhas e alimentado a poderosa, faminta e perversa indústria das fake news.

Em 2018, o diretor e ator Jordan Peele fez um vídeo com imagens do ex-presidente dos EUA, Barack Obama, utilizando-se de tecnologias deepfake como forma de chamar atenção para as controversas e mesmo gravíssimas implicações da popularização desses tipos de tecnologia. No vídeo, Peele usou como base uma gravação original de Obama e inseriu digitalmente sua própria voz e imagens dos movimentos de sua boca, fazendo com que Obama "dissesse" coisas que jamais havia dito, por exemplo, alguns chamamentos pouco lisonjeiros direcionados ao então presidente e adversário político, Donald Trump.

Peele, BuzzFeed e Monkeypaw Productions se utilizaram de um software facilmente disponível para manipular a edição do vídeo e criar o deepfake. De acordo com o site *BuzzFeed* (2018), os deepfakes são criados, em sua maioria, com o *FakeApp*, que é um software de IA gratuito, e que foi popularizado em fóruns dedicados ao compartilhamento de vídeos falsos no Reddit e no Discord. O vídeo de Peele viralizou na internet, muitas vezes reeditado, sem o trecho original em que ele aparece dividindo a tela com a imagem de Obama e no qual ele mostra claramente se tratar de um deepfake. Como tudo que circula nas redes pode ser facilmente adulterado digitalmente, essa não seria uma exceção. A viralização do vídeo gerou grande repercussão nas mídias em geral, e a indústria das fake news encontrou um prato cheio para alimentar os ativistas da direita mais raivosa (e perigosa) que apoiava Trump, à época.

De 2018 para cá, o desenvolvimento, aperfeiçoamento e popularização de várias outras tecnologias deepfake continuam a surgir e, em mãos erradas, inegavelmente, têm poder atômico de espalhar ainda mais desinformação e mentiras travestidas de realidade e verdade; e alimentar com ainda mais sofisticação as teorias da conspiração mais bizarras que, atualmente, já têm o poder de ameaçar de maneira sórdida, não só a política, a lei, a economia, mas também a ciência — o que em épocas como a nossa, de pandemia global de Covid-19, representa um risco potencial e adicional à vida de muitas pessoas, especialmente aquelas com tendências a comportamentos negacionistas. Frente a isso, as tecnologias deepfake tornarão ainda mais complexa e perigosa a guerra da desinformação, ao fazerem com que qualquer distinção entre verdadeiro e falso ou real e fictício colapse completamente.

É importante frisar que há muitas categorias de deepfake com propósitos diversos, que vão desde a pesquisa acadêmica em diversas áreas das ciências da computação e suas interseções; passando pelo desenvolvimento e usos comerciais; até chegar ao uso amador, que serve a inúmeros fins, sejam eles estritamente pessoais ou de grupos, com consequências divertidas (como os memes), nefastas e tudo o que pode haver entre esses dois extremos. De maneira geral, neste artigo, abordo as deepfakes com vistas às consequências de sua aplicação indiscriminada e/ou sem algum tipo de caracterização explícita e/ou controle, e não a partir de seu desenvolvimento técnico ou seus propósitos científicos. O objetivo deste artigo é contextualizar as deepfakes no âmbito do problema da realidade sintetizada/simulada e relacioná-lo com o problema da verdade.

Tecnologia e realidade simulada

As discussões sobre a realidade acompanham a história da filosofia ocidental. Aliás, a própria filosofia surge como uma busca de explicação racional para o mundo e a natureza, e como contraponto às crenças mitológicas sobre a criação do mundo. Os gregos, desde os pré-socráticos, já tratavam sobre o problema da realidade, mas a intenção, aqui, não é traçar um panorama histórico ou cronológico do tema na história da filosofia — até porque isso seria impossível—, mas apenas contextualizar brevemente o leitor menos familiarizado com a disciplina e/ou com a discussão.

Para o espanto de muitos, as reflexões filosóficas sobre as tecnologias também já animavam discussões acaloradas na Grécia antiga. Platão, Demócrito e Aristóteles já se ocupavam com temas filosóficos relaciona-

dos às tecnologias. Aristóteles, por exemplo, dentre outras contribuições que atravessaram os séculos, fez uma importante distinção ontológica entre as coisas naturais e os artefatos. Para ele, as coisas naturais (animais, plantas e os quatro elementos) movem-se, crescem, mudam e se reproduzem em função de causas finais internas e estão, portanto, de acordo com os propósitos da natureza. Os artefatos, por outro lado, sem a intervenção humana, perdem suas formas artificiais e se decompõem em outros materiais e/ou desaparecem por completo.

Tales de Mileto, por exemplo, afirmava que a água era o princípio de todas as coisas; para Anaxímenes, o ar seria a origem da terra, da água e do fogo; Parmênides foi o primeiro a mostrar que a Terra era esférica, a afirmar que espírito e alma eram a mesma coisa e a considerar que havia dois tipos de filosofia: uma que se referia à verdade e outra à opinião. Foi ainda Parmênides que afirmou que nada nasce do nada, e que nada do que existe se transforma em nada, além de afirmar que as transformações que observava na natureza não eram mudanças "reais". Parmênides defendia a doutrina da existência de uma única realidade (monismo), e deu início ao que mais tarde ficou conhecido como racionalismo — que é a ideia de que é a razão humana, e não os sentidos, é a fonte primária do conhecimento do mundo. Heráclito, ao contrário de Parmênides, acreditava que tudo muda o tempo todo, que a realidade era fluida e que as percepções são fonte de conhecimento da realidade.

Platão, por sua vez, já elaborava aspectos importantes das diferenciações entre a realidade do mundo e a realidade "simulada" pelas imagens. Na Alegoria da Caverna, Livro VII de A República (2006, p. 267-272), ele nos mostrou como as imagens podem ser fontes de distorções e como elas são capazes de simular a realidade e, ao mesmo tempo, diferenciar-se dela. Platão entendeu, como poucos, o imenso poder que as imagens têm de afetar os sentidos e a mente, e também sobre como elas ficam indelevelmente gravadas em nós.

No século XVII, Descartes construiu seu argumento cético baseado na ideia de existência de um gênio demoníaco dedicado a enganá-lo. Para lidar com ele, precisaria duvidar de tudo, até mesmo dos próprios sentidos que, de acordo com ele, haviam lhe ensinado tudo o que sabia até então sobre a realidade, e suspender o juízo sobre aquilo que já havia considerado verdadeiro:

> Sou obrigado a confessar que, de todas as opiniões que no passado considerei verdadeiras, não existe nenhuma da qual hoje não possa duvidar, não por alguma falta de consideração ou imprudência, mas por razões muito fortes e refletidas: de modo que é preciso que de agora em diante suspenda meu juízo a respeito de tais pensamentos, e que não mais lhes dê crédito, como faria com as coisas que se me afiguram manifestamente falsas, se desejo encontrar algo de imutável e de indubitável nas ciências. (DESCARTES, 1999, p. 254)

A importância de Descartes para a hipótese da realidade simulada é que ele nos possibilitou duvidar de tudo que existe, além de ter elaborado de maneira muito sofisticada o argumento platônico de que não podemos confiar em nossos sentidos, e que, portanto, a realidade que percebemos pode ser simulada.

A filosofia e a ciência costumam dialogar com frequência, mesmo que estes diálogos sejam repletos de dissensos, e desde Descartes e Newton, ou seja, desde a era considerada clássica na física, "o universo físico aberto a nossas pesquisas explodiu", como apontou Prigogine (1984, p. 163).

No início do século XX, em mais um movimento de disrupção que costuma acontecer de tempos em tempos na ciência, Einstein apresentou ao mundo a teoria da relatividade, e em companhia de tantos outros físicos importantes, como Heisenberg, Planck, Bohr, Schrödinger, von Neumann, entre outros, ajudou a estabelecer as bases da mecânica quântica. A mecânica quântica mudou radicalmente a forma de a ciência lidar com o problema da realidade, provando através de experimentos a probabilidade de universos paralelos. A partir da mecânica quântica vimos surgir a computação quântica. A computação quântica, por sua vez, está por trás de inúmeros avanços contemporâneos espantosos nos campos não só das ciências, mas das tecnologias, especialmente a IA.

Nick Bostrom, em seu artigo "Are Your Living in a Computer Simulation?" (2003) desenvolveu e expandiu o argumento filosófico sobre a probabilidade de aquilo que chamamos de realidade ser uma simulação. Em seu argumento, Bostrom mostra que pelo menos uma das seguintes proposições é verdadeira: (1) a espécie humana tem grande probabilidade de se extinguir antes de atingir um estágio "pós-humano"; (2) é extremamente improvável que qualquer civilização pós-humana execute um número significativo de simulações de sua história evolutiva (ou variações dela); (3) quase certamente estamos vivendo em uma simulação de computador. Segue-se que a crença de que há uma chance significativa de um dia nos tornarmos pós-humanos que executam simulações de seus ancestrais é falsa, a menos que estejamos atualmente vivendo em uma simulação.

O argumento da simulação, de acordo com Bostrom, é fundamentalmente diferente dos argumentos filosóficos tradicionais sobre a realidade, inclusive do argumento cético cartesiano, porque não estabelece um problema cético como desafio às teorias epistemológicas e ao senso comum; mas diz que temos razões empíricas interessantes para acreditar que é possível, em algum nível, uma afirmação disjuntiva sobre a realidade do mundo. O argumento da simulação, de acordo com ele, depende crucialmente de premissas empíricas que não são óbvias sobre as tecnologias futuras e nossas habilidades com elas. Além disso, a conclusão do argumento da simulação não é simplesmente que não podemos ter certeza de que não estamos vivendo em uma simulação.

Bostrom (2003) afirma ainda que o argumento da simulação também é diferente dos argumentos do "cérebro numa cuba", isto é, do cenário de um hipotético cérebro isolado do corpo do seu dono humano, cujos neurônios são conectados por fios a um supercomputador e fornece impulsos elétricos idênticos aos que o cérebro humano recebe. Em vez disso, o ponto de partida é que as coisas são do jeito que acreditamos que são e, em seguida, embora possamos justificá-las e atribuir crédito inicial a essa crença, ele nos mostra que temos razões empíricas específicas para revisar nossas crenças iniciais e não para nos tornarmos agnósticos sobre a existência de um mundo externo, mas para aceitar a possibilidade de uma conclusão disjuntiva. Assim, a melhor comparação do argumento da simulação não é com o argumento cético (que nos tornaria ainda mais agnósticos), mas com um argumento que nos fizesse aumentar nossa crença em uma disjunção particular, e, consequentemente, diminuir nossa crença em sua negação. O objetivo do argumento da simulação é dizer algo sobre o mundo, em vez de mostrar que sabemos menos sobre ele do que pensávamos.

O interessante entre o diálogo da filosofia com a ciência é que, enquanto a filosofia se ocupa da construção sólida dos argumentos (mesmo que o argumento seja especulativo), a ciência precisa formular experimentos que lhes forneçam provas empíricas. No caso, apesar de a hipótese da realidade simulada receber apoio de parte da comunidade científica, muitos físicos teóricos a refutam (mas não a descartam completamente) por entenderem que, atualmente, não há computadores qualificados para a tarefa de processar e projetar uma simulação perfeita de todo o universo. Neste contexto, convém lembrar o que o "efeito Hall" revela. Essa teoria diz que a simulação de algo se torna mais complexa e improvável conforme a quantidade de partículas, átomos ou dados em questão au-

mentam. No caso, para simular tal sistema, seria necessário diagonalizar a matriz em questão num computador, o que é praticamente impossível com a tecnologia que se tem disponível hoje, porque a quantidade de metadados do universo ultrapassaria a própria capacidade de armazenamento do universo.

Verdadeiro e simulado?

Na filosofia, o problema da verdade é um dos mais antigos e abrangentes, e é abordado há milhares de anos sob diversas perspectivas: pela metafísica, lógica, epistemologia, ética etc.; além de serem muitas as teorias da verdade: teorias da coerência, da correspondência, deflacionista, pragmatista etc. (cf. Haack, 2002). Assim como o problema da realidade, o problema da verdade atravessa a história da filosofia, e recentemente, com os avanços das ciências e das tecnologias nas últimas décadas, que têm modificado radicalmente nossas possibilidades e capacidades de interação no mundo e com a realidade, as discussões sobre o tema têm ganhado novos e interessantes contornos.

As teorias contemporâneas pluralistas da verdade têm suas raízes no pragmatismo de William James (2003). Em linhas gerais, James considerava que as crenças verdadeiras eram aquelas que serviam a algum propósito, porém, ele apontou para o fato de que há muitas maneiras diferentes de as crenças servirem a propósitos, e que muitas vezes isso depende do campo em que as estamos tratando, por exemplo, na moral, na matemática, entre outros.

Subjacente às formas de pluralismo está a ideia de que o conceito de verdade pode exigir um tratamento diferenciado em diferentes campos expressos nos domínios do discurso; ou seja, na ideia de que tanto o pensamento quanto os discursos humanos podem abranger um grande número de temas diferentes, em contextos também diferentes. Por exemplo, podemos discutir se determinada piada é engraçada; ou ainda se a própria atitude de contar essa piada é moralmente condenável. Como esses debates pertencem a campos diferentes (retórica e moral, respectivamente), precisaríamos levar isso em consideração. Para um pluralista, qualquer relato completo sobre a natureza da verdade precisa considerar um domínio específico, em contraste ao que constitui a verdade *per se*.

As teorias pluralistas da verdade assumem que há mais de uma maneira de algo ser verdadeiro. É muito importante manter o cuidado aqui, porque assumir que algo pode ser verdadeiro em mais de uma maneira é algo bem diferente de assumir que há muitas "verdades", que é uma reivindicação do(s) relativismo(s).

O pluralismo, no geral, é tido como uma alternativa ao monismo (conceito único de verdade) e ao relativismo (não há verdade, pois, cada pessoa ou grupo de pessoas define suas próprias verdades, escolhe seus próprios valores etc.). No entanto, como em filosofia tudo pode ser ainda mais complicado do que parece à primeira vista, Susan Haack, no seu artigo "The Unity of Truth and the Plurality of Truths", apresenta-nos uma argumentação refinada sobre a possibilidade de haver um conceito de verdade não ambíguo, não relativo, e ainda assim muitas e variadas proposições verdadeiras (expressões de "verdades"):

> Dizer que uma coisa é verdadeira é dizer (não que alguém, ou todos, acreditem nisso, ou que decorra desta ou daquela teoria, ou que existe uma boa evidência disso, mas) simplesmente que as coisas são como dizem; as muitas verdades, [são] reivindicações empíricas particulares, teorias científicas, proposições históricas, teoremas matemáticos, princípios lógicos, interpretações textuais, declarações sobre o que uma pessoa quer, acredita ou pretende, sobre a gramática e regras legais, etc. (HAACK, 2005, p. 88)

Como se pode notar, movemo-nos em campo minado. Entretanto, torna-se cada vez mais difícil evitar essas discussões, dado o fato de que já temos que lidar com problemas éticos e morais — só para citar os campos mais evidentes aos olhos de qualquer pessoa — que emergiram juntamente com os avanços tecnocientíficos das últimas décadas, e que eram inexistentes e/ou inimagináveis para as gerações anteriores, dentre eles, as deepfakes.

Deepfakes e verdade na realidade simulada

A manipulação de som e imagem não constitui novidade. A história nos mostra que, mesmo antes da fotografia e do cinema, meios passíveis de manipulações técnicas as mais diversas, artistas financiados pela realeza e classes abastadas já falseavam a "realidade" da aparência de seus patronos com o objetivo de torná-los mais bonitos, mais altos, mais heroicos, e assim por diante.

No cinema, há muito as técnicas de manipulação/simulação da realidade vêm sendo utilizadas – primeiro com a introdução do *chroma key*, uma técnica que permite substituir uma cor sólida (verde ou azul) por qualquer outra; depois com os avanços da computação gráfica, que permitiu, entre outras coisas, que o gênero épico ganhasse novas possiblidades nas narrativas cinematográficas, se utilizando de poucos atores que se transformavam em milhares de guerreiros lutando em batalhas campais

memoráveis; ou, ainda, em tempos mais recentes, "rejuvenescendo" digitalmente atores já idosos para fazerem papel deles mesmos em suas versões mais jovens (Robert De Niro, em *O Irlandês*, de Martin Scorsese, 2019).

Mas o aperfeiçoamento, desenvolvimento e aplicações de novas tecnologias sempre supera o que conseguimos imaginar a princípio, e hoje as tecnologias deepfake, que estão baseadas em *deep learning* (um ramo da aprendizagem de máquina) e que aplicam simulações de redes neurais a conjuntos massivos de dados (*big data*) são capazes de criar, por exemplo, pessoas que nunca existiram no "mundo real". Essas criações sintéticas podem estrelar no próximo filme de Scorsese, contracenar com De Niro ou qualquer outro ator de carne e osso, em qualquer língua, ter perfil no Instagram, no Facebook, interagir conosco no Twitter, sem que tenhamos ideia de que elas não existem na vida "real". Será que não estamos presenciando a inauguração de uma nova era da realidade sintetizada, que transbordará para todos os campos da vida humana?

O maior problema não parece estar nem no aperfeiçoamento e nem no uso circunscrito de tecnologias de simulação da realidade – as ciências, o entretenimento, as engenharias, entre outros campos do conhecimento, já usufruem dessas possibilidades há tempos, com diferentes finalidades –, mas em seu fácil acesso e popularização indiscriminada disfarçada de democratização.

Além disso, o fato de vivermos numa época de supervalorização do *DIY* (*Do it yourself*), de crescente desacreditamento das ciências e na qual os relativismos imperam, vemos surgir um cenário propício para o desastre se formando no horizonte, pois as armas para a "desestabilização da civilização" (BOSTROM, 2019, p. 455) já estão se espalhando como rastros de pólvora pelo globo terrestre.

Há inúmeros casos recentes de DIY com tecnologias promissoras e importantes para a ciência, que levantaram questões éticas e morais cruciais que precisam ser observadas e tratadas com a atenção e o rigor que merecem. Um dos casos notórios foi o do *biohacker* Aaron Traywick, que foi encontrado morto pela polícia, imerso num tanque de terapia de flotação. Traywick ficou famoso após injetar-se, ao vivo, num evento público, com um composto artesanal, resultado de pesquisa caseira para tratamento e cura de HIV-Aids e herpes.

A prática de *biohacking* conquistou o Vale do Silício e está geralmente associada à ideia de as pessoas terem mais controle sobre seus corpos e, para isso, vale tudo: das dietas mais mirabolantes, aos experimentos com

drogas e substâncias estranhas e sem comprovação científica; ou ainda manipulação genética com sofisticadas técnicas CRISPR – o kit com todo o material necessário para executar experimentos domésticos com bactérias pode ser legalmente adquirido no internet.

Os adeptos do *biohacking* usam e abusam do argumento político da defesa da liberdade de escolha sobre o que fazer com o próprio corpo. Ideias como autonomia e liberdades individuais precisam ser defendidas, sem dúvidas, mas o contexto importa. Quando há o risco real de pessoas comuns estarem criando em suas casas, sem nenhum tipo de regulamentação ou restrição, armas biológicas com potencial de destruição em massa a partir de um kit comercializado pela internet, está claro que isso não pode ser defendido como um problema restrito às escolhas individuais sobre o que fazer com o próprio corpo, simplesmente porque as consequências ultrapassam as ameaças à integridade física apenas do *biohacker*. O indivíduo tem papel importante na sociedade, mas o bem-comum e os outros indivíduos também os tem, e, neste caso, o argumento da liberdade individual em relação ao próprio corpo torna-se falacioso. Entretanto, como o tempo dos desenvolvimentos tecnocientíficos geralmente corre mais rápido do que o das instituições sociais constituídas, pessoas como Josiah Zayner não podem ser nem caladas ou punidas, porque, como ainda não há previsão desse tipo de crime na legislação vigente nos EUA, elas não estão cometendo crimes aos olhos da Justiça.

A pandemia de COVID-19 tem nos ensinado muito sobre como pode ser perigosa e maléfica a mistura promíscua de informações distorcidas e/ou remixadas (fake news + deepfakes) como tentativas de desacreditamento da ciência e relativização de suas descobertas. Presenciamos, todos os dias, pessoas crendo nas promessas de cura rápida e barata do kit Covid, que foi promovido por leigos, religiosos de todos os matizes, mas também por médicos e cientistas pouco afeitos aos rigores e cautelas da ciência. Assistimos, atônitos, à disseminação massiva de informações falsas e pânico tanto sobre a doença, como sobre as medidas protetivas e preventivas que deveríamos adotar a fim de nos proteger e proteger outras pessoas de ficarem doentes e até morrerem. Quando surgiram as primeiras vacinas, vimos o roteiro das fake news e das deepfakes se repetir com requintes de crueldade.

A mistura de medo, ignorância, negacionismo e interesses políticos os mais diversos é sempre muito perigosa. Se hoje a humanidade encontra-se ainda de joelhos frente à epidemia de Covid-19, causada pelo vírus SARS-COV-2 que ceifou a vida de milhões de pessoas em pouco mais de

um ano em circulação, e sobre o qual não há consenso acerca da origem e de como varreu o mundo em pouquíssimo tempo, como imaginar um mundo em que pessoas, sem formação mínima em biologia ou outras ciências, possam manipular o código genético de organismos, os mais diversos, nos quintais de suas casas? Será a era da ciência de garagem? Será a volta da alquimia e do curandeirismo?

As tecnologias deepfakes podem parecer pertencer a outra categoria — e realmente pertencem se considerarmos apenas suas aplicações restritas aos campos da imagem científica ou do entretenimento. No entanto, se considerarmos as possibilidades amplas de aplicação nos campos da informação e da comunicação e sua utilização indiscriminada e sem controle, veremos que a consequência pode ser o colapso completo das crenças nas instituições que mantêm as sociedades humanas relativamente funcionais, tornando cada vez mais complexa e perigosa a guerra da desinformação, e fazendo com que qualquer distinção entre verdadeiro e falso ou real e fictício colapse. O colapso do real e da verdade – mesmo que aceitemos sua pluralidade ou a pluralidade de suas expressões – tem consequências práticas perigosas, que, neste caso, podem mesmo significar a extinção da humanidade. A pergunta que persiste é: há saída?

Uma luz no fim do túnel?

No parágrafo final de *Sapiens: Uma Breve História da Humanidade*, Harari escreve: "Existe algo mais perigoso do que deuses insatisfeitos e irresponsáveis que não sabem o que querem?" (2015, p. 427-28). Ele se refere a nós, humanos, que apesar de sermos muito poderosos e ambiciosos, não conseguimos saber o que queremos para o futuro deste mundo em que vivemos e não nos damos conta do estado e do grau de vulnerabilidade em que estamos vivendo.

A hipótese do Mundo Vulnerável – *Vulnerable World Hypothesis* – também foi formulada por Bostrom (2019, p. 457) e professa, em linhas gerais, que se o desenvolvimento tecnológico mantiver o ritmo atual, um conjunto de capacidades será alcançado, em algum momento, que tornará a devastação da civilização provável, a menos que esta consiga sair, de forma suficiente, do padrão semianárquico em que se encontra. Sobre os problemas a serem enfrentados, ele aponta: (1) capacidade limitada de políticas preventivas; (2) capacidade limitada para governança global; (3) motivações diversas.

De modo geral, a renúncia aos avanços tecnológicos parece improvável – pois isso exigiria cessar a atividade inventiva e criativa em todo o mundo – e isso não é uma hipótese realista (*ibid.*, p. 462). Bostrom propõe, então, que haja melhor direcionamento do progresso científico e tecnológico, além da adoção de contramedidas possíveis que possam mitigar a vulnerabilidade civilizacional. Ele aponta, por exemplo: evitar que informações perigosas se espalhem; restringir o acesso aos materiais, instrumentos e a infraestrutura; dissuadir potenciais malfeitores, aumentando as chances de serem descobertos e punidos; maior cautela e trabalho de avaliação de riscos; vigilância e mecanismos de fiscalização que tornariam possível interditar tentativas de levar a cabo atos destrutivos (*ibid.*, p. 464). Ideias como essas precisam começar a ser levadas a sério e discutidas em nível global e também local, porque é notório que os riscos à vida do e no planeta são iminentes e que as consequências podem ser-nos fatais.

Em *Ciência, Razão e Paixão* (2009, p. 104), Prigogine afirma que a mais importante característica da vida, em geral, é a preocupação com o futuro e que essa preocupação atinge o ápice com a espécie humana, pois sua previsão desempenha papel central em suas decisões. No entanto, as tentativas de previsão, no geral, obedecem a um modelo linear e de certa maneira determinista, o que é o contrário do que se pode observar quando analisamos o passado e o presente da humanidade. O autor diz,

> A sociedade é inteiramente não linear pois aquilo que eu faço influencia o que os outros fazem e vice-versa. Na verdade, produzir modelos não-lineares é, porém, algo mais difícil do que modelizar supondo-se uma evolução linear. A emergência da não-linearidade fica bastante clara no momento das crises. As crises são um efeito da não-linearidade. Quanto mais complexa for a sociedade, mais importantes são os efeitos não-lineares, mais numerosos os pontos de bifurcação. (PRIGOGINE, 2009, p. 105)

Parece que, para lidar responsiva e responsavelmente com as questões emergentes que o desenvolvimento tecnocientífico tem-nos colocado, precisamos explorar e ampliar nossas alternativas, a partir de novos modelos não-lineares de convivência com eles. No entanto, para pensarmos o futuro que queremos, e não o aceitar como uma mera consequência inevitável de nosso semianárquico presente, devemos, então, imaginá-lo diferente. Afinal, a imaginação, a criatividade e a engenhosidade sempre foram nossas capacidades mais admiráveis e o nosso poder mais contundente.

Referências

BAGHRAMIAN, Maria; CARTER, J. Adam. Relativism. *In*: ZALTA, Edward N. (ed.). *Stanford Encyclopedia of Philosophy* (Spring 2021 edition). Disponível em: plato.stanford.edu/archives/spr2021/entries/relativism/. Acesso em: jul. 2021.

BOSTROM, Nick. Are you living in a computer simulation? *Philosophical Quarterly*, Oxford, v. 53, n. 211, p. 243-255, 2003. Disponível em: simulation-argument.com/simulation.html. Acesso em: jul. 2021.

BOSTROM, Nick. The vulnerable world hypothesis. *Global Policy*, Durham, v. 10, n. 4, 2019. Disponível em: nickbostrom.com/papers/vulnerable.pdf. Acesso em: jul. 2021.

CURD, Patricia. Presocratic philosophy. *In*: ZALTA, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy* (Fall 2020 edition), Disponível em: plato.stanford.edu/archives/fall2020/entries/presocratics/. Acesso em: jul. 2021.

DESCARTES, René. *Descartes* (Coleção Os Pensadores). São Paulo: Abril Cultural, 1999.

FAGAN, Kaylee. A viral video that appeared to show Obama calling Trump a 'dips' shows a disturbing new trend called 'deepfakes'. *Insider*, Apr 17, 2018. Disponível em: businessinsider.com/obama-deepfake-video-insulting-trump-2018-4. Acesso em: 15 jul. 2021.

FRASSEN, Maarten; LOKHORST, Gert-Jan; POEL, Ibo van de. Philosophy of technology. *In*: ZALTA, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy* (Fall 2018 edition). Disponível em: plato.stanford.edu/archives/fall2018/entries/technology/. Acesso em: jul. 2021.

HAACK, Susan. The unity of truth and the plurality of truths. *Principia*, v. 9, n. 1-2, 2005, p. 87-109. Disponível em: Dialnet-TheUnityOfTruthAndThePluralityOfTruths-5251200.pdf. Acesso em: jul. 2021.

HAACK, Susan. *Filosofia das lógicas*. Tradução: Cezar Augusto Mortari, Luiz Henrique de Araújo Dutra. São Paulo: Editora UNESP, 2002.

HARARI, Yuval. *Sapiens*: Uma breve história da humanidade. Tradução: Janaína Marcoantonio, 3. ed. Porto Alegre: L&PM, 2015.

JAMES, William. *Pragmatism*: a new name for some old ways of thinking. New York, NY: Barnes & Noble, 2003.

MCDONALD, Glen. We are not living in a simulation. Probably. *Fast Company*, 03/13/2018. Disponível em: fastcompany.com/40537955/we-are-not-living-in-a-simulation-probably. Acesso em: jul 2021.

PEDERSEN, Nikolaj; LINDING, Jang Lee; WRIGHT, Cory. Pluralist theories of truth. *In*: ZALTA, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Disponível em: plato.stanford.edu/archives/win2018/entries/truth-pluralist/. Acesso em: jul. 2021.

PLATÃO. *A República*. Tradução: Anna Lia Amaral de Almeida Prado, introd. Roberto Bolzani Filho. São Paulo: Martins Fontes, 2006.

PRIGOGINE, Ilya. *Ciência, razão e paixão*, org. Edgard de Assis Carvalho, Maria da Conceição de Almeida, 2. ed. rev. e ampl. São Paulo: Editora Livraria da Física, 2009.

PRIGOGINE, Ilya; STENGERS, Isabelle. *A nova aliança*: a metamorfose da ciência. Brasília: Editora Universidade de Brasília, 1984.

RADCLIFF, Edmonds. Plato's virtual reality. *IAI Magazine*, v. 91, 08/09, 2020. Disponível em: iai.tv/articles/platos-virtual-reality-auid-1632. Acesso em: 18 jul. 2021.

SAMUEL, Sigal. Is biohacking ethical? It's complicated. A new Netflix Series explain why. *Vox*, Oct. 24, 2019. Disponível em: vox.com/future-perfect/2019/10/22/20921302/netflix-unnatural-selection-biohacking-crispr-gene-editing. Acesso em: 20 jul. 2021.

VAIANO, Bruno. Agora você pode editar o DNA de bactérias em casa. *Superinteressante*, 25/05/2018. Disponível em: super.abril.com.br/ciencia/agora-voce-pode-editar-o-dna-de-bacterias-em-casa/. Acesso em: 17 jul. 2021.

COELHO, Patrícia Margarida Farias; HILDEBRAND, Hermes Renato. Estratégias de criação de deepfake: uma análise semiótica. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 119-134.

Recebido em: 2 mar. 2021
Aprovado em: 31 mar. 2021

dx.doi.org/10.23925/1984-3585.2021i23p119-134

Estratégias de criação de deepfake:

uma análise semiótica

Patrícia Margarida Farias Coelho[1]

Hermes Renato Hildebrand[2]

Resumo: Em um cenário marcado pela expansão da internet, a população cada vez mais vai mudando seus hábitos e migrando suas ações para a rede. Com o aumento de informações trocadas nas redes sociais, as deepfakes também se proliferam, tornando-se um perigoso instrumento de persuasão e provocando consequências negativas. O presente artigo investiga esse fenômeno a partir dos estudos da semiótica francesa. Temos dois objetivos: (i) compreender como as estratégias discursivas são produzidas nas deepfakes e (ii) realizar uma análise do percurso gerativo de sentido tripartido nos níveis narrativo, discursivo e fundamental. A metodologia aplicada é a de caráter descritivo e teórico, conforme prevê a tradição semiótica. Como resultado, compreende-se que a deepfake é um fenômeno em ascendência, pois está intrinsecamente associada às novas práticas possíveis do universo digital, divulgando conteúdos que parecem ser verdadeiros, mas não são.

Palavras-chave: Semiótica francesa. Sentido. Deepfake.

1 Mestra em Letras (Universidade Presbiteriana Mackenzie), Doutora em Comunicação e Semiótica pela Pontifícia Universidade Católica de São Paulo (PUC-SP). É coordenadora e professora permanente no Programa de Mestrado em Ciências Humanas da Universidade Santo Amaro e Professora do Programa de Mestrado e Doutorado em Educação da Universidade Metodista de São Paulo (UMESP desde 2018). ORCID: orcid.org/0000-0002-1662-1173. CV Lattes: lattes.cnpq.br/1087625657694882. E-mail: patriciafariascoelho@gmail.com.

2 Graduado em Matemática, Mestre em Multimeios pela UNICAMP e Doutor em Comunicação e Semiótica pela PUC-SP. Membro do Coletivo Artístico - SCIArts – Equipe Interdisciplinar. Professor das disciplinas relacionadas à Arte e Tecnologia, Pensamento Computacional, Tecnologias Emergentes, Semiótica e Propaganda e Marketing na UNICAMP e PUCSP. ORCID: orcid.org/0000-0002-3714-6295. CV Lattes: lattes.cnpq.br/6263913436052996. E-mail: hrenatoh@gmail.com.

Deepfake creation strategies: a semiotic analysis

Abstract: In a scenario marked by the expansion of the internet, newsreaders are changing their habits and migrating their actions to the Internet. With the increase of information exchanged on social networks, deepfakes are becoming a dangerous instrument of persuasion. The study investigates this phenomenon from the perspective of French semiotics. It has two objectives: (i) To understand how the discursive strategies are produced in deepfakes. (ii) To carry out an analysis of the generative path of tripartite meaning at the narrative, discursive, and fundamental levels. The methodology applied is descriptive and theoretical in character, as foreseen by the French semiotic tradition. It is understood that deepfake is a phenomenon on the rise because it is intrinsically associated with new possible practices through the digital universe, disseminating contents that seem to be true but are not.

Keywords: French semiotics. Sense. Deepfake.

Introdução

A internet conecta diariamente milhões de pessoas e transforma-se cada vez mais em uma necessidade, tornando-se primordial nesse novo contexto, o da *Sociedade Digital* (CASTELLS, 2011), pois grande parte das atividades que eram realizadas presencialmente migraram para os ambientes virtuais das redes, como, por exemplo, a educação, que mudou do presencial para o ensino on-line, lojas físicas que migraram para o atendimento digital ou restaurantes que só serviam refeições no local e agora passaram a realizar entregas. Além disso, as reuniões entre pessoas, que anteriormente aconteciam presencialmente, agora são realizadas, em grande número, por meio das plataformas digitais, como Zoom, Meet, Whatsapp. As relações sociais tornaram-se mais líquidas (BAUMAN, 2003), não no sentido da superficialidade, mas em relação ao contato entre esses indivíduos, que se tornou cada vez menos frequente.

A partir desse novo panorama social, proliferaram-se nas redes sociais e na internet as fake news e as deepfakes. Ressaltamos que essas técnicas, atualmente utilizadas pela Inteligência Artificial, não são novidades, pois elas já existiam. No entanto, aumentaram e ganharam força, principalmente, com as ações da população que cada vez mais incorpora a internet ao cotidiano. Delmazo e Valente explicam que as fake news são:

> Notícias falsas, histórias fabricadas, boatos, manchetes que são risco de cliques (as chamadas clickbaits) não são novidade. Darnton (2017) relembra o surgimento dos pasquins, na Itália do século XVI, que se transformaram em um meio para difundir notícias desagradáveis, em sua maioria falsas, sobre personagens públicos. Também recorda o surgimento dos Canards, gazetas com falsas notícias que circularam em Paris a partir do século XVII. (DELMAZO; VALENTE, 2018, p. 156)

As fake news existem há muito tempo (SANTAELLA 2018, 2020), mas com a facilidade de disseminação permitida pelo ambiente on-line, as circulações de notícias falsas e o anonimato permitido pelas redes fizeram essa forma de comunicação aumentar significativamente, produzindo falsas informações, particularmente nos dias de hoje, sobre a pandemia. Já as deepfakes, tema deste trabalho, incluem fake news em formato de vídeo (MORAES, 2019), caracterizando-se, segundo informação da jornalista Isabela Cabral, no website da *Tech Tudo*, como

> uma tecnologia que usa inteligência artificial (IA) para criar vídeos falsos, mas realistas, de pessoas fazendo coisas que elas nunca fizeram na vida real. A técnica que permite fazer as montagens de vídeo já gerou desde conteúdos pornográficos com celebridades até discursos fictícios de políticos influentes. Circulam agora debates sobre a ética e as consequências da tecnologia, para o bem e para o mal. (Cabral, 2018)

Moraes, complementando as ideias de Cabral, explica que as deepfakes são técnicas que visam substituir o rosto de uma pessoa por outra, em um vídeo. Segundo a autora:

> Em questão de datas, o primeiro ocorreu no Outono de 2017 utilizado para gerar conteúdos adultos. Posteriormente, essa técnica foi melhorada por uma pequena comunidade para criar nomeadamente uma aplicação chamada "FakeApp". O processo para gerar deep fake consiste em imagens que reúnem rostos alinhados de duas pessoas diferentes, nas quais há a reconstrução do rosto de uma em conjunto de dados de imagens faciais das outras e se autocodifica para então reconstruir rostos com as imagens faciais. Na prática, os resultados são impressionantes, o que explica a popularidade da técnica. O último passo é levar o vídeo ao alvo, extrair e alinhar a face do alvo a partir de cada quadro, utilizando software ou aplicativos "FaceApp". (MORAES, 2019, p. 3)

A hipótese que sustenta este trabalho é a de que a possibilidade de invisibilidade permitida pelos algoritmos faz com que as deepfakes se multipliquem na rede, pois permitem a *ilusão de impunidade* aos desenvolvedores que distribuírem esses conteúdos falsos.

A partir dessa contextualização, são propostos dois objetivos a serem alcançados neste estudo, a saber: (i) compreender como as estratégias discursivas são produzidas nas deepfakes e (ii) realizar a análise do plano do conteúdo nos níveis narrativo, discursivo e fundamental, a fim de verificar quais são os valores que elas disseminam.

O *corpus* utilizado nesta pesquisa é o vídeo intitulado *Você não vai acreditar no que o Obama disse!* (SILVERMAN, 2018). A justificativa para selecionar esse vídeo deve-se, especificamente, ao fato de ele permitir: (i) compreender o que é uma deepfake e (ii) como ela pode influenciar o comportamento dos internautas.

O vídeo que compõe o nosso *corpus* foi publicado no site *BuzzFeed. News* e divulgado pelo *media editor* Craig Silverman (2018). Lembramos, ainda, esse vídeo não é exatamente uma deepfake, e, sim, uma apresentação das estratégias utilizadas na criação desse tipo de produção, que é o que nos interessa compreender neste estudo.

Nesta pesquisa, utilizamos como arcabouço teórico os estudos sobre semiótica francesa alicerçados em Greimas (2014), Greimas e Courtés (2008), Fiorin (1999, 2016) e Barros (2005, 2015, 2019, 2020). A proposta de trabalharmos a partir dos estudos da semiótica discursiva parte, principalmente, de duas propostas, a saber: em primeiro lugar, por Greimas (2014) se debruçar a refletir sobre *a verdade*, pois para o pesquisador a verdade é um constructo:

> eminentemente contratual e decorre da fidúcia – "[...] toda comunicação humana, toda tratativa, mesmo que não verbal, está fundada sobre um mínimo de confiança mútua e que ela vincula os protagonistas ao que chamamos contrato fiduciário" (GREIMAS, 2014, p. 134). Ela, portanto, é construída em cada texto, sem que isso diminua sua força. Nessa linha, a semiótica discursiva filiou-se às correntes que derrubaram o edifício da verdade universal e ontológica, e até mesmo da verossimilhança, como confirmação de um referente externo para reconhecer na crença o pressuposto de todo saber. "Percebeu-se [...] que o eu penso que, que serve de suporte para o discurso interior do sujeito, quando este quer exteriorizá-lo, não é um 'eu sei', mas um 'eu creio'. [...] o saber dito científico seria apenas um parêntese ou [...] um efeito de sentido que se constitui em condições a serem determinadas" (GREIMAS, 2014, p. 128). A semiótica não se ocupa do que é a verdade, mas da maneira como o verdadeiro se faz. (KALIL FILHO, 2019, p. 208)

Em segundo lugar, de acordo com Barros (2019, p. 126), é fundamental buscar compreender as "questões de veridicção e de modalização pelo *saber* e pelo *crer* na produção e interpretação de fake news". Ainda que a autora dedique seu olhar a refletir sobre as fake news, alicerçamo-nos em suas pontuações para aplicá-las às deepfakes. A pesquisadora explica, por meio da semiótica francesa, questões sobre a verdade dos discursos nas pesquisas da modalização (modalização pelo saber e pelo crer, modalização veridictória):

> Na veridicção, as relações modais entre o ser e o parecer, que determinam os discursos como verdadeiros, mentirosos, secretos ou falsos, e levam seus destinatários a neles acreditar ou não, têm na internet características próprias. Se os textos de "histórias de pescador" são, por definição, interpretados como falsos, isto é, que nem parecem nem são verdadeiros, os da internet são, em geral, considerados verdadeiros, ou seja, que parecem e são verdadeiros, e, mais do que isso, que eles são discursos que desmascaram a mentira ou revelam o segredo. O destinatário de discursos interpreta-os a partir de seus conhecimentos e crenças e da capacidade de persuasão do destinador-manipulador. No caso da internet (BARROS, 2015), a interpretação como discurso verdadeiro e também o desmascaramento da mentira e a revelação do segredo decorrem do efeito de sentido de grande quantidade de saber armazenado pela internet ("que sabe tudo") e do de interatividade acentuada, pois, com esses atributos de seu discurso, o desti-

> nador é colocado, pelo destinatário interpretante, na posição de sujeito do saber. Soma-se a esses procedimentos de persuasão e interpretação dos discursos na internet, o de o destinatário, devido à interatividade intensa já mencionada, deles se considerar, em boa parte, também como "autor-destinador". Assim construído, o destinatário acredita e confia nos discursos que também são "seus". A pós-verdade pode ser entendida, no quadro da veridicção, como resultante de interpretação baseada, sobretudo ou apenas, nas crenças e emoções do destinatário interpretante. Dessa forma, por mais absurdos que pareçam, os discursos cujos valores estiverem de acordo com as crenças e sentimentos do destinatário serão por ele considerados verdadeiros. (BARROS, 2020, p. 126-127)

É a partir dessa perspectiva teórica, fundamentada na semiótica francesa, que esta pesquisa se desenvolve. Dessa forma, entendemos que a deepfake é uma técnica perigosa, pois propaga, por meio de imagens e sons, mensagens *não verdadeiras* que comprometem as informações distribuídas nas redes, como, por exemplo, segurança, educação, saúde, economia e, especialmente, políticas, dentre outras. Por isso, é importante estudá-la compreendendo suas características e valores concretizados nas redes sociais digitais.

As estratégias discursivas da deepfake: mentira, verdade, segredo ou falsidade?

Para evidenciar as estratégias discursivas propostas pela semiótica francesa, investigamos neste tópico três imagens recortadas do vídeo *Você não vai acreditar no que o Obama disse!* (SILVERMAN, 2018), que auxiliam a nossa discussão. Os recortes dessas imagens devem-se a uma estratégia metodológica para apresentar a sequência das cenas que aparecem no vídeo. Ressaltamos mais uma vez que as imagens retiradas deste vídeo não são exatamente uma deepfake, mas sua escolha para compor o *corpus* se deve ao fato de elas nos permitirem identificar as estratégias usadas para a criação desse tipo de conteúdo.

Selecionamos o recorte dessas imagens porque elas nos oferecem uma gama de sentidos que não estão restritos à materialidade linguística, e convocam o internauta a se posicionar diante do que está sendo apresentado, uma vez que "toda comunicação, publicidade, vídeo é manipulação" (COELHO; COSTA, 2013, p. 111), que direciona o usuário a acreditar na *verdade* que lhe está sendo oferecida.

Figura 1. Disponível em: buzzfeed.com/br/craigsilverman/como-identificar-deepfake-video-obama-peele. Acesso em: 9 jun. 2021.

Figura 2. Disponível em: buzzfeed.com/br/craigsilverman/como-identificar-deepfake-video-obama-peele. Acesso em: 9 jun. 2021.

Figura 3. Disponível em: buzzfeed.com/br/craigsilverman/como-identificar-deepfake-video-obama-peele. Acesso em: 9 jun. 2021.

Nas Figuras 1 e 2 destacamos a fala do protagonista, o ex-presidente dos Estados Unidos Barack Obama, falando sobre o também ex-presidente dos Estados Unidos Donald Trump. Já na Figura 3, próxima ao final do vídeo, a imagem se divide em dois planos, com Barack Obama e o cineasta Jordan Peele. Barros (2016, p. 12) explica que os "discursos da internet são, em geral, considerados verdadeiros, ou seja, que parecem e são verdadeiros, e, mais do que isso, que eles são discursos que desmascaram a mentira ou revelam o segredo". Segue a transcrição do vídeo realizada pelos autores:

> Estamos entrando em uma Era na qual nossos inimigos podem fazer com que qualquer um pareça estar dizendo qualquer coisa a qualquer momento, mesmo que eles nunca tenham dito isso. Então, por exemplo, poderiam me fazer dizendo coisas como... Não sei... "Killmonger estava certo", ou, "Ben Carson está no lugar profundo do 'Corra!', ou, que tal, simplesmente: "O presidente Donald Trump é um total e completo merda". Agora, vejam, eu nunca diria essas coisas. Pelo menos não em um discurso público, mas outra pessoa o faria. Alguém como Jordan Peele. Este é um momento perigoso. Ao avançar precisamos ficar atentos com aquilo que acreditamos na internet. É uma época em que precisamos contar com fontes confiáveis de notícia. Pode soar básico, mas a forma com a qual avançamos na Era da Informação fará a diferença entre nós sobrevivermos, ou nos tornarmos um tipo de distopia. Obrigado, e fiquem atentos, seus putos. (SILVERMAN, 2018)

A partir da afirmação da pesquisadora, verificamos que a deepfake busca criar a ilusão de verdade (FIORIN, 1999) por meio da estratégia de *parecer* que o discurso *parece e é verdadeiro*. A verdade, para Greimas (2014, p. 122), é um "*fazer-parecer-verdadeiro*", pois "A verdade em si não existe, ela é uma (re)criação da interação entre o enunciador, aquele que enuncia, e o enunciatário, aquele que interpreta. Estes são pares intercambiáveis e coautores do enunciado produzido" (COELHO; COSTA; FONTANARI, 2015, p. 7). Completando as afirmações dos pesquisadores, Kalil Filho (2019, p. 212) explica:

> A verdade no discurso não é mais da ordem da verossimilhança ou da verdade ontológica, mas anui à veridicção, uma "operação que se exerce como um saber sobre os objetos (do mundo)" (p. 87). Para o "juízo epistêmico definitivo" (GREIMAS; COURTÉS, 2008, p. 533), a certeza e a verdade devem ser delineadas na imanência, no interior do texto e na relação entre enunciador e enunciatário. Mais do que isso, o enunciatário deve julgar o estatuto da imanência, do "ser" do objeto, em seu contato com a manifestação, o "parecer" do objeto. (KALIL FILHO 2019, p. 212)

Dessa maneira, verificamos, na Figura 3, que o desenvolvedor desse vídeo revela, com estratégias discursivas (verbo-visuais), a *mentira* por meio da imagem e das palavras, mostrando os efeitos de sentido e as estratégias sincréticas que *podem* e que são *criadas* por meio das deepfakes divulgadas nas redes, compartilhando informações falsas.

Portanto, a Figura 3 evidencia, conforme informações disponibilizadas no site (SILVERMAN, 2018), que o vídeo foi produzido utilizando um vídeo original com a fala de Barack Obama, em que o ator Jordan Peele sobrepôs outro discurso ao do ex-presidente. O sucesso dessa deepfake deveu-se ao fato de que quanto mais tempo o vídeo ficou no ar, mais tempo houve para processar a junção da cabeça de Obama com a fala de Jordan Peele. Essa união entre imagem e fala é que faz parecer que o vídeo *pare*ça verdadeiro (FIORIN, 2016).

Greimas (2014, p. 123) explica que quando a verdade prescinde de um referente externo, ela se torna um *dizer verdadeiro*, ou seja, basta *parecer verdadeiro* para que *verdadeiro seja*. É por isso que parte da população, ao ver o vídeo, reconhece-o como legítimo, pois esse tipo de fala, que aparenta ser realmente do ex-presidente, direciona o destinatário a crer no conteúdo compartilhado. Contudo, sabe-se que Barack Obama jamais realizaria aquelas afirmações em público, ainda que acreditasse na *verdade* dessa afirmação.

Portanto, ressaltamos que não é a internet que compartilha conteúdo falso e enganoso, ou que causa desinformação e problemas, mas, sim, os discursos construídos e divulgados nas deepfakes compartilhadas nas redes, com imagens e sons (falas) que são, de alguma forma, reconhecidos pelos internautas como verdadeiros. As inverdades das deepfakes que circulam na rede, em geral, causam desinformações aos destinatários, fazendo-os crer que o conteúdo divulgado é real.

Kalil Filho (2019) defende que as modalidades veridictórias inscrevem-se no texto na relação entre o fazer persuasivo e o fazer interpretativo, uma vez que, "a verdade é um julgamento em que a manifestação parece ser e a imanência é. Um texto será falso quando não parece ser e não é. O objeto será mentiroso se parece ser, mas não é" (KALIL FILHO, 2019, p. 212).

Assim sendo, podemos entender que a deepfake selecionada para este artigo se apresenta como um segredo – não parece, mas é. Por meio dessas reflexões, verificamos que *a verdade* de um discurso é definida pela veridicção (BARROS, 2019, 2020), ou seja, pelas estratégias, procedimentos e recursos utilizados para construir um determinado discurso, nesse caso, o vídeo.

Percurso gerativo de sentido: um olhar panorâmico

a. Nível narrativo

Barros (2005, p. 11) explica que "no nível das estruturas narrativas, os elementos das oposições semânticas fundamentais são assumidos como valores por um sujeito e circulam entre sujeitos graças à ação também de sujeitos". No primeiro tópico, tratamos das Figuras 1 e 2, referentes ao vídeo, analisando como foi construída a narrativa de um sujeito (ex-presidente Barack Obama), manipulando outro sujeito (internauta) por tentação, pois trata-se de um manipulador *forte*, o ex-presidente dos Estados Unidos, que afirma sobre devermos tomar cuidado com os conteúdos distribuídos na internet. O sujeito-internauta é representado pelos eleitores e simpatizantes do político.

Os internautas que aceitaram e aceitam a proposta de Barack Obama são reconhecidos como *corajosos* e *não manipuláveis pela grande mídia*. Barros (2005, p. 20) esclarece ainda que *o programa narrativo* ou sintagma elementar da sintaxe narrativa define-se como um enunciado de fazer, que rege um enunciado de estado. Assim, constroem-se diferentes programas narrativos: (i) aquisição, ou (ii) privação.

Natureza da função	Denominação	Exemplo
Aquisição	Doação	1° O ex-presidente Barack Obama doa o objeto de valor ao internauta
Privação	Espoliação	2°: O ex-presidente Barack Obama não doa o objeto de valor ao internauta, pois o internauta pertence a outro seguimento político.

Quadro 1. Relação entre as naturezas de funções. Elaborado pelos autores.

Neste estudo, temos o primeiro programa narrativo, o da aquisição, pois o internauta recebe do ex-presidente Barack Obama o objeto-de-valor *cuidado com os conteúdos vistos e compartilhados na internet* (o sujeito do fazer é o ex-presidente, que sugere ao internauta ter atenção sobre o tipo de conteúdo visualizado na rede de computadores; o sujeito de estado tem sua situação alterada, caso aceite como verdade a fala do ex-presidente e passe a se atentar sobre o que compartilha e acredita dos vídeos e conteúdos encontrados na internet.

No percurso do *destinador-manipulador*, encontramos no vídeo, deepfake, Barack Obama (destinador) doando a *competência modal*, ou seja, do *querer-fazer*, do *dever-fazer*, do *saber-fazer* e do *poder-fazer* durante a manipulação. No entanto, para que possa ocorrer a doação de competência modal, é necessário que o destinatário-sujeito *acredite* e *aceite* o contrato oferecido pelo destinador.

Se os internautas simpatizantes do ex-presidente Barack Obama *acreditam,* por exemplo, na *verdade propagada* pelo vídeo, e aceitam o contrato como verdadeiro, eles passam a ter *cuidado com os conteúdos vistos e compartilhados na internet*. O perigo da deepfake está em compartilhar um conteúdo manipulado que possa prejudicar uma boa parcela da população. Nas cenas do vídeo, verificamos que ele foi criado para mostrar *as marcas* de inverdades nele contidas. No entanto, nem sempre é ou será tão fácil reconhecer essas marcas nas narrativas de outros vídeos de deepfake, por exemplo.

b. Nível discursivo

Fiorin (1999, 2016) pontua que nesse nível se analisa como o texto utiliza categorias como tempo e espaço, e como os elementos narrativos são concretizados com os temas e figuras no nível discursivo. Barros (2015, 2019, 2020) explica que todo enunciado apresenta um *sujeito da enunciação* que se desdobra em um enunciador falando a um enunciatário.

No vídeo selecionado como *corpus*, temos um enunciador, ou seja, a voz e a imagem do ex-presidente dos Estados Unidos, que fala aos enunciatários-internautas para tomarem consciência sobre o conteúdo acessado e disponibilizado na rede. Nas cenas do vídeo (Figuras 1 e 2), verificamos uma semelhança não só na voz, mas também na forma de se expressar, bem semelhante à utilizada pelo ex-presidente Barack Obama, que utiliza uma *embreagem enunciativa (eu-aqui-agora)*, pois o vídeo traz Barack Obama falando em primeira pessoa, Eu, aqui (agora-Estados Unidos), no tempo presente, para criar o efeito de *verdade e aproximação* com o que está sendo dito.

Barros (2020) explica que essa fala similar ao do ex-presidente dos Estados Unidos tem uma intencionalidade explícita, pois busca levar o enunciatário a acreditar no que está sendo dito, uma vez que o discurso aparenta ter credibilidade, já que o sujeito que fala é *confiável,* e foi o representante de um dos países mais importantes do mundo. Dessa forma, no vídeo há uma ancoragem actancial, temporal e espacial explícita, uma vez que houve a indicação exata de pessoa, tempo e espaço no discurso, permitindo a credibilidade dos fatos narrados.

A imagem e a voz, semelhantes às de Barak Obama, utilizadas em outros canais de comunicação para compor o vídeo, ratificam valores e ideologias transmitidos pelo ex-presidente, de um homem íntegro e comprometido com os valores propagados pelos meios de comunicação, que auxiliam a convencer os enunciatários-internautas de que o vídeo é verdadeiro.

Barros (2005, p. 206) explica que o uso de temas e figuras são "enriquecimentos semânticos empregados para dar o acabamento estético almejado pelo enunciador". No vídeo em análise, encontramos a leitura temática, a saber: tema sobre a propagação de conteúdo falso, enganoso e mentiroso compartilhado na internet, por meio da manipulação do uso de sons e imagens de celebridades do aplicativo *FakeApp*, como nos é revelado a partir da Figura 3.

c. Nível fundamental

Greimas e Courtés (2008) explicam que é no nível das estruturas fundamentais que se encontram as oposições semânticas do texto que parte das estruturas mais simples, e que vão para as mais complexas. É nesse nível que o quadrado semiótico se situa. Fiorin (1999) explica que o quadrado semiótico se caracteriza como a representação visual da articulação lógica de qualquer categoria semântica, partindo da oposição de pelo menos dois termos que se estabelecem por uma combinatória das relações de contradição e asserção. No vídeo estudado, temos o *Ser* (ex-presidente) em oposição ao *Parecer* (ex-presidente), conforme representado por meio do quadrado semiótico:

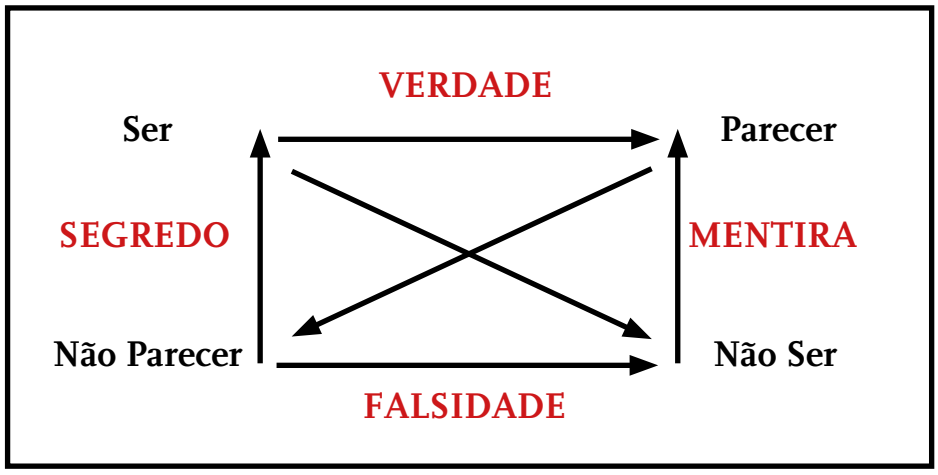

Figura 4. Articulações veridictórias sobre o Quadro Semiótico. Fonte: Greimas (2014, p. 66). Relação entre as naturezas de funções. Elaborado pelos autores.

Baldan (1999) esclarece que somente entendemos as noções de *verdade* ou *mentira* em um texto quando o compreendemos (lemos) a partir de uma prática social, pois um texto somente faz sentido para um sujeito na medida em que este está relacionado com os enunciados. Dessa forma, a pesquisadora defende que *verdade* e *mentira* são os efeitos de sentido construídos em todo e qualquer ato de interpretação discursiva, uma vez que:

> interpretar implica apreender um sentido enquanto saber produzido pelo discurso-enunciado, tal como ele emerge da cooperação dos dois fazeres implicados no ato da enunciação: o do enunciador e o do enunciatário; desse modo, a mensagem surgirá como o lugar de uma prática significante, o espaço em que ocorre um ato enunciativo que, visto do pólo do enunciador, produz o discurso, a unidade semiótica dotada de um fazer informativo – um fazer saber –, mas que, quando visto do pólo do enunciatário, se manifesta como um texto, unidade semiótica dotada de um fazer interpretativo, produtora de um fazer saber sobre aquele fazer informativo. (BALDAN, 1999, p. 49)

Assim sendo, no vídeo, a oposição se manifesta na voz e na imagem do ex-presidente dos Estados Unidos, que nos possibilita compreender que:

> (a) verdade – "aquilo que é e que parece ser isso que é" (produção do saber autêntico).
> (b) falsidade – "aquilo que nem é (o que é) nem parecer ser (isso que é)" (produção do não-saber).
> (c) mentira – "aquilo que parece ser (o que é) mas não é" (produção de simulação do saber-parecer saber).
> (d) segredo – "aquilo que é (o que é) mas não parece ser" (produção de dissimulação do saber-parecer não-saber). (BALDAN, 1999, p. 51)

Por meio dessas reflexões, verificamos o discurso *mentiroso* do vídeo, pois o protagonista *parece* ser, na imagem, na voz e no discurso, *mas não é* o ex-presidente Barack Obama e, sim, sua imagem com a voz e a fala manipuladas pelo ator Jordan Peele. Portanto, verificamos que a deepfake busca nos convencer de que o vídeo *parece* e é *verdadeiro*, pois utiliza recursos visuais, sonoros e verbais que se assemelham à *realidade* que a sociedade (re)conhece compartilhada *nas* e *pelas* mídias sobre esse ex-governante.

Considerações finais

Nesta pesquisa, refletimos sobre como as deepfakes, por meio da Inteligência Artificial, constroem estratégias discursivas mentirosas com o intuito de enganar os sujeitos-internautas, utilizando o verbal, o visual e o sonoro para criar vídeos com conteúdos que *parecem,* mas não são

verdadeiros. Nossa hipótese foi confirmada, pois as deepfakes se espalham na internet sem que se saiba quem foi o desenvolvedor do conteúdo e representam, ainda, um risco *sem precedente*s para a sociedade, já que esses *vídeos parecem* verdadeiros e seus desenvolvedores não apresentam ética ao compartilharem conteúdos enganosos.

Nossos dois objetivos foram alcançados. No primeiro tópico, compreendemos como as estratégias discursivas produzidas na deepfake revelaram ser mentira, verdade, segredo ou falsidade. Já no segundo, realizamos uma análise do plano do conteúdo, partindo do nível narrativo para os níveis discursivo e fundamental. Por meio do estudo semiótico, foram desveladas as estratégias de manipulação utilizadas na construção da deepfake para disseminar conteúdo enganoso nas redes sociais.

A semiótica francesa é uma teoria profícua para a análise de vídeos. No vídeo selecionado como *corpus* de nossa pesquisa, o discurso do ex-presidente evidencia a manipulação de imagens e sons criados pela ferramenta gratuita *FakeApp,* que permite criar vídeos utilizando rostos de famosos com as falas manipuladas. Foi possível, também, verificar por meio do discurso criado para o vídeo, a advertência do ex-governante sobre a legitimidade dos conteúdos disponibilizados na internet e a seriedade de se verificar a autenticidade do que é encontrado e visto na rede.

Concluímos que a deepfake divulga imagens e conteúdos comprometedores com recorrência a falas e discursos carregados de inverdades, trazendo sérias e perigosas consequências à sociedade. Finalizamos este trabalho, almejando que, em um futuro próximo, as deepfakes sejam menos compartilhadas, diminuindo os mecanismos que propagam e disseminam conteúdos mentirosos e que podem prejudicar grande parte da população.

Referências

BALDAN, Maria de Lourdes Ortiz Galdin. Veridicção: um problema de verdade. *Alfa – Revista de Linguística*, São Paulo, v. 32, p. 47-52, 1988. Disponível em: periodicos.fclar.unesp.br/alfa/article/view/3797/3505. Acesso em: 9 jun. 2021.

BARROS, Diana Luz Pessoa de. *Teoria semiótica do texto*. São Paulo: Editora Ática, 2005.

_____. A complexidade discursiva na internet. *Cadernos de Semiótica Aplicada*, São Paulo, v. 13, n. 2, p. 13-31, 2015. Disponível em: doi.org/10.21709/casa.v13i2.8028. Acesso em: 9 jun. 2021.

_____. Algumas reflexões sobre o papel dos estudos linguísticos e discursivos no ensino-aprendizagem na escola. *Estudos Semióticos*, v. 15, n. 2, p. 1-14, 2019. Disponível em: doi.org/10.11606/issn.1980-4016.esse.2019.165195. Acesso em: 9 jun. 2021.

_____. Ser uma semioticista ontem e hoje. *Revista Entrepalavras*, Fortaleza (CE), ano 10, v. 10, n. especial, p. 113-132, maio 2020. Disponível em: entrepalavras.ufc.br/revista/index.php/Revista/article/view/1797. Acesso em: 9 jun. 2021.

BAUMAN, Zygmunt. *Modernidade líquida*. Tradução de Plínio Dentzien. Rio de Janeiro: Zahar, 2003.

CABRAL, Isabela. O que é deepfake? Inteligência artificial é usada para fazer vídeo falso. *TechTudo* 28/07/2018. Disponível em: techtudo.com.br/noticias/2018/07/o-que-e-deepfake-inteligencia-artificial-e-usada-pra-fazer-videos-falsos.ghtml. Acesso em: 9 jun. 2021.

CASTELLS, Manuel. *A sociedade em rede*. Tradução de Roneide Venâncio Majer. São Paulo: Paz e Terra, 2011.

COELHO, Patrícia Margarida Farias; COSTA, Marcos Rogério Martins. Publicidade e contos de fadas: reflexões semióticas. *Acta Semiótica et Lingvistica*, João Pessoa, v. 18, n. 1, p. 110-124, 2013.

_____; FONTANARI, Rodrigo. O parecer do sentido: a perspectiva semiótica. *Razón y Palabra*, Monterrey, v. 92, p. 1-18, 2015.

DELMAZO, Caroline; VALENTE, Jonas C. L. Fake news nas redes sociais online: propagação e reações à desinformação em busca de cliques. *Media & Jornalismo*, Lisboa, v. 18, n. 32, p. 155-169, 2018.

FIORIN, José L. Sendas e veredas da semiótica narrativa e discursiva. *DELTA*, São Paulo, v. 15, n. 1, 1999. Disponível em: scielo.br/scielo.php?script=sci_arttext&pid=S0102-44501999000100009&lng=en&nrm=iso. Acesso em: 9 jun. 2021.

_____. *As astúcias da enunciação*. São Paulo: Ática, 2016.

GREIMAS, Algirdas J. *Sobre o sentido II*: ensaios semióticos. Tradução de Dilson Ferreira da Cruz. São Paulo: Nankin, Edusp, 2014.

_____; COURTÉS, Joseph. *Dicionário de semiótica*. Tradução de Alceu Dias Lima, et al. São Paulo: Contexto, 2008.

KALIL FILHO, Marcos da Veiga. Fake news e democracia: contribuições da semiótica discursiva acerca da verdade e da informação na internet. *Caderno Letras UFF*, Niterói, v. 30, n. 59, p. 205-219, 2019.

MORAES, Cristiane Pantoja. "Deepfake" como ferramenta manipulação e disseminação de "fake news" em formato de vídeo nas redes sociais. *IX Encontro Ibérico EDICIC*, Barcelona, 2019. Disponível em: doi.org/10.31219/osf.io/mf7t6. Acesso em: 9 jun. 2021. Disponível em: techtudo.com.br/noticias/2018/07/o-que-e-deepfake-inteligencia-artificial-e-usada-pra-fazer-videos-falsos.ghtml. Acesso em: 9 jun. 2021.

SANTAELLA, Lucia. *A pós-verdade é verdadeira ou falsa?* (Coleção Interrogações). Barueri: Estação das Letras e Cores, 2018.

_____. A semiótica das fake news. *Verbum – Cadernos de pós-graduação*, v. 9, n. 2, p. 9-25, 2020.

SILVERMAN, Craig. Como identificar um "deepfake" como este vídeo do Barack Obama. *BuzzFeed News Media* 19/04/2018. Disponível em: buzzfeed.com/br/craigsilverman/como-identificar-deepfake-video-obama-peele. Acesso em: 9 jun. 2021.

EXTRA DOSSIÊ

COLAPIETRO, Vincent; NÖTH, Winfried; CESTARI, Guilherme H. de O. NETO BITTENCOURT, Levy Henrique. How can we change habits? *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 136-155.

Recebido em: 1 fev. 2021
Aprovado em: 15 fev. 2021

dx.doi.org/10.23925/1984-3585.2021i23p136-155

How can we change habits?[1]

Reflections by Vincent Colapietro[2]

in dialogue with Winfried Nöth[3]

edited by Guilherme Henrique de Oliveira Cestari[4] and Levy Henrique Bittencourt Neto[5]

W.N.: Good afternoon, we are here to speak about a very difficult topic: how to change habits. Peter Sloderdijk launched a book, translated into English as "You must change your life".[6] We were more respectful and put it in the form of a question: How do we change habits? But the presupposition of this question is the same, you cannot ask how you change a habit if you are not convinced that you must change habits, and life is habits, is it not? Well, of course, the background of our dialogue is Peirce's conception of life and habits, but perhaps not only. We are here to meet and to speak in a series of online reflections. This is the third. A week from now we have a fourth. Without further ado, Professor Vincent Colapietro, whom you must know by now, is ready to speak about this topic, and we thank him very much for being with us once more.

1 The dialogue took place online on the channel @TIDDigital youtu.be/It_I1wip2G8, on September 18, 2020.

2 University of Rhode Island, Department of Philosophy, Kingston, RI, USA. ORCID: orcid.org/0000-0002-8411-9601.

3 Professor of the Postgraduate Program in Technologies of Intelligence and Digital Design (TIDD) at Catholic University of São Paulo (PUC-SP). ORCID: orcid.org/0000-0002-2518-9773. CV Lattes: lattes.cnpq.br/7221866306191176. E-mail: wnoth@pucsp.br.

4 PhD – TIDD (PUC-SP). ORCID: orcid.org/0000-0002-8411-9601. CV Lattes: lattes.cnpq.br/7987513983124155. E-mail: ghocestari@pucsp.br.

5 PhD – TIDD (PUC-SP). CV Lattes: lattes.cnpq.br/9788577794623417. E-mail: nikolai.streisky@gmail.com.

6 Peter Sloterdijk, *Du musst dein Leben ändern:* über Anthropotechnik, Frankfurt / Main: Suhrkamp, 2009. – English: *You must change your life:* on anthropotechnics, trad. Wieland Hoban, Cambridge: Polity Press, 2013.

V.C.: Thank you so much for this opportunity. The title of the book to which you refer actually quotes a line from Rainer Maria Rilke, if I recall correctly.[7] This implies that changing one's habits can extend all the way to changing one's life.

I want, however, to begin by talking about habit change in a very general way because it seems to me that the title of our exchange calls for that wider context. The *can* in our title points to the scope of our agency. *Can* we change our habits – does our agency extend to the alteration of our habits and, if so, how is this most effectively accomplished? The focus of our concern is really the *deliberate* alteration of habits. We, as agents, are oftentimes in some ways rather severely restricted by bad habits and, as a result, we want to change those habits. For example, somebody might want to quit smoking, or somebody might want to correct their posture. – I do not sit properly and this causes me backaches, so I have tried, with limited success, to alter my somatic dispositions. The focus on the deliberate alteration of habits is of course very important, but I want, at the outset, to step back from that specific topic and spend some time considering more generally the processes of habit change.

Habits are in fact changing all the time. Even when a habit remains the same, it is in some sense changing. I do not mean to be unduly paradoxical here. All I mean by this seeming paradox is that, when the exercise of a habit is fluid and unimpeded, what happens – in however imperceptible or slight a way – is that the habit is strengthened. In a sense, the habit does not remain the same. It is getting stronger, however imperceptibly. In general, then, habits are either strengthening or weakening. They are never staying absolutely the same, no matter how much it might look that way.

Two other points about habit change in general are especially pertinent here. First, habits often change willy-nilly; they just change because of the spontaneity of firstness, among other things. While habit change might result from deliberation, it mostly happens without intention or even consciousness. Habit change is an inescapable fact about the natural world. Deliberate alteration of habits is only a very small subset of habit change. Second, habit change need not be a function of repetition. Often it is, though not always. In fact, a single event might generate a deeply entrenched habit. Think here of a traumatic event, specifically, think about a very young child being attacked by a vicious dog. A single event,

7 The line is from Rilke's sonnet "Archaic Torso of Apollo", first published in 1908. Sloterdijk discusses the poem in the first chapter of his book. See also: Rachel Corbett, *You must change your life:* the story of Rainer Maria Rilke and Auguste Rodin, New York, NY: Norton, 2016.

the attack of the child by the dog, might cause the child to have a lifelong fear of dogs. This fear is at bottom a disposition to respond to certain animals, in certain ways. The child might be so traumatized that they are unconsciously and automatically fearful in the presence of *any* dog. In sum, habit change is inexorable, often unintended, and not always the result of repetition.

As we have just noted, the formation of a habit might be generated by a single event; a traumatic event would be an example of that. As we have also stressed, habits are always changing, even when they appear to remain the same. As important as outward or overt action as well as unanticipated and especially traumatic experience may be, they do not constitute the whole story about habit change. Dramatic imagination can play a critical role here (dramatic imagination being the capacity to imagine a scene –e.g., a dress accidentally catching on fire – and then envisaging various responses to a dramatic situation). There is a very interesting anecdote that Peirce recalls[8] (one I quote it at the beginning of chapter five in my book on *Peirce's Approach to the Self*).[9] It is an episode that is an actual recollection of his childhood, which involves his younger brother, who was very young at the time.[10] I am not sure how old Herbert was at the time, though I would say he was probably six, seven, or eight. Peirce's brother Herbert,[11] who went on to become a diplomat and was for part of his career a United States ambassador in Oslo (at the time, Christiana). Peirce's household was the center of not merely the scientific community, but the literary in the cultural community as well. It was one of the most important gathering places in Cambridge, Massachusetts. While Peirce was a youth, one of the guests who used to come to Peirce's household was the American poet and professor at Harvard, Henry Wadsworth Longfellow.[12] On July 11, 1861, Longfellow's wife Fanny had been the victim of a domestic fire accident from which she actually died as a result of her injuries. The accident had been a well-known and much discussed event

8 Charles S. Peirce. *Collected Papers*, vols. 1-6, ed. Hartshorne, Charles & Paul Weiss, vols. 7-8, ed. Arthur W. Burks, Cambridge, MA: Harvard Univ. Press, 1931-58 (in the following: quoted as CP, followed by volume and paragraph number), here CP 5.487, "A survey of pragmaticism", c.1907.

9 Vincent Colapietro. *Peirce e a abordagem do self*: uma perspectiva semiótica sobre a subjetividade humana. Newton Milanze, trad. São Paulo: Intermeios, 2014, p. 151.

10 Peirce comments on this episode in "Reason's Rules" of c.1902 (CP 5.538).

11 Herbert Henry Davis Peirce (1849-1916).

12 1807-1882.

in the Peirce family.[13] Some time later, there was a dinner party at Peirce's household. And low and behold, Peirce's mother, having spilled some burning spirits on her skirt also caught on fire. Brother Herbert, picking up a rug, moved immediately to the rescue, as though he had been practicing this action, and he did it with tremendous speed and accuracy.[14] Afterwards, Charles asked Herbert about his immediate and deft response. The child in effect replied, "Well, you know, I heard the story and I had practiced in my imagination what I would do were such an event to occur in my presence."

The point is that dramatic rehearsal of an action, in imagination, merely in imagination, can be the basis for the formation of a habit. So, when the event took place at Peirce's dinner table and the dress caught on fire, Herbert's disposition was to respond immediately, as though he had physically, outwardly, practiced saving a woman from such a disaster. What he really had done was to rehearse this scenario in his imagination. "It was." as Charles later claimed, "a striking example of a real habit produced by exercises in the imagination". What enabled Herbert to respond so quickly and aptly was that "he had often run over in imagination all the details of what ought to be done in such an emergency".[15] In its most rudimentary sense, this is a description of deliberation (to turn over in imagination what one ought to do, what line of conduct would be most fitting in such and such a situation).

---

13 For the report on the accident in the New York Times of July 12, 1861, see the reprint in the NYT archives at nytimes.com/1861/07/12/archives/the-death-of-mrs-longfellow.html; accessed March 10, 2021.

14 Peirce's *Collected Papers* have two references to this episode. In "Reason's Rules", c.1902 (CP 5.538), Peirce writes: "I remember that one day at my father's table, my mother spilled some burning spirits on her skirt. Instantly, before the rest of us had had time to think what to do, my brother, Herbert, who was a small boy, had snatched up the rug and smothered the fire. We were astonished at his promptitude, which, as he grew up, proved to be characteristic. I asked him how he came to think of it so quickly. He said, 'I had considered on a previous day what I would do in case such an accident should occur.'" – Five years later, the reference is to "a lady" (in a footnote to CP 5.487, "Survey of Pragmaticism", c.1907): "I well remember when I was a boy, and my brother Herbert, now our minister at Christiania, was scarce more than a child, one day, as the whole family were at table, some spirit from a "blazer," or "chafing-dish," dropped on the muslin dress of one of the ladies and was kindled; and how instantaneously he jumped up, and did the right thing, and how skillfully each motion was adapted to the purpose. I asked him afterward about it; and he told me that since Mrs. Longfellow's death, it was that he had often run over in imagination all the details of what ought to be done in such an emergency. It was a striking example of a real habit produced by exercises in the imagination."

15 Ibid.

This brings me, I think, in some sense, to the heart of the story. For Peirce, human rationality is, first and foremost, *deliberative* rationality and, in turn, deliberative rationality involves is the dramatic imagination exercised for the sake of norms and ideals, logical, moral, and otherwise. What makes us rational or reasonable is our capacity to deliberate, to think through alternative lines of conduct. Although the expression is John Dewey's,[16] there is almost exactly the same expression in one of Peirce's unpublished manuscripts, and it is an expression used to define, or at least characterize, what "deliberation" is.[17] Deliberation is the dramatic rehearsal in imagination of various scenarios: you imagine you are going to do this, you imagine that you are going to that, and you try to, as best you can, think about the consequences.[18] If you do this, what consequences follow; if you do that, what consequences follow.

In sports today, and especially in sports psychology, there is this process called "imaging". Before a game, players will prepare for it, oftentimes by imaging it.[19] For instance, they will be imagining their opponents' particular tendencies and, in light of imagining, rehearse in their imagination what their opponent is going to do and, in turn, what they are going to do in response. So, what Peirce saw in the case of his brother is actually part of common sense, as it is also very widespread in contemporary sports.

---

16 John Dewey, *Human nature and conduct*, New York, NY: Modern Library, 1922, p. 190: "Deliberation is a dramatic rehearsal (in imagination) of various competing possible lines of action. It starts from the blocking of efficient overt action, due to that conflict of prior habit and newly released impulse to which reference has been made. Then each habit, each impulse, involved in the temporary suspense of overt action takes its turn in being tried out. Deliberation is an experiment in finding out what the various lines of possible action are really like."

17 In his *Peirce e a abordagem do self* (see fn. 4), p. 84, Colapietro quotes from Peirce's manuscript MS 649 of April 11, 1910 (p. 26): "When I speak of a man's Real Self, or True Nature, by which I mean the Very Springs of Action in him, which means how he would *act*, not when in haste, but after 'due consideration', I mean such deliberation as shall give him time to develop."

18 To illustrate how we rehearse in imagination the scenario of something that we expect to occur Peirce gives the example of what we expect to happen when we use a vending machine: "Suppose for example that I slip a cent into a slot, and expect on pulling a knob to see a little cake of chocolate appear. My expectation consists in, or at least involves, such a habit that when I think of pulling the knob, I imagine I see a chocolate coming into view. When the perceptual chocolate comes into view, my imagination of it is a feeling of such a nature that the percept can be compared with it as to size, shape, the nature of the wrapper, the color, taste, flavor, hardness and grain of what is within" ("Minute Logic", CP 2.148, c.1902).

19 Cf., e.g., Sandra Moritz et al., "What are confident athletes imaging?: an examination of image content", *The Sport Psychologist*, v. 10, n. 3, p. 171-179, 1996.

We can think about this process in a social context. For example, we are going to a meeting and there is going to be, at the gathering, a particular person who completely drives us crazy. We know that she is over-officious, rather self-important and causes us to lose our temper. So, animated by our commitment to self-control and civility, we prepare ourselves for this person in order not to lose our temper. At its heart, then, rationality is an exercise of self-control that takes the form of a dramatic rehearsal in imagination, where we try out different lines of conduct and see what ensues from those lines of conduct.[20] Peirce is quite explicit about the intrinsic link between human rationality (or intelligence) and imagination: "The whole business of ratiocination, and all that makes us intellectual beings, is performed in imagination".[21]

We are finally in a position to address squarely our titular question, "Can we change our habits?"[22] This question might be translated into: How can we change our habits deliberatively, that is, imaginatively? And this question invites another: Is habit change always goal-directed? There is always some goal, or ideal, governing the process of deliberation. For example, we are animated by the ideal of not losing our temper, not becoming angry, and not becoming one's worst self. You deliberate about what this person is going to do in this meeting, and you do so in light of the ideal of emotional self-control. You, as an athlete, are preparing against your opponent, and you know that this opponent has certain tendencies. He extremely quickly shows you the ball and takes it away. If you go for the ball, he is going to go by you. So, you imagine the move, one of his signature moves, that he puts the ball out, just close enough he baits you, he seduces you into going for it, and then he goes around you, and you deliberate your image, and say, "Do not take the fake". He puts the ball out, and you keep your position, you do not go for it.

---

20 See Vincent Colapietro, "Peirce's guess at the riddle of rationality: Deliberative rationality as the personal locus of human practice", in *Classical American pragmatism: Its contemporary vitality*, ed. Sandra B. Rosenthal, Carl R. Hausman, and Douglas R. Anderson, Urbana, IL: University of Illinois, 1999, p. 15-30.

21 "Grand Logic", 1893, CP 6.286. See also "Lessons from the History of Science", c.1896, CP 1.46-48.

22 "Among the things which the reader, as a rational person," Peirce stresses in "What Pragmatism Is" (1905), "does not doubt, is that he not merely has habits, but also can exert a measure of self-control over future actions; which means, however, not that he can impart to them any arbitrarily assignable character, but, on the contrary that a process ff self-preparation will tend to impart to action (when the occasion for it shall arise, one fixed [or recognizable] character" (EP 2, 337). "Now the theory of Pragmaticism was originally based," Peirce claims in "Issues of Pragmatism" (1905), "upon a study of thsat experience of the pheneomena of self-control which is common to all grown men and women" (EP 2, 348).

Each one of these examples is predicated on the self-controlled agent governed by a certain ideal and also relying upon the capacity of imagination to help that agent dramatically anticipate consequences – for the sake of realizing that ideal. To repeat: How can we change our habits? Not easily, certainly not directly. Agents cannot in the present immediately and instantaneously instill within themselves any given habit. On this point, Peirce is very close to Aristotle. There are at least three great philosophers – there are more than three theorists of habit since there are psychologists, anthropologists, and all kinds of folks who have written very insightfully about habits –, but there are at least three of the greatest *philosophers* who have made habit really quite central to their concern. They are Hegel, and Peirce. Aristotle, even more than Peirce, supposes that we have extremely limited control over our habits. Peirce stresses how limited our control over our habits is,[23] and how it is difficult to change habits.[24] Aristotle uses a rather remarkable metaphor here, he says the formation of a habit is originally in our power, but once the habit becomes deeply rooted in the character of the agent, that habit is almost, virtually, outside of our control. Aristotle uses the metaphor of throwing a stone. Once the stone is out of your hand, you cannot influence its course any more. The stone in flight cannot be trained to "behave" otherwise. The actions that generate human habits, by contrast, are in your control, but once the habit is formed, Aristotle says, you have little or no control over them anymore.[25]

---

23 Cf., for example, Peirce's claim that "most men are incapable of strong control over their minds. Their thoughts are such as instinct, habit, association suggest, mainly" ("Telepahy and Perception", CP 7.606, 1891).

24 In his paper "The Fixation of Belief", Peirce criticizes the incapacity of some contemporaries to change their habits (of belief) as their adherence to the method of tenacity: "A man may go through life, systematically keeping out of view all that might cause a change in his opinions. [...] But this method of fixing belief, which may be called the method of tenacity, will be unable to hold its ground in practice. The social impulse is against it. [...] Unless we make ourselves hermits, we shall necessarily influence each other's opinions; so that the problem becomes how to fix belief, not in the individual merely, but in the community" (CP 5.377-78, 1877).

25 Aristotle, *Nicomachean Ethics* 1103a (H. Rackham, trad.): "Moral or ethical virtue is the product of habit (ethos), and has indeed derived its name, with a slight variation of form, from that word. And therefore it is clear that none of the moral virtues formed is engendered in us by nature, for no natural property can be altered by habit. For instance, it is the nature of a stone to move downwards, and it cannot be trained to move upwards, even though you should try to train it to do so by throwing it up into the air ten thousand times; nor can fire be trained to move downwards, nor can anything else that naturally behaves in one way be trained into a habit of behaving in another way." Available at: perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0054%3Abekker+page%3D1103a; accessed March 10, 2021.

The argument sounds a little fatalistic, which does not do justice to other ideas of Aristotle's, with which Peirce was more in sympathy.[26] But the point is merely how severely limited habit change is as well as how our capacity to alter our habits is likely quite small. This is most evident when we are dealing with bad habits: smoking or various kinds of habits we want to abandon. It is extremely hard to do so. Peirce is deeply appreciative of this topic. Habit change is indeed possible, but it requires the sustained exercise of deliberative imagination, animated by some transformative ideal. An old dog *can* learn new tricks, but only by becoming in some respect a young pup! As far as learning goes, the necessity of becoming childlike cannot be stressed too much.[27]

W.N.: May I ask a question at this point? When you raise the question whether or how *we* can change habits, you consider agents who do or do not change their habits in a self-controlled way. Habit change of this kind belongs to the domain of thirdness, of agency guided by reason. However, in face of the present pandemia, we are all confronted with habit change imposed on us from circumstances beyond our self-control. So, should not the role of secondness, of habit change beyond our self--control be considered, too, of interruptions of habits by others, by external circumstances or even catastrophic events?

V.C.: You raise a very important point. While nominalists are in danger of eliminating thirdness altogether, Peirce and those inspired by him are sometimes at risk of exaggerating the role of thirdness. Here as everywhere else, we have to bring in secondness as well as firstness. Of course, we need to bring in all three of the categories, but for the moment, let us just limit our attention to firstness and secondness. Because I think that what happens in the cases to which you are so insightfully calling our attention is a combination between rupture and spontaneity. So, I am trying to make this computer work, and I am increasingly experiencing pure secondness. An incomprehensible opposition is thwarting my brute exertions. This thing, whatever I am doing, is not working. I am in effect bumping up against the wall. What happens, it seems to me in such ca-

26 For some other respects in which Aristotle's philosophy is relevant to Peirce, see Philip H. Hwang, "Peirce and Aristotle on chance", in E. C. Moore (ed.), *Peirce and the Philosophy of Science*, Tuscaloosa, AL: University of Alabama Press, 1993, p. 262-75. Or: Demetra Sfendoni-Mentzou, "C. S. Peirce and Aristotle on time", *Cognitio: revista de filosofia*, São Paulo, v. 9, n. 2, p. 261-280, 2008. Or: Jorge Alejandro Flórez, "Peirce's commentaries on Aristotle's accounts of induction", *Discusiones Filosóficas*, v. 31, no. 2, p. 41-57, 2017.

27 In a famous passage from Matthew, "unless you change and become as little children, you will by no means enter the kingdom of heaven" (18:3).

ses, is that we become – at least *I* become – very quickly and intensely frustrated, and I do not have the presence of mind to conceive the full array or a wider spectrum of possibilities. We (at least some of us) keep doing the same thing and getting the same results, which is almost a definition of craziness or madness. I think what happens, in the best cases, is that there is the rupture, there is the arrest of a habit. That opens space for spontaneity, and we try this, we try that; the rupture, the arrest of the habit can, optimally – does not always, does not necessarily, perhaps does not typically, but it can – open up spaces of experimentation, of possibilities we did not previously imagine or anticipate.

I think that the exercise of habit, when it leads to an impasse – when we become stuck or blocked, so that we do not know how to go on, – can be a frustrating experience, just that and nothing more. But such an experience can also be the beginning of learning. And learning involves the explosion of a greater degree of spontaneity than was previously available to us. Let me draw a parallel here. When Peirce is talking about evolution, he, of course, thinks there are three different types or forms of evolution.[28] He thinks that Charles Darwin only captures one dimension of evolution. Chance variation and radical spontaneity, which were in the focus of Charles Darwin, offer only a partial explanation of the evolutionary process. In some sense, Darwin also focused upon struggle and conflict, but even so, this is not the whole of it. Those aspects of evolution run parallel to the processes of changing habits. The agonistic dimension of conflict, also the spontaneous dimension of chance, are all part of the process.

W.N.: Evolution by chance and spontaneity implies habit change under the influence of firstness, while evolution by struggle and conflict means habit change under the influence of secondness. But how about evolution and habit change under the influence of thirdness? Isn't it important in evolution, too?[29]

---

28 For Peirce's philosophy of evolution, see his 1893 essay "Evolutionary Love" in *The Monist*, vol. 3, pp. 176-200 (also in: CP 6.287-317) and Carl R. Hausman, *Charles S. Peirce's Evolutionary Philosophy*. Cambridge: Cambridge University Press, 1993. And: Brioschi, Maria Regina, "Does continuity allow for emergence? An emergentist reading of peirce's evolutionary thought", *European Journal of Pragmatism and American Philosophy*, vol. 11, no. 2, 2019. Available at: journals.openedition.org/ejpap/1647; accessed March 10, 2021.

29 Peirce introduces his theory of the three types of evolution in his paper "Evolutionary Love" of 1893 (CP 6.302-303): "Three modes of evolution have thus been brought before us: evolution by fortuitous variation, evolution by mechanical necessity, and evolution by creative love. We may term them tychastic evolution, or tychasm, anancastic evolution, or anancasm, and agapastic evolution, or agapasm. The doctrines which represent these as severally of principal importance we may term tychasticism, anancasticism, and agapasticism. On the other hand the mere propositions that absolute chance, mechanical necessity, and the law of love are severally operative in the cosmos may receive the names of tychism, anancism, and agapism.

V.C.: I do of course think the influence of thirdness is extremely important. It ultimately culminates in *agapistic* evolution, potentially radical change through creative love.

W.N.: So, could you say a word about this third kind of evolution?

V.C.: Peirce does not write extensively on this topic, which is unfortunate because when he does write on this topic, it is deeply insightful.[30] What is crucial is that he is talking about a process of ongoing transformation, more precisely, a more or less radical self-transformation, wherein the self, in relationship to others, is being transformed in a rather dramatic, open-ended manner, and the way in which such self-transformation works is, in part, because of *agápe*.[31] Now, *agápe* means that I care for the other as though the other were myself. He is not, Peirce is clear though subtle, engaged in an act of in self-abnegation or self-annihilation. The realization of the other and the realization of myself are, in some complicated not altogether obvious way, connected. This is a process of self-alteration, of self-overcoming, in which the dialectic, or the relationship of self and other, is at the center. It is precisely my ability to immerse myself in the other which enables me to grow. So, you, Winfried, as a linguist, transform yourself by immersing yourself in the study of this language and that language: it is precisely by your deep devotion to understanding this thing, which is not yours when you first come to it; it is quite foreign. When you are studying a foreign language, it is quite foreign. But it is precisely your immersion in that, which is other than you, that causes you to become transformed, and it goes for all of the disciplines. That really is an instance of *agápe*, solicitude for the other, for the other's sake, but not in such a way as to involve a negation of one's self.

W.N.: I have two questions, and they are very different. You returned to the self several times, so my question in this context is: "Is the self a habit?" The second question is: Peirce has an almost paradoxical expression, "the habit of habit change"[32]. What does it mean? Do you have a clue on how we can overcome our perplexity at these two questions?

---

30 For example, CP 1.107 (1896) and 1.348 (1903).

31 Peirce's use of the concept of *agápe* is inspired by its biblical sense of 'unconditional love' and as an argument against the 19th century doctrine of evolution by the Spencerian principle of "survival of the fittest": "The gospel of Christ says that progress comes from every individual merging his individuality in sympathy with his neighbors. On the other side, the conviction of the nineteenth century is that progress takes place by virtue of every individual's striving for himself with all his might and trampling his neighbor under foot whenever he gets a chance to do so. This may accurately be called the Gospel of Greed" ("Evolutionary Love", CP 6.294, 1893).

32 Cf. Winfried Nöth. "Habits, habit change, and the habit of habit change according to Peirce". *In* Donna E. West; Myrdene Anderson (eds.), *Consensus on Peirce's concept of habit*, New York, NY: Springer, 2016, pp. 35-63.

V.C.: Well, I do not know that I can answer these questions, but I certainly think that I can make several steps towards addressing them. I would hesitate to say without qualification that the self is a habit. While the self is inconceivable apart from habits, it is not immediately or univocally identifiable as a habit. I would say the self is an incredibly complex network of virtually countless habits, which are more or less integrated. On my account (and I take this to be essentially a Peirce account), the self is not a habit, but *a network of habits*. It is never a fully integrated (or harmonious) system or network; it is rather a *more or less* integrated system or network of habits. Peirce uses the expression – it is commonplace in the 19th century[33] – that a human being is a "bundle of habits"[34], and that is important, but what it misses, of course, is the extent to which a person or a self is a unified or integrated bundle of habits. We are not just some random collection; we are not just some completely chaotic number of disparate tendencies and dispositions. We are, to repeat, more or less integrated. Now, what about us, grounds or ensures, at least in a minimal way, the integration of our habits? I do not want this to be construed as a dualism, but there are distinct levels of functional unity Simply as an organism, there is by virtue of physiology – thus of metabolism – a unified being. In order to have the functional unity of a living organism, the habits of my being are more or less integrated, and if they are going off, in all kinds of completely different directions, I will not survive, I could not live as such an organism. Indeed, I could not *be* an organism were my metabolic functions not knit together into a functional unit.

Thus, it seems to me, at least, that we have, at the biological or the organic level, a more or less integrated functional unity of habits. Rather early, the human organism acquires self-consciousness and various reflexive capacities. Self-consciousness is obviously a reflexive capacity, but self-consciousness provides the basis for self-criticism. I am aware of myself doing this, and I am critical of myself doing that. Self-criticism is not idle or purposeless, or it need not be idle or purposeless. It might have a point and purpose, and that point or purpose might be self-control. I take this triad to be extremely important, so the human organism, as a social actor, acquires certain dispositions, that cluster around these three capacities: the capacity for self-consciousness, the capacity for self-criticism,

33 William James, in the first sentence of chapter 4 of his *Principles of Psychology* of 1890, writes: "When we look at living creatures from an outward point of view, one of the first things that strike us is that they are bundles of habits."

34 "Each personality is based upon a 'bundle of habits', as the saying is that a man is a bundle of habits. But a bundle of habits would not have the unity of self-consciousness. That unity must be given as a centre for the habits" ("Notes for Eight Lectures", CP 6.228, 1898).

and the capacity for self-control. The capacity for self-control ultimately makes reference to some ideal. It could be a religious ideal, it could be a scientific ideal, it could be a moral ideal, could be a purely cultural ideal. This ideal allows me to integrate my habits more fully and finely than they would be in purely random biology.

So, to your first question, the self is not a habit, but a more or less integrated network of habits, as a purely biological being, that is one thing on the basis of the functional unity of the living organism. However, we become self-conscious, self-critical, and self-controlling agents and we do so in light of certain ideals, to which we devote ourselves. Without any entailment of dualism, then, the functional unity of reflexive agents depends on, but goes beyond the metabolic unity of the living organism.

W.N.: The second question was very different, the paradox of the habit of habit change. Is it missing to some people, who do not see the dangers of the future ahead?

V.C.: Yes, I do think so. I have been reading a fair amount of Albert Einstein.[35] I am teaching this semester, and Einstein has some wonderful short essays, pedagogically effective texts. One of the things he stresses in these essays is the need to maintain or rekindle our childlike wonder. Einstein was not a very good student, as a matter of fact; more bluntly, he was a somewhat bad student, in certain respects. That caused him to reflect upon education, in a really deeply thoughtful way. And one of the things that he keeps on stressing in these writings, especially in education, is that this childlike wonder is really crucial for the human animal, for the human learner.[36] One of the tragedies, and Sigmund Freud says this too, is how early and seemingly irreparably the childlike wonder of human-animal gets extinguished, or gets maimed.[37] The most delightful and successful people, it seems to me, have ways of renewing and rekindling their childlike wonder.

---

35 1879-1955.

36 Albert Einstein, *Ideas and opinions*, ed. Carl Seelig and Sonja Bargmann, New York, NY: Crown, 1954, contains several essays (p. 54-67) specifically dedicated to education. On p. 63, Einstein writes: "The point is to develop the childlike inclination for play and the childlike desire for recognition and to guide the child over to important fields for society; it is that education which in the main is founded upon the desire for successful activity and acknowledgment."

37 In *The Future of an Illusion* (New York, NY: Norton, 1961), Sigmund Freud asks his readers, "Think of the depressing contrast between the radiant intelligence of a healthy child and the feeble intellectual powers of the average adult" (p. 60). He blames religious education for "a large share of the blame for this relative atrophy" (ibid.). Whatever the cause, the contrast is stark and indeed disheartening.

It probably has to do with fear, with a deep kind of fear of change, anyway, I do not know what the cause is necessarily, but something has extinguished or deeply maimed this capacity to wonder. We have the disposition to acquire dispositions by constitution. Ideally, our habit of habit change and our capacity to change habits grow. In the course of our life, we can become more and more able – optimally, ideally – to acquire habits. The actual record, however, is rather disheartening, dispiriting. As the adage goes, "You cannot teach an old dog new tricks". The point is that the older we get, the harder it so often it is to change our habits. As folks get older, indeed, it is harder and harder for them to acquire new habits – harder, but not impossible. If you think about artists, Pablo Picasso would be one, but Miles Davis would be another example of how the habit of habit change can grow. Both were always in search of trying to play in new ways. They were always breaking down the ways in which they performed their art, trying to become like children again, learning and learning anew what it is to be a trumpet player, or what it is to be a visual artist. This capacity to reclaim our childhood, to be childlike, but not to be childish, is absolutely crucial, and it has to do with the habit of habit change.

W.N.: Thank you. We have meanwhile touched quite a number of topics, but we should no longer remain in twoness, to use Peirce's expression.[38] If you allow, I now open the floor for questions. Here we have the first, from Lucia Santaella of TIDD, "Which connection do you see between today's topic and our current pandemical crisis?"

V.C.: I see any number of connections. I think, the disposition to deny reality is a very deeply entrenched disposition in the human animal. One of the things I see in the situation of the pandemic is the disposition toward denial, and that might even be constitutional to the human animal. It seems to me that some people have the courage to confront the awful or the difficult, and other people tell themselves and tell others fairy tales. They are enamored of this.

It seems to me that there is an essential link between the search for truth and courage, and it goes to the very first thing you said. Because most of us, if not all of us, some of the time, lack the courage to change our lives.[39] Some people, all of the time, manifestly lack the courage to

38 Actually, Peirce uses this expression occasionally in the derogatory sense of 'dualism', e.g., in "Immortality in the Light of Synechism" of c.1892 (CP 7.570): "Synechism [...] can never abide dualism. [...] It does not wish to exterminate the conception of twoness [...]. But dualism [...] is most hostile to synechism."

39 In "The Fixation of Belief" (1877), Peirce gives the example of Kepler's courage in trying out one irrational hypothesis after another until he finally succeeded in changing the habits of scientific thought and the course of astronomy (CP 5.362).

change their lives. What would be the appropriate changes in the face of such a disaster, such an illness? It seems to me that we need to be both realistic and imaginative. It is not enough simply to be realistic, and it is not enough to be imaginative.[40]

I think we can tease out any number of connections, and I think it is a very important question. Without being a Pollyanna, it seems that the pandemic gives us the possibility of reimagining much of our lives. Reimagining work, reimagining school, so, yes, it is awful. Yes, it is terrible! The number of deaths is just unconscionable, and those deaths must be attributed, to a great extent, to irresponsible and inattentive political leaders. Having said that, it is an awful situation, but what might we make of it? What good might we draw out of it? Among the possibilities, it seems to me, is reimagining work. It has perhaps been time we did that.

I love the question, but it requires a much, much fuller, more detailed, more nuanced answer than I can give now, but it is worthy of long hard thought.

W.N.: There are other questions, here is one, from Alexandre Quaresma, Rio de Janeiro, "The virus is on the frontier between the living and the non-living, is it capable of semiosis?"[41]

V.C.: That is a good question. In our world, at least as far as we can ascertain, there are only fuzzy borders. There are no absolutely sharp lines of demarcation. So, the living and the dead, the self and the other, culture and nature, at certain points, the borders between these are fuzzy. They are irreducibly fuzzy, and you cannot say this is this and that is that. Anytime you have an exchange, in which something on one side is transmitted or communicated to the other side, it seems to me you have an instance of semiosis. In direct answer to the question, it is a very good question, I would say yes! The virus, even if strictly speaking, the virus is not semiosis, although I have to think about that, it lends itself to being described and explained in semiotic terms. We would be at a disadvantage if we denied ourselves the semiotic terminology to describe and explain the kinds of interactions, transactions, and transmissions that are going across these fuzzy borders.

---

40 In "The Fixation of Belief", Peirce also gives an example, from the history of chemistry, of how imagination can bring about habit change: "Lavoisier's method was not to read and pray, but to dream that some long and complicated chemical process would have a certain effect, [and...] to dream that with some modification it would have another result, and to end by publishing the last dream as a fact" (CP 5.363).

41 The question is the topic of Kalevi Kull and Winfried Nöth, "Virus semiosis", *TECCOGS: Revista digital de tecnologias cognitivas*, v. 22, p. 13-20.

So, whether or not the virus, strictly speaking, is an instance of semiosis, what is going on needs, invites, and warrants being described and explained in semiotic terms.

W.N.: So, what are the other questions? Geane Alzamora from Belo Horizonte asks: "How can we prevent fake news from become habits?"[42]

V.C.: Well, this is a very difficult question. I think that when we say "fake news" – and I assume that you are using this expression in this way –, it seems to me that we need to be clear that all news is perspectival, all news is, in certain respects, biased; it is from a certain point of view. If you read one paper, there might be a liberal bias, and when you read another paper, there might be a conservative bias. Biases in and of themselves do not make news fake. Fake news is a very different matter than simply biased news. All news is biased, some are less, others are more. News is less biased precisely because the journalists who circulate them are more conscious of their bias and try to counteract. But the issue of fake news is an extremely difficult one. It seems to me that it is a failure of education and a failure, more broadly, of a culture that allows some folks just to bombard each other with manipulated images and to shout slogans at one another. We have allowed this to gain the degree of centrality and legitimacy that it has gained in our culture. Is it possible, is it conceivable, is it imaginable to reform, in a radical way, the human discourse, the human dialogue, where we address the other in a certain manner that exhibits, displays, my respect for the humanity and the otherness of that person? It seems that we have our work cut out for us. The only way of getting rid of fake news is by generating, in an attractive way, using our media savvy, using our detailed knowledge of social psychology, the conditions for a genuine dialogue. And that is long, patient, hard work, but nothing short of that is going to make a difference.

W.N.: Here is another question, from Gilmar Hermes, de Pelotas, Rio Grande do Sul: "Is it possible to talk about habits in social life, in some phenomenological manifestation between firstness and secondness, without self-control?"

V.C.: Thank you! I remember fondly of your time here in Rhode Island. I think it is indeed ultimately, an incomplete discussion insofar as we do not bring in thirdness. But there might be very good methodo-

42 This question is also much discussed at TIDD, see: Lucia Santaellea, *A pós-verdade é verdadeira ou falsa*, São Paulo: Estação das Letras e Cores, 2018.

logical reasons not to bring in self-control, or thirdness.[43] Because what happens too often is that we rush, and we rush over the nuances, and the details, and so on. If we talk about the habits of social life, we might need to get down and dirty for a lot longer with the aspects of firstness and secondness, without thinking of them so exclusively, or even primarily, in reference to self-control and thirdness. Think of Peirce's suggestion for the defining qualities of a good phenomenologist in his second Lowell Lecture on Pragmatism of 1903. For an artist, he lectures, the first quality is the faculty to see what stares us in the face.[44] We do not see. We see what is supposed to be there, but we do not see what is phenomenologically there. If we have an overriding concern with self-control and thirdness, we are, all together, all too likely, not to attend carefully enough to the wild spontaneous ways in which social habits operate and some of the hidden imperceptible forms of conflict. So, by all means, let us give these first two categories, these first two phenomenological categories theirs full due. Now, ultimately, I think there has to be a reference to self-control, but methodologically, that might be suspended for a good long time.

W.N.: Thank you. Further questions? Here is one by Soraya Ferreira, Juiz de Fora in Minas Gerais: "New York is beautifully painted with the phrase 'Love is the answer'. What changes or is changing in terms of habits?"

V.C.: Thank you for your question. So, again, to go back to *agape*, and it might seem that I should have made more progress in my life by now, but I have not. The language of the ancient Greeks was oftentimes richer than ours. They had three words for "love", *philia*, *agape*, and *eros*. These words might not have been altogether different in the sense that

43 Self-control is one of the phenomena of thirdness in Peirce's system of categories, for Peirce a criterium of reasoning as well as moral conduct: "The phenomena of reasoning are, in their general features, parallel to those of moral conduct. For reasoning is essentially thought that is under self-control, just as moral conduct is conduct under self-control. Indeed reasoning is a species of controlled conduct and as such necessarily partakes of the essential features of controlled conduct" ("Lowell Lectures" I.1, 3rd draught, 1903, CP 1.606).

44 "What we have to do, as students of phenomenology, is simply to open our mental eyes and look well at the phenomenon and say what are the characteristics that are never wanting in it, whether that phenomenon be something that outward experience forces upon our attention, or whether it be the wildest of dreams, or whether it be the most abstract and general of the conclusions of science. The faculties which we must endeavor to gather for this work are three. The first and foremost is that rare faculty, the faculty of seeing what stares one in the face, just as it presents itself, unreplaced by any interpretation, unsophisticated by any allowance for this or for that supposed modifying circumstance. This is the faculty of the artist who sees for example the apparent colors of nature as they appear" ("Lowell Lectures" II, 2nd draught, CP 5.41-42, 1903).

they might not be altogether separable. What does "love" have to do with this? In the final analysis, and in the beginning, I am disposed to say everything. Because one of the principal reasons why we are so maimed as human beings is that we were not loved in the way we might have been loved well, according to our parent's and others' best lights, their best intentions. And we are so complicated, delicate, and multifaceted that even the most loving parent might not have loved us in precisely the way we most needed to be loved. So, if love, in the beginning, is inadequate, that has reverberation throughout a lifetime, and in the end, it seems to me that it is precisely my willingness to give up on absolutely brute force in the face of brute opposition and try to find a way in which I can recognize the humanity of the other, even when that being is acting in the most inhumane ways. In the immediate circumstance, I have to get out of it and defend myself, but I ought not to allow myself to form attitudes toward the other that are fundamentally predicated on the negation of the other. Somehow, someway, I need to reimagine the situation such that the other as human and the other as other comes rather clearly into focus.

W.N.: Here is another question. This time from Monica Allan, São Paulo: "Would the use of rationality in the process of habit change not imply hypocrisy? How can we distinguish between truth and fake with respect to the self in this context?"

V.C.: I do not know, but this is a very good, although difficult question. One of the problems is that the word "rational" is ambiguous. I am not sure that we have to back and forth about the meaning of the words and the force of the question, but it seems to me that one of the tendencies we have is to equate logicality with rationality and then, also, to think that each one of those is the equivalent of "reasonable". What I would argue is that there are actually distinct meanings here. "Rational" is not simply a synonym for the logical. Rational is always more than merely logical. As paradoxical as it might sound in English, or Portuguese, or German, "reasonable" does not carry the same nuances and the same valences as "rational". It is one thing to be rational, and it is another thing, at least slightly different, to be reasonable.

There is the notion that we can come up with a set of rules, that we can identify an algorithm, a finite set of explicit rules, or that there are rules that can in principle made explicit. This way of thinking seems to be a commitment, a defining commitment, of the rationalistic mind. It might not be reasonable at all to suppose that algorithms are at the root of everything. It may be that there are flexible, fluid, integrated, nuanced

habits that can, in some sense, be specified in the form of rules, that can, in some ways, be captured in the formula of algorithms. The lexical definitions in the dictionary or the various works used by any linguistic community only capture part of those linguistic habits. The dictionary is an attempt to distill the essence of the habit, but the distillation is never complete. There is always more to the habit that gets down on the page. It is precisely the habits that are primordial. The codes and the algorithms are secondary and derivative. It is precisely my disposition that makes me reasonable.

Peirce has a manuscript entitled "Reason's Conscience".[45] Among the things he implies there, one is that whereas a moral conscience issues mainly in imperatives, a logical conscience mainly issues in questions.[46] A moral conscience will issue negative and positive injunctions. Do not engage in acts of cruelty directed towards sentient beings. There are do's and don'ts. It seems to me that the cultivation of reason is the cultivation of self-critical habits. I do not see that as narrowly or mechanically logical. I see it as a kind of fluid, artistic, sensitivity.

If we aim at reasonable self-control over our thought, our feelings, and our actions, what is then the question that the moment most calls for? In some sense, we imagine too often both that we know what the question is and that we will know what the answer is. But part of the

45 Charles S. Peirce, "Reason's Conscience: A Practical Treatise on the Theory of Discovery; Wherein logic is conceived as Semeiotic" (1904). The manuscript is described in detail in Richard S. Robin (comp.), *Annotated catalogue of the papers of Charles S. Peirce*, Amherst, MA: University of Massachusetts Press, 1967. It has been edited in part in: Charles S. Peirce, *New Elements of Mathematics*, vol. 4, ed. Carolyn Eisele, Bloomington, IN: Indiana Press, 1976, p. 185-215, and in Charles S. Peirce, *Historical Perspectives on Peirce's Logic of Science*, vol. 2, ed. Carolyn Eisele, The Hague: Mouton, 1985, p. 801-851. – The page references in the following are to Eisele's edition of 1985.
The manuscript deals with two issues of relevance to the question of rationality, (1) the relationship between logic and other kinds of reasoning and (2) the relationship between logic and ethics. On (1), Peirce writes, "Many of our reasonings [...] are performed instinctively, and it must not, for an instant, be supposed that I should recommend that such modes of action be given up in favor of theoretical procedures, except to compare theory with practice [...]. Other reasonings, although not exactly instinctive, have become so habitual as to resemble instinctive actions. In many cases, the habits have come to us from tradition" (p. 803). – On (2), Peirce writes, "The business of ethics is to [...] find out [what] the familiar but confused idea of moral goodness really consists in. [...] Moral conduct is conduct which is self-controlled so as to be steadily directed toward a sort of purpose which ethics will define. [...] Logic, developing its own purpose in a similar way, soon finds that it is essential to the action of reasoning that it should be self-controlled: for without that, all criticism of it, as good or bad, is idle. It would, therefore, be nothing but an application of ethics to a particular kind of conduct" (p. 832).

46 E.g., What is the evidence for this claim? Is there an ambiguity hidden here?

problem, it seems, is that we are not deeply Socratic enough. We are not really in possession yet of the question. What is going on? Yeah, we have some inklings, we have some intuitions, we have some intimate intimations of what is going on, but we do not know deeply, fully, finally enough what the very questions are that we ought to be posing and addressing. It is precisely that I take reason as the capacity to ask the question that we have not yet asked and to turn the whole discussion around in new directions.

Ludwig Wittgenstein has this wonderful analogy. Someone is in a room and is trying to get out. The window is too high. The door is closed, and seems to be locked, but all the person in the room has to do is turn around and realize that, behind him, "the door has been open all the time."[47] We often have this feeling of being stuck or entrapped or even imprisoned, and we cannot find our way out. But what might be required is *metanoia*,[48] to have our minds, our souls, turned around in a new direction. Then, and only then, will we find a way out. Then, and only then, will we find the questions we need to be posing.

W.N.: I believe this is a wonderful conclusion. There may be more questions, but I cannot imagine a better ending than your last insights into habits and habit change. Therefore, I suggest that we continue thinking about our reflections instead of asking new questions. Nevertheless, I would like to give you the last word.

V.C.: In conclusion, then, I want to say two things. First, I would like to remind us of Rilke's letters to a young poet, in particular, his advice in which he says, "Do not try to answer the question, you are not yet in the position to answer the question, you must first live the question".[49] I think that is a very important piece of advice, not merely for an aspiring poet, but all of us. That we have to have the patience, the humility,

---

47 Norman Malcolm, *Ludwig Wittgenstein: A Memoir* (Oxford University Press, 1972), 52; cf. Philosophical Investigations, #108, #123, #309. Also, Ludwig Wittgensetin, *Culture and value*, ed. Georg Henrik Von Wright and Heikki Nyman, Chicago, IL: University of Chicago Press, 1980, p. 42e: "A man will be *imprisoned* in a room with a door that's unlocked and opens inwards; as long as it does not occur to him to *pull* rather than push it."

48 The Greek μετάνοια means 'after-thought' or 'beyond-thought'. It has the biblical meanings of 'transformative change of heart' and 'repentance'. For its use in philosophy, see Norman Wirzba, "From maieutics to metanoia: Levinas's understanding of the philosophical task", *Man and World*, vol. 28, no. 12, p. 9-144, 1995.

49 Rainer Maria Rilke, *Briefe an einen jungen Dichter*, Leipzig: Insel, 1929. English: *Letter to a young poet*, trans. M. D. Herter Norton, London: Penguin, 2011, Letter #4 of July 16, 1903: "Don't search for the answers, which could not be given to you now, because you would not be able to live them. And the point is, to live everything. Live the questions now." Also available at: carrothers.com/rilke4.htm; accessed March 10, 2021.

and the courage to live the questions of our time before we rather frantically and aggressively try to answer these questions. The most important thing before trying to answer any question is to live it more fully than we have thus far. Only then will we avoid superficiality or glibness. Second, I would like to express my deep gratitude. Winfried, I love conversing with you, I love the way you conduct an interview: I always feel that I am better than I usually am and I feel this is so because of the quality and depth of your questions. Of course, my gratitude extends to each of those in the audience. As in our previous exchanges, the questions from the audience have been consistently of the highest quality, and so I am very grateful to the members of the audience, all of them, for simply their attention and for those who posed those wonderful questions. A simple expression of a deep and encompassing gratitude is truly my last word: Obrigado.

W.N.: Thank you once more, dear Vincent, and thank you, Luis Felipe, for the technical organization of our meeting. Last but not least, let us also thank Lucia Santaella, the *spiritus rector* of this series of reflections.

RESENHAS

LIAO, Matthew. *Ethics of artificial intelligence*, por Dora Kaufman. *TECCOGS* – Revista Digital de Tecnologias Cognitivas, n. 23, jan./jun. 2021, p. 157-163.

Recebido em: 22 abr. 2021
Aprovado em: 05 mai. 2021

LIAO, Matthew (org.). *Ethics of artificial intelligence*. Oxford: Oxford University Press, 2020.

dx.doi.org/10.23925/1984-3585.2021i23p157-163

Resenha do livro *Ethics of artificial intelligence*, de Matthew Liao

Dora Kaufman[1]

A empresa inglesa *DeepMind Technologies*, fundada em 2010 e adquirida pelo Google em 2014, é uma referência no campo da Inteligência Artificial (IA); em 2016, seu programa *AlphaGo* não apenas venceu por quatro a um o campeão mundial do milenar jogo chinês Go, o sul-coreano Lee Sedol, como o fez com jogadas inéditas. O feito repercutiu na comunidade de IA, impulsionando o reconhecimento do papel estratégico da tecnologia pela China. No mesmo ano, não por coincidência, o *Center for Mind, Brain, and Consciousness* da NYU, sob a coordenação dos filósofos David Chalmers e Ned Block, reuniu cerca de trinta palestrantes, dentre pesquisadores de tecnologia e de ciências humanas, na conferência "AI Ethics".

No empenho de identificar como introduzir nos sistemas inteligentes princípios éticos e valores humanos, os painéis abordaram conceitos tais como moralidade e ética das máquinas, moralidade artificial e IA amigável. Ao longo da conferência, contudo, estabeleceu-se um consenso de que a ética pertence à esfera humana, ou seja, permeia as escolhas de desenvolvedores e usuários. Coube ao filósofo sueco Nick Bostrom, autor do livro "Superintelligence" (2014), abrir o evento alertando sobre os benefícios e riscos da concretização da "máquina inteligente" no século XXI. Além de Bostrom, a conferência contou com palestras de Peter Asaro, John Basl, Meia Chita-Tegmark, Kate Devlin, Vasant Dhar, Virginia Dignum, Mara Garza, Daniel Kahneman, Adam Kolber, Yann LeCun, Gary Marcus, Steve Petersen, Francesca Rossi, Stuart Russell, Ronald Sandler, Jürgen Schmidhuber, Susan Schneider, Eric Schwitzgebel, Frans Svensson, Jaan Tallinn, Max Tegmark, Wendell Wallach, Stephen Wolfram e Eliezer Yudkowsky.

1 Doutora na Escola de Comunicações e Artes pela USP. ORCID: orcid.org/0000-0001-7060-4887. CV Lattes: lattes.cnpq.br/8045171889826285. E-mail: dkaufman@usp.br.

Essa é a origem da coletânea de artigos organizada por S. Matthew Liao (doze dos trinta colaboradores foram palestrantes na conferência). Composta de 17 ensaios inéditos, produzidos por cientistas e filósofos, agrupados em quatro seções, a coletânea aborda dilemas-chave para uma IA ética, dentre outros, os impactos da automação inteligente no mercado de trabalho; o viés contido nos dados que perpetuam os preconceitos da sociedade; a ética envolvida em aplicações como carros autônomos, sistemas de vigilância, armas autônomas; robôs sexuais; direitos e consciência da IA; e status moral. Numa perspectiva futura, os ensaios da terceira seção refletem sobre os riscos da "superinteligência".

No ensaio inicial, "A Short introduction to the ethics of Artificial Intelligence", S. Matthew Liao subdivide as abordagens éticas em dois conjuntos: (a) as associadas à eficiência da técnica em alguns domínios, implicando que os humanos podem se sentir vulneráveis ao lidar com esses sistemas, denominadas por ele de "vulnerabilidades humanas", e (b) as associadas às limitações da técnica, denominadas por ele de "vulnerabilidades no aprendizado de máquina". O primeiro conjunto de questões éticas abarca, dentre outras externalidades negativas, a ameaça ao suposto "livre arbítrio" dos indivíduos, função da capacidade dos algoritmos de IA extrair dos dados conhecimento inédito sobre os usuários das plataformas/dispositivos tecnológicos e, com base nele, elaborar estratégicas para influenciar/alterar/manipular o comportamento humano; a privacidade por conta da disseminação dos sistemas de monitoramento e vigilância com o uso de técnicas de reconhecimento facial; o aperfeiçoamento das fake news com as deepfakes e sua capacidade de distorcer imagem e voz, simulando falas, imagens e vídeos de pessoas reais com forte aproximação da realidade; o deslocamento do trabalhador humano por sistemas inteligentes mais rápidos e mais eficientes e a um custo menor. No segundo conjunto de questões éticas, destacam-se o problema do viés nos modelos de IA e o problema da não explicabilidade de como os modelos chegaram ao resultado final.

A indagação central de Liao é como criar sistemas de IA que sejam justos e não gerem resultados tendenciosos inadvertidamente; outro aspecto abordado no ensaio é se é factível atribuir status moral aos sistemas de IA. Sobre os resultados tendenciosos, é importante ter em mente que a maior parte das aplicações atuais de IA é baseada na técnica de *machine learning* denominada Redes Neurais de Aprendizado Profundo (*Deep Learning Neural Networks* – DLNNs), em que os algoritmos "aprendem" estabelecendo correlações a partir de grandes conjuntos de dados

(*big data*). Nessa técnica, o viés deriva (a) da codificação de estruturas e padrões mentais existentes, filtrados pelos desenvolvedores dos sistemas ao definir variáveis iniciais, arquiteturas, base de dados; (b) de dados tendenciosos, no caso da base de dados de treinamento dos algoritmos não representar o universo do objeto em questão; (c) da realidade ser enviesada, quando os dados refletem os preconceitos existentes na sociedade; e (d) de previsões baseadas em dados do passado, efeito minimizado em séries não temporais (*computer vision*/*image recognition* e NLP – *Natural Language Processing*). O tema do status moral dos sistemas de IA é abordado por outros autores, e retomado por Liao no último ensaio da coletânea.

Parte I: Construindo Ética em Máquinas, com cinco ensaios

No primeiro ensaio, "Ethical learning, natural and artificial", Peter Railton alerta para os impactos éticos diretos e indiretos, na medida em que a IA altera as capacidades e os potenciais benefícios e riscos de outras tecnologias. Reconhecendo o crescente protagonismo dos sistemas artificiais na tomada de decisão que afetam a vida, o autor investiga a possibilidade desses sistemas se tornarem sensíveis às questões éticas, definido essa sensibilidade como a "capacidade robusta e confiável de detectar e responder apropriadamente a características eticamente relevantes de situações, ações, agentes e resultados" (p. 45). No segundo ensaio, "The use and abuse of the Trolley Problem: self-driving cars, medical treatments, and the distribution of harm", F. M. Kamm apresenta os casos comumente utilizados para ilustrar os dilemas éticos e os julgamentos morais padrão associados a condutas permissíveis. Kamm diferencia as questões morais dos carros autônomos do padrão de *trolley problem*, atribuindo responsabilidade aos programadores por danos a pedestres, motoristas e passageiros. No terceiro ensaio, "The moral psychology of AI and the ethical opt-out problem", Jean-François Bonnefon, Azim Shariff, e Iyad Rahwan argumentam que a promessa da IA de melhorar as decisões humanas só pode se tornar realidade se incorporar os *trade-offs* morais exclusivos dos humanos, tarefa da competência dos cientistas comportamentais que terão que adaptar os métodos de psicologia moral a domínios técnicos complexos.

No quarto ensaio, "Modeling and reasoning with preferences and ethical priorities in AI Systems", Andrea Loreggia, Nicholas Mattei, Francesca Rossi e K. Brent Venable defendem a premência de construir sistemas inteligentes que se comportem moralmente (alinhados com valores humanos), pré-condição para torná-los confiáveis, particularmente no caso dos "robôs cuidadores". Os autores propõem uma modelagem

para detectar prioridades éticas, e os possíveis desvios com referência nos valores da comunidade de usuários desses sistemas. O quinto ensaio, "Computational law, symbolic discourse, and the AI constitution", Stephen Wolfram defende a viabilidade de criar uma linguagem de discurso simbólica geral e aplicá-la para construir uma estrutura para o direito computacional, incluindo no futuro as "IAs": que ética e quais princípios, e como inseri-los nos sistemas.

Os cinco ensaios atribuem um agenciamento inexistente nos sistemas atuais de IA, que são "meros" modelos estatísticos de probabilidades aos quais não pode ser atribuída a condição de agente moral, pressuposto corroborado por vários filósofos. Mark Coeckelbergh (2019, 2020) argumenta que as tecnologias de IA podem ser agentes, no sentido de atuar no mundo, mas não atendem aos critérios tradicionais de agente moral mesmo reconhecendo que as decisões automatizadas com IA podem não ser moralmente neutras. Wendell Wallach e Colin Allen (2009) cunharam o termo "moralidade funcional", para designar uma moralidade intermediária, nem plena nem neutra. Bostrom e Yudkowsky (2014) recusam conceder aos atuais sistemas de IA, ainda de competência restrita a um único domínio, o status moral, mesmo que seus algoritmos exerçam funções cognitivas anteriormente atribuídas aos seres humanos.

Parte II: O Futuro Próximo da Inteligência Artificial com quatro ensaios

No primeiro ensaio, "Planning for mass unemployment: precautionary basic income", Aaron James trata do efeito da automação sobre o emprego e potenciais iniciativas para evitar o desemprego em massa. No segundo ensaio, "Autonomous weapons and the ethics of Artificial Intelligence", Peter Asaro alerta sobre o potencial das "armas autônomas" transformarem radicalmente a guerra, o policiamento e o entendimento de direitos humanos relacionados a máquinas e algoritmos de IA. Diante da "imoralidade" dessas armas, o autor argumenta a favor da supremacia dos direitos e deveres morais sobre as razões utilitárias.

No terceiro ensaio, "Near-Term Artificial Intelligence and the ethical matrix", Cathy O'Neil e Hanna Gunn argumentam que os problemas no curso prazo dos sistemas de IA são problemas morais, presentes desde as decisões de design dos algoritmos. Por meio de estudos de caso, as autoras buscam mostrar que os interesses humanos não estão sendo contemplados no desenvolvimento e uso desses sistemas, sugerindo a criação de uma "matriz ética" aos moldes da ética de ciências de dados.

No quarto ensaio, "The ethics of the artificial lover", Kate Devlin aborda as tecnologias de sexo, ainda de consumo de nicho, mas com potencial de expansão ao promover experiências robóticas multissensoriais. Para a autora, essas tecnologias podem proporcionar vidas sexuais mais gratificantes ao romper barreiras fisiológicas, psicológicas e discriminatórias, sem negligenciar a ética (segurança de dados, privacidade e controle e consentimento do usuário).

São inúmeras as externalidades negativas, éticas e sociais, dos sistemas atuais de IA, algumas gerais e outras relacionadas ao setor e/ou tarefa de aplicabilidade (a natureza e grau de impacto ético associados a sistemas de recomendação de filmes/música são radicalmente distintos, por exemplo, de sistemas bélicos). A técnica de DLNNs, como todo modelo estatístico de probabilidade, é intrinsecamente incerta, ademais, sendo baseada em grandes conjuntos de dados, agrega os vieses contidos nos dados. Essas e outras características dos modelos de IA implicam em questões éticas a serem equacionadas, ao menos minimizadas, pela sociedade. Do âmbito social, a automação inteligente configura-se como a maior ameaça ao eliminar funções repetitivas e cognitivas em distintos setores econômicos.

Parte III: Impactos de longo prazo da superinteligência com quatro ensaios

No primeiro ensaio, "Public policy and superintelligent AI: a vector field approach", Nick Bostrom, Allan Dafoe e Carrick Flynn chamam a atenção para uma série de circunstâncias especiais que podem cercar o desenvolvimento e a implantação da "IA superinteligente"; com base em uma abordagem de "campo vetorial" para a análise normativa, os autores buscam extrair implicações de política direcional dessas circunstâncias especiais, implicações caracterizadas como um conjunto de "desiderata" (p. 313-314). "Propostas de política" referem-se a documentos oficiais do governo e planos desenvolvidos por atores privados interessados no desenvolvimentos de longo prazo da IA. O segundo ensaio, "Artificial Intelligence: a binary approach", Stuart Russell aborda o problema de controle da IA, o que envolve a construção de sistemas mais poderosos que os humanos com a garantia de que os mesmos serão benéficos. Russell distingue sistemas "melhores para tomar decisão" de sistemas que tomam as melhores decisões.

No terceiro ensaio, "Alignment for advanced Machine Learning systems", Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire e Andrew Critch identificam oito áreas de pesquisa direcionadas para projetar sistemas de IA robustos e confiáveis, ressaltando que as soluções devem contemplar os sistemas atuais e os sistemas altamente inteligentes do futuro, bem como devem funcionar na teoria e na prática. No quarto ensaio, "Moral machines: from value alignment to embodied virtue", Wendell Wallach e Shannon Vallor partem das "Três leis para robôs" de Isaac Asimov para discutir leis mais adequadas à complexidade da Inteligência Artifical Geral (*General Artificial Intelligence* - GAI), mesmo reconhecendo que a IA ainda continuará a ser projetada para contextos morais limitados nas próximas décadas, requerendo engenharia, teste, vigilância, supervisão e refinamentos iterativos. No quinto ensaio, "Machines Learning Values", Steve Petersen, considerando factível a possibilidade de os humanos criarem uma superinteligência artificial com valores próprios (capaz, inclusive, de exterminá-los), defende projetar essa superinteligência com valores fundamentais semelhantes aos humanos, conhecido como "alinhamento de valores". O autor reconhece a complexidade desses valores para serem programados explicitamente, mas não para serem "aprendidos" pelas técnicas de *machine learning,* identificando três obstáculos e suas potenciais soluções inter-relacionadas.

As questões dessa seção remetem a previsões sobre o futuro da IA que, dada as limitações atuais das técnicas de *machine learning*, carecem de evidências científicas de que serão (ou não) concretizadas.

Parte IV: Inteligência Artificial, Consciência e Status Moral com três ensaios

No primeiro ensaio, "How to catch an AI zombie: testing for consciousness in machines", Susan Schneider alerta para a premência de antecipar os futuros problemas quando a IA superará os humanos em múltiplos domínios. A autora aborda de diferentes ângulos o conceito de "consciência", inclusive o que seria uma "IA consciente" aferidos por meio de vários testes ou marcadores. No segundo ensaio, "Designing AI with rights, consciousness, self-respect, and freedom", Eric Schwitzgebel e Mara Garza ponderam que no futuro será possível criar entidades com AI que mereçam tanta consideração moral quanto os seres humanos. Os autores convocam filósofos e formuladores de políticas para antecipar a discussão dos princípios éticos associados, com o pressuposto de que a IA merecedora de consideração moral equivalente à humana deve ser projetada com valores próprios (não necessariamente humanos).

No terceiro ensaio, "The moral status and rights of Artificial Intelligence", S. Matthew Liao retoma a questão do status moral dos sistemas de IA defendendo sua aplicabilidade (a) no caso de IA "vivas, conscientes ou sencientes" (capaz de sentir dor, ter desejos); (b) no caso de ter "base física"; (c) no caso de base física por emulação do cérebro; (d) no caso de seres humanos interessados em se tornar IAs por substituição gradual (em vez de emulação do cérebro); e (e) no caso de IA com direitos autorais. Segundo o autor, alguns dos direitos serão semelhantes aos direitos dos seres humanos, como o direito à vida e à liberdade e o direito a igual proteção, além de direitos exclusivos de sua natureza, como o direito de controlar sua taxa subjetiva de tempo. Liao inclui as formas de vida artificiais em uma lista de nove entidades com potencial de ter status moral: objetos inanimados (rochas, obras de arte, edifícios, o meio ambiente); coisas vivas terrestres não humanas (plantas e animais); seres humanos com funcionamento normal; seres humanos feridos (gravemente deficientes mentais); seres humanos no início da vida (fetos, bebês, crianças pequenas); possíveis seres humanos futuros (gerações futuras); seres humanos não vivos (seres humanos mortos); espécies extraterrestres não humanas de seres vivos (alienígenas, seres do espaço sideral); e formas de vida artificiais (androides, robôs, computadores, algoritmos).

São múltiplas as externalidades negativas dos sistemas atuais de IA, éticas e sociais. O desafio é como mitigá-las preservando as externalidades positivas intrínsecas aos modelos de negócio baseados em dados (*data-driven models*). Esse cenário recomenda manter o foco na busca por soluções de curto-médio prazo, deixando o longo prazo para a esfera da ficção científica.

Referências

BOSTROM, Nick; YUDKOWSKY, Eliezer. The Ethics of Artificial Intelligence. In: FRANKISH, Keith; RAMSEY, William (eds.). *The Cambridge Handbook of Artificial Intelligence*. New York, NY: Cambridge University Press, 2014. Disponível em: cambridge.org/core/books/cambridge-handbook-of-artificial-intelligence/ethics-of-artificial-intelligence B46D2A9DF7CF3A9D92601D9A8ADA58A8. Acesso em: 12 maio 2021.

COECKELBERGH, Mark. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 2019. Disponível em: link.springer.com/article/10.1007/s11948-019-00146-8. Acesso em: 5 abril 2021.

_____. AI Ethics. Cambridge, MA: MIT Press, 2020.

WALLACH, Mendell; ALLEN, Colin. Moral Machines: Teaching Robots Right from Wrong. New York, NY: Oxford University Press, 2009.






Diretrizes para autores – **TECCOGS**

A *TECCOGS – revista digital de tecnologias cognitivas* é um periódico do Programa de Pós-Graduação em Tecnologias da Inteligência e Design Digital (TIDD) da Pontifícia Universidade Católica de São Paulo (PUC-SP). As edições são semestrais e exclusivamente digitais, disponíveis em pucsp.br/pos/tidd/teccogs.

A TECCOGS recebe artigos e resenhas de doutores ou de especialistas, mestrandos, mestres e doutorandos em coautoria com doutores.

Título, **subtítulo**, **resumo** (com no mínimo 1000 e no máximo 2500 caracteres com espaços) e **palavras-chave** (de três a seis termos) do artigo deve aparecer em português ou espanhol (caso o artigo esteja escrito nessa língua) e, logo em seguida, traduzidos para o inglês.

O(s) **nome(s) do(s) autor(es)** deve(m) estar logo abaixo do subtítulo do artigo, acompanhado de uma nota de rodapé (escrita em fonte *Times New Roman* tamanho 11 pt, espaçamento simples) contendo currículo e biografia (formação, vínculo acadêmico, área de atuação e e-mail) com, no máximo, cinco linhas.

Cada artigo deve possuir no mínimo 20.000 e no máximo 50.000 caracteres com espaços.

Resenhas devem possuir no mínimo 8.000 e no máximo 13.000 caracteres com espaços.

O **corpo do texto** deve ser configurado em fonte *Times New Roman* tamanho 12 pt, espaçamento 1,5 linhas, parágrafo alinhado à esquerda, sem hifenização. **Citações diretas com quatro linhas ou menos** devem aparecer entre aspas (“”) incorporadas ao corpo do texto, indicando a fonte entre parênteses no modelo “(SOBRENOME [em maiúsculas], ano de publicação, p. [número da página])”, conforme a Norma Brasileira (NBR) 10520 (ago. 2002) da ABNT.

As **citações diretas com mais de quatro linhas** devem ter recuo à esquerda de 4 cm, sem aspas, com fonte *Times New Roman* tamanho 11 pt, espaçamento simples, parágrafo justificado e sem hifenização.

Imagens (fotografias, ilustrações, diagramas, tabelas, gráficos) precisam ter resolução de, no mínimo, 100 dpi/ppi (*pixels* por polegada) e devem estar integrados ao corpo do texto, com imagem e legenda centralizadas e fonte especificada (para imagens da *internet*: “Disponível em: “<site>”. Acesso em: “dia mês abreviado ano”).

O texto deve respeitar o **Novo Acordo Ortográfico da língua portuguesa**, vigente desde 2009. De acordo com a Base XIX da Nova Ortografia, termos como “Inteligência Artificial”, “Psicologia Cognitiva”, “Informática” e “Filosofia” (quando se trata da área de conhecimento) devem iniciar com maiúsculas. Segundo a política de direitos autorais da revista, os autores se responsabilizam pelos direitos de uso de todas as imagens.

Para elaboração de resumos, citações e referências, a revista segue as NBR 6023 (ago. 2002), 6028 (nov. 2003) e 10520 (ago. 2002) da ABNT. Não são permitidas notas de fim. Notas de rodapé devem ser usadas o mínimo possível, exclusivamente para adicionar observações pontuais, nunca para indicar referências bibliográficas. Em fontes da *internet*, a autoria do texto deve ser indicada entre parênteses, bem como o ano de publicação e endereço e data de acesso.

Todas as obras mencionadas nas referências devem estar citadas ao menos uma vez no texto e, do mesmo modo, toda e qualquer obra mencionada no texto deve constar nas referências.

A TECCOGS disponibiliza um arquivo formato .DOC que serve de *template* com instruções e exemplificações e estilos detalhados para escrever o artigo. Baixe o modelo aqui.