REVISTA RUMOS DA PESQUISA EM CIÊNCIAS EMPRESARIAIS, CIÊNCIAS DO ESTADO E DA TECNOLOGIA

Revista Rumos da Pesquisa em Ciências Empresariais, Ciências do Estado e da Tecnologia
ISSN: 2525-278X

http://revistas.unicerp.edu.br/

O IMPACTO DISRUPTIVO QUE O DEEPFAKE PODE CAUSAR NA SOCIEDADE

Everton Ferreira Silva
everton102010@live.com
https://orcid.org/0000-0003-0190-8662
UNICERP, Patrocínio, MG, Brasil

Lucca de Barros Casalenovo
debarroscasalenovo@gmail.com
https://orcid.org/0000-0002-0121-8330
UNICERP, Patrocínio, MG, Brasil

Cássio Aparecido do Amaral
cassio@unicerp.edu.br
https://orcid.org/0000-0001-7371-7414
UNICERP, Patrocínio, MG, Brasil

Recebido em: 01/11/2022
Aprovado em: 14/07/2023

DOI: http://dx.doi.org/10.17648/2525-278X-v1n7-6

Correspondência:
Everton Ferreira Silva
Rua Jacob Marra, n°544, Centro, Patrocínio, Minas Gerais, Brasil

RESUMO

INTRODUÇÃO: As mídias críveis, que são geradas por uma rede neural, possuem características cognitivas análogas aos indivíduos e são capazes de aparentar que alguém disse ou fez algo mesmo sem o consentimento da pessoa alvo. Esta tecnologia pode ser usada para praticar crimes e ser colocada a serviço da desinformação, transformando indivíduos comuns em vetores ativos da sua propagação.

OBJETIVO: O objetivo geral é determinar o perigo que o Deepfake acarreta e seus potenciais danos à sociedade. Os objetivos específicos são descrever a violação de direitos fundamentais; determinar se a legislação vigente consegue regular a disseminação dessas mídias online e caracterizar os danos oriundos da desinformação acarretado pela tecnologia.

MATERIAL E MÉTODOS: O tipo da pesquisa é descritiva-qualitativa e seu método é o hipotético-dedutivo que confronta duas hipóteses. A pesquisa teve como técnica o levantamento bibliográfico.

RESULTADOS: A confecção dos Deepfakes por intermédio de duas abordagens (Autoencoders e GANs) torna possível gerar mídias que se valem de plataformas digitais e fenômenos; como a atração por notícias falsas, cognição cultural, informação em cascada, bolhas online de informação e a verdade ilusória; para violar direitos fundamentais e espalhar desinformação. Projetos de regularização desta tecnologia emergem no mundo e somado ao avanço tecnológico de sistemas de identificação de detecção automatizada da integridade de uma mídia digital proporciona o combate aos Deepfakes.

CONCLUSÃO: O Estado poderá não ser baseado em evidências empíricas e será possível criar uma legislação para proteger as pessoas contra os usos irrestritos da tecnologia.

PALAVRAS-CHAVE: Deepfake. Ameaça. Tecnologia.

REVISTA RUMOS
DA PESQUISA EM CIÊNCIAS EMPRESARIAIS, CIÊNCIAS DO ESTADO E DA TECNOLOGIA

THE DISRUPTIVE IMPACT DEEPFAKE CAN HAVE ON SOCIETY

ABSTRACT

INTRODUCTION: The believable media generated by a deep neural network that has cognitive characteristics similar to ordinary people let us to believe that one person said or did something without the consent of them. This technology can be used to commit crimes and be available to the service of disinformation, turning ordinary individuals into active vectors of its propagation.

OBJECTIVE: The overall objective is to determine the danger that Deepfake can have on society. The specific objectives are to describe the violation of fundamental rights; determine if the current legislation can regulate the dissemination of these online media and characterize the damage arising from the disinformation caused by technology.

METHODS: The research is descriptive-qualitative and its method is hypothetical-deductive, which confronts two hypotheses. The research had as technique the bibliographic survey.

RESULTS: The making of Deepfakes depends on two approaches (Autoencoders and GANs) that makes it possible to generate media that use digital platforms and phenomena, such as the attraction of fake news, cultural cognition, cascading information, filter bubbles and the illusory truth; to violate fundamental rights and spread disinformation. Projects to regulate this technology that are emerging in the world and added to the technological advance of identification systems of the automated detection of the integrity of a digital media makes it possible to combat Deepfake.

CONCLUSION: The State may not be based on empirical evidence and it will be possible to create legislation to protect people against the unrestricted uses of technology.

KEYWORDS: Deepfake. Threat. Technology.

INTRODUÇÃO

O avanço exponencial tecnológico no mundo foi capaz de possibilitar a criação dos Deepfakes que já estão dispostos na vida comum das pessoas. Utilizando-se de mídias digitais hiper-realistas, que são geradas por uma rede neural profunda e possuem características cognitivas análogas aos indivíduos capazes de criar conteúdo parecendo que alguém disse ou fez algo tendo ou não o consentimento da pessoa alvo. Está tecnologia pode ser usada para a prática de diversos tipos de crimes e pode ser colocada a serviço da desinformação, transformando indivíduos comuns em vetores ativos de propagação de noticia falsas.

Uma mãe na Pennsylvania fez o uso de um Deepfake para manipular e forjar fotos/vídeos que demonstrasse a companheira de equipe de sua filha fumando, bebendo e em cenas de nudez, uma vez que eram animadoras de torcidas rivais (GUARDIAN, 2021). A tecnologia foi capaz de atingir uma jornalista após fazer uma campanha para uma vítima de estupro chamada Kathua. Ela teve o seu rosto inserido em vídeos pornográficos na internet (DELHI, 2018). Os Deepfakes, ainda, foram utilizados para aplicar um golpe em um CEO em $243, 000 (DAMIANI, 2019) e em um caso envolvendo o Azmin Ali, ministro de assuntos econômicos da Malázia, que se envolveu em um escândalo com o vazamento de cenas em que participava de relações sexuais com um membro do seu partido que configurava um crime (AJDER, CAVALLI, GRAPHIC, PATRINI, 2019) em seu país.

Nesta pesquisa, investiga-se as repercussões negativas do uso negativo dos Deepfakes e possui como objetivo geral identificar o perigo que o Deepfake acarreta e seus potenciais perigos para a sociedade. Os objetivos específicos são descrever a violação de direitos fundamentais; determinar se a legislação vigente consegue regular a disseminação dessas mídias online e caracterizar os danos oriundos da desinformação acarretado pela tecnologia.

MATÉRIAS E METÓDOS

O tipo da pesquisa é descritiva-qualitativa e seu método é o hipotético-dedutivo que confronta as seguintes hipóteses: “Se não existem maneiras efetivas de combater os impactos nocivos dos Deepfakes e os indivíduos não identificam eficientemente mídias falsas, fere-se o

Estado Democrático de direito que não irá mais ser pautado em evidências empíricas" e "Se o estado não é baseado em evidências empíricas, será necessário a criação de uma legislação especifica para delimitar o uso do Deepfakes".

O trabalho cientifico terá como técnica o levantamento bibliográfico, utilizando-se da releitura de livros, artigos acadêmicos e sites pertinentes para a temática abordada.

RESULTADOS E DISCUSSÃO

A gênesis e o desenvolvimento dos deepfakes

Anos após a criação da IA e como seu produto, surgem os *Deepfakes* que é formado pela junção de outras duas palavras: '*deep learning*' e '*fake*', podendo ser caracterizada como sendo uma "mídia crível gerada por uma rede neural profunda (LEE, MIRSKY, 2020) - que são regularmente confeccionados por meio de duas abordagens: *Autoencoders* e *Generative Adversarial Networks* (GANs).

Deste modo, a primeira abordagem citada (*Autoencoders*) pode ser dividida em três etapas. Primeiro, deve-se analisar em cada quadro de um vídeo a fisionomia e expressões do rosto do material de origem e destino, ordenando os dois. Segundo, entende-se as particularidades do rosto de um indivíduo específico e como um determinado segmento do rosto se relaciona e compõe as expressões faciais (a sua postura, a luz, a cor da pele e etc). Por último, após encerrar o treinamento do programa é possível simular qualquer expressão do alvo com a análise do material de origem, isto é, aprende a modelar o rosto de uma pessoa indo além do material existente (BOUCHER, 2021; KIETZMANN, KIETZMANN, LEE, MCCARTHY).

A Rede Adversária Generativa - GANs, que é um tipo específico de aprendizado profundo (*Deep learning*), foi criado após sucessivas evoluções tecnológicas, no ano de 2014, por Ian Goodfellow e outros pesquisadores da universidade de Montreal (SPIVAK, 2019). Esta abordagem baseia-se em duas redes neurais treinadas dentro do mesmo banco de dados e que trabalham juntas para criar vídeos, textos, sons ou imagens com uma enorme similaridade com a realidade: são as redes geradora e discriminadora.

A primeira rede sintetiza um conteúdo verossímil a sua amostra de dados (por exemplo, imagens artificiais da pessoa X conforme os dados reais de X), a segunda, por outro lado, busca

identificar os erros da criação artificial, avalia-la e informar a rede geradora sobre os resultados de sua análise. Em conjunto as redes se aperfeiçoam e aprimoraram, dando a possibilidade de se produzir Deepfakes cada vez mais sofisticados. Em que pese a geração de Deepfakes seja necessárias vastas informações, pesquisadores já conseguem gerar falsificações por meio de apenas uma foto (CHELSEY, CILTRON, 2019; WESTERLUND, 2019).

> Enquadram-se na sintetização de vídeos em algumas categorias principais: *Facial expression manipulation*, *Face morphing*, *Face replacement/swap*, *Face generation* e *Full body puppetry* (...) A tecnologia consegue expressar o estilo da pessoa na sua maneira de escrever e falar algo (...) A arquitetura comumente utilizada é a NPL (*Natural Language Processing*) (...) que foi utilizado para criar o GPT-3 (*Generative Pre-trained Transformer 3*) que pode criar artigos de notícias sintéticos que avaliadores tem dificuldade de diferencia-los dos produzidos por humanos (BOUCHER, 2021, p. 8-13, tradução nossa).

> Com a síntese de voz é possível renderizar passagens de um texto em um áudio. Um exemplo desta síntese é o sistema WaveNet do Google que foi treinado por falantes profissionais por meio de um banco de dados que continham 24,6 horas de falas em inglês (...) para a avaliação de desempenho dessas gravações foram usados testes subjetivos de comparação emparelhada e teste de pontuação média de opinião (MOS). Os indivíduos deveriam classificar a voz sintética em uma escala de naturalidade (1: Ruim, 2: Ruim, 3: Regular, 4: Bom, 5: Excelente). Os resultados foram uma avaliação de 4,21 em média das vozes sintetizadas e a fala do ser humano foi avaliada como 4,55, com uma diferença de cerca de 8% (DIAKOPOULOS, JOHSON, 2021, p. 3 tradução nossa).

Cerca de dois anos após a criação das GANs, a academia tomou conhecimento sobre os Deepfakes na *Conference on Computer Vision and Pattern Recognition*, logo que Justus Thies apresentou sua pesquisa sobre recomposição e captura de rosto em tempo real (RUITER, 2021). Entretanto, a tecnologia ficou amplamente conhecida pelo público em 2017, quando em novembro um usuário do Reddit @deepfakes publicou diversos vídeos com os rostos de atrizes famosas (Gal Gadot, Emma Watson, Scarlett Johansson) nos corpos de mulheres em cenas pornográficas. Este usuário também forneceu o software livre ("*FakeApp*"), oportunizando que pessoas comuns com os seus computadores domésticos e sem possuírem formação em ciência da computação pudessem produzir Deepfakes, o que levou a uma enorme acessão na produção de conteúdos falsos. Assim, a mídia começou a também utilizar o termo 'DeepFake' para se referir a vídeos deste gênero (DONOVAN, PARIS, 2019; RINI, 2020; SPIVAK, 2020). Esta foi a primeira demonstração de quais poderiam ser os alarmantes impactos causados na sociedade.

Novas formas de praticar antigos crimes

As mídias sintetizadas podem ser utilizadas para fins positivos e negativos. Dentre suas aplicações benéficas, o Deepfake pode criar obras de artes; experiências únicas; inovações de aplicações para tratamentos médicos, psicológicos e possibilitando novas pesquisas na área médica; fazer documentários, campanhas e inovações educacionais, podendo trazer indivíduos que ficaram marcados na história para dialogar com uma sala de aula, criar vídeos de coisas que ainda não aconteceram e recriar momentos históricos que ocorreram, mas não foram gravados (CHELSEY, CILTRON, 2019; FALLIS, 2021).

> Por meio da tecnologia é possível ajudar o tratamento de Alzheimer e o contato com alguém que este indivíduo se lembre, a pessoas transgêneros a se aceitarem e se verem no seu gênero, também conseguiria trazer digitalmente um ente próximo já falecido de volta, oportunizando conversas virtuais e uma despedida de amigos e familiares a pessoa morta (...) O documentário chamado "*Welcome to Chenya*" em que indivíduos LGBT que foram perseguidos na Rússia puderam contar suas histórias sem serem identificados graças o deepfake (WESTERLUND, 2019, p. 41, tradução nossa).

> Ainda o Javier Valdez, escritor e jornalista, "voltou a vida" para cobrar o presidente do México, Lopes Obrador, para lutar com mais veemência contra o crime organizado e a corrupção (...) Como Javier foi morto durante as suas investigações sobre o crime organizado, o programa "*Defending Voices Program for the Safety of Journalist*" o trouxe de volta em um vídeo em que clama por justiça aos jornalistas desaparecidos (BRENNER, FILKUVOVÁ, LANGGUH, POGORELOV, SCHROEDER, 2021, p. 8, tradução nossa).

Os impactos danosos que a tecnologia pode causar são variados, não criando novos problemas, mas deteriorando ainda mais as vulnerabilidades pré-existentes nas instituições, sistemas e estruturas democráticas. Dentre os usos maliciosos que o Deepfake pode acarretar, enumera-se: fazer indivíduos adquirirem crenças falsas, o aumento significativo da desconfiança existente em gravações, práticas de bullying, assédio, incitar violência, ameaça à paz global, capaz de espalhar falsidades sobre líderes mundiais, forjar mapas, alterar imagens de satélites, o uso de evidências falsas em tribunais, a erosão da verdade, o aumento da desinformação, roubo de identidade, fraude, (s)extorsão, danos à democracia, perigo a segurança nacional; possibilidade de causar fraudes financeiras, golpes financeiros, manipulação de preços de ações e entre outros (BOUCHER, 2021; CHELSEY, CILTRON, 2019; FALLIS, 2021; HARTZOG, SILBEY, 2019; MAHMUD, SHARMIN, 2020; WESTERLUND).

Questões diversas emanam com o surgimento dos Deepfakes nas democracias liberais em que pode ocorrer a redução da confiança nas instituições democráticas e o agravamento de problemas sociais e políticos (RUITER, 2021). Os malefícios mais comuns seriam o uso da imagem de mulheres para a criação de vídeos pornográficos e imagens sexuais e, no campo do jornalismo e da comunicação, o funcionamento pleno da democracia é tolhido pela propagação de desinformação.

Desinformação e a criação de uma realidade falsificada

A existência das fake news remonta de séculos atrás não sendo um termo novo. A origem do vocábulo é discutida por autores sobre seu surgimento com o nascimento da imprensa, ou datado de período anterior quando o homem desenvolveu sua retórica e política, ou até mesmo quando começou a desenvolver sua comunicação. Ainda, no ano de 1894, um cartunista chamado Frederick Burr Opper, já havia demonstrado por meio de uma ilustração um indivíduo empunhando um jornal com a palavra fake news (ABBOUD, CAMPOS, NERY, 2022; TANDOC, LIM, LING, 2017; RAIS, 2020).

Uma dificuldade de definição semântica de fake news reside na atualidade. A expressão adquiriu cada vez maior diversidade de significados, sendo considerado pela *High-Level Group on Fake News and Online Disinformation* como um termo inadequado, porquanto é usado como uma forma de políticos escaparem de acusações que poderiam prejudica-los (COMISSION, 2018). A palavra fake news seria aproximado do direito - “uma mensagem propositadamente mentirosa capaz de gerar dano efetivo ou potencial em busca de alguma vantagem” (RAIS, 2020, n.p), e se insere em um campo mais amplo de desordens informacionais, recaindo dentro do conceito de desinformação.

Os autores Wardle e Derakhshan (2017) dividem os distúrbios informacionais em três tipos: *misinformation* (informações falsas compartilhadas sem a existência de querer causar dano), *malinformation* (informações verdadeiras são publicadas com o fim de prejudicar) e *desinformation* (informações sabidas como falsas são compartilhadas para causar males). Dentro deste artigo a palavra desinformação irá alcançar as descrições feitas por Wardle e Derakhshan de *malinformation*, *misinformation* e o significado já demonstrado de *desinformation*.

R. Rumos da Pesquisa, Patrocínio, v. 1, n. 7, p. 490-508, Julho, 2023.

O cenário atual midiático com o ciberespaço oportuniza o compartilhamento de notícias online, uma emergente importância no aumento do espalhamento de desinformação e da comunicação visual.

> Através do contato visual é possível de uma forma eficiente transmitir informações, se lembrar com mais facilidade de mensagens visuais do que apenas sonoras (...) a comunicação visual consegue ser percebida com menos esforço (...) processar informações melhor do que apenas mensagens verbais (...) que as maneiras visuais de comunicação são integradas melhor no cérebro do que outras formas de dados sensoriais (...) Imagens e vídeos tem um potencial maior de enganar o público devido a heurística do realismo, pois os indivíduos tratam os áudios e as imagens como podendo ser mais verdadeiros do que textos por ressoarem melhor no "mundo real" da experiência cotidiana (CHADWICK, VACCARI, 2020, p. 2, tradução nossa).

> o sistema visual do cérebro, apesar de ser bastante robusto em ambientes naturais, pode ser alvo de percepções errôneas. Exemplos clássicos incluem ilusões de ótica e figuras biestáveis, como o conhecido pato-coelho Jastow e os rostos-vaso Rubin que podem ser vistos de duas maneiras diferentes (...) se vemos algo com nossos próprios olhos, acreditamos que exista ou seja verdade, mesmo que seja improvável, como foi o caso dos exemplos de Deepfake" (KIETZMANN, KIETZMANN, LEE, MCCARTHY, 2019, p. 3, tradução nossa).

Ainda, na atualidade, existe uma crescente popularização das mídias digitais com cada vez mais indivíduos se informando através dela. Dentro deste ambiente alguns fenômenos tornam as plataformas online um solo fértil para os Deepfakes.

O primeiro destes fenômenos é atração por notícias falsas. Há uma tendência natural do compartilhamento de informações falsas por atraíram a nossa atenção a ameaças negativas e o intuito de espalha-las. Contrário à crença convencional, existe uma maior difusão de mensagens e desinformação por humanos do que por robôs. Em uma pesquisa, percebeu-se que robôs aceleraram a disseminação de informações verdadeiras e falsas em uma equânime proporção (ARAL, ROY, VOSOUGHI, 2017).

Em um estudo elaborado por pesquisadores do MIT (*Massachusetts Institute of Technology*) demonstraram que notícias falsas possuem 70% mais de chances de ser reetweetada do que uma notícia verdadeira e se espalham 6 vezes mais rápido. Outro fator relevante encontrado nesta pesquisa é que existe uma propensão maior de pessoas compartilharam notícias inovadoras e novas do que verdadeiras (ARAL, VOSOUGHI, 2017), garantindo um espaço para que possam intensificar e estimular a difusão deste tipo de mensagem por bots.

O segundo fenômeno é dividido em dois fatores: cognição cultural e as informações em cascada. Aquele é descrito como a tendência de indivíduos analisarem riscos e fatos ao seu redor através de seus valores pessoais, reforçando dinâmicas que não seriam saudáveis. Pessoas costumam acreditar em fatos que reforçam crenças pré-existentes, usualmente podem ignorar contradições as suas convicções e acreditar em informações ambíguas que a fortaleçam. Mesmo tendo a opção, elas estariam dispostas a concordar com notícias que as agradam (CARVALHO, CASTERLFRANCHIL, FAGUNDES, MALCHER, MASSARANI, MENDES, MIRANDA, LOPES, 2021). Este é um fato conveniente para a aplicação do Deepfake. Ainda, sobre as informações em cascata, ocorre quando pessoas param de crer em suas próprias informações para acreditar na avaliação de confiança de outrem sobre algo, repassando a mensagem para outros indivíduos. O ciclo se repete dando força a cascata (CHELSEY, CILTRON, 2019).

Os últimos fenômenos são a potencialização do espalhamento de desinformação por meio das bolhas online. Pessoas rodeiam-se de informações que possuem compatibilidade com suas crenças e as redes digitais intensificam este aspecto capacitando sujeitos a compartilhar e endossar novamente o conteúdo. O algoritmo ainda os cerca de informações compartilhadas por pessoas próximas e populares que possuem pensamentos parecidos, criando assim grupos homogêneos perfeitos para a utilização de micro direcionamento de Deepfakes.

> se os *Deepfakes* fossem micro direcionados para descredibilizar um político, conseguiriam modificar atitudes e decisões de um indivíduo sobre um político e seu partido (...) Foi descoberto que de fato, é possível encenar um escândalo político com um Deepfake (...) Embora especialmente a atitude em relação ao político seja diretamente afetada pelo Deepfake, as atitudes em relação ao partido do político são afetadas apenas condicionalmente (...) Mas foi descoberto também que o micro direcionamento afetou uma porção menor do que se esperava (DOBBER, METOUIL, THRILLING, HELBERGERLS VREESE, 2020, p. 82, tradução nossa).

Todos estes fatores fortalecem o uso dos Deepfakes que geram impactos no âmago do funcionamento da democracia, dos direitos fundamentais e dos direitos humanos. A democracia pode ser atingida de algumas formas: indivíduos podem acolher desinformações espalhadas online que podem se alastrar pelas redes; o exaurimento do pensamento crítico devido as pessoas não saberem com certeza quais informações são falsas ou verdadeiras, gera-se uma inabilidade dos mesmos conseguirem decidir questões políticas informadas; o dividendo do mentiroso pode dar a oportunidade de personalidades políticas negarem a responsabilidade alegando a falsidade da mídia (mesmo sendo verdadeiro); diminui a qualidade da democracia

minando a confiança das pessoas sobre as instituições que a compõe; aumenta a polarização online e distorce processos eleitorais (COLOMINA, MARGALEF, YOUNGS, 2021).

Afeta-se também direitos, como a privacidade, liberdade de pensamento, o direito à liberdade de expressão, direito a ter opiniões sem possuir interferências, direito a participar de eleições e de assuntos políticos.

O direito à privacidade pode ser atingido pela desinformação de duas maneiras: causar dano a privacidade ao determinado indivíduo a que se refere e a sua reputação, e desrespeitar o público em que a mensagem foi destinada. A tecnologia amplifica e facilita o impacto às violações a privacidade online acontecendo de forma variada em ambientes públicos e privados como também por meio de fronteiras físicas e nacionais. Possuindo novas vulnerabilidades no mundo digital, o direito à privacidade recebe agressões nas redes sociais por meio de utilização e coleta de dados pessoais online para o micro direcionamento de mensagens.

A criação de padrões mínimos para reger o tratamento de dados contra possíveis ameaças é necessário, visando um processamento de dados transparente, justo e buscar a proteção de pessoas contra a desinformação e suas consequências (danos a reputação, incitamento de violência e discriminação ou hostilidade com um grupo determinado). Assim, investigando a solução para estes problemas, ensejou-se a criação do Regulamento Geral Sobre Proteção de Dados na União Europeia (Regulamento 2016/679), que começou a vigorar em 2018, e da Lei Geral de Proteção de Dados Pessoais no Brasil (Lei nº 13.709/2018), já em vigor. Estás legislações impõem limites na utilização de dados pessoas por terceiros (ABBOUD, CAMPOS, NERY JR, 2022; COLOMINA, MARGALEF, YOUNGS, 2021).

A liberdade de expressão é outro direito fundamental que funciona como um dos pilares da democracia que inclui o direito de acesso a liberdade de imprensa e informações. O direito humano de transmitir ideias e informações não está limitado a declarações ‘corretas’, este direito também protege manifestações que possam chocar, ofender e perturbar. Assim, pode-se dizer que “as proibições gerais de divulgação de informações baseadas em ideias vagas e ambíguas, incluindo ‘notícias falsas’ ou ‘informações não objetivas’, são incompatíveis com os padrões internacionais de restrições à liberdade de expressão” (EUROPE, 2017, p. 1-5, tradução nossa).

> No entanto, a liberdade de expressão deve ser exercida em harmonia com os demais direitos e valores constitucionais. Ela não deve respaldar a alimentação do ódio, da intolerância e da desinformação. Essas situações representam o exercício abusivo

> desse direito, por atentarem, sobretudo, contra o princípio democrático, que compreende o "equilíbrio dinâmico" entre as opiniões contrárias, o pluralismo, o respeito às diferenças e a tolerância (ABBOUD, CAMPOS, NERY JR, on-line, 2022).

As decisões em uma democracia devem ser baseadas em fatos e evidências empíricas. A qualidade do diálogo se deteriora com a difusão de notícias falsas e desinformação (CHELSEY, CILTRON, 2019).

Como as representações feitas pelo Deepfakes se utilizam de marcadores centrais da personalidade e identidade do indivíduo que mexem com a estima, o respeito que rodeiam a pessoa no ambiente social que vive e o reconhecimento do valor social que possui perante terceiros; então viola-se a honra objetiva da vítima e afeta a forma que as pessoas percebem as virtudes e qualidades de seus semelhantes (PEDROSO, 2010). Visando proteger o nome, a fama e as relações interpessoais, o Código Penal prevê os delitos de calúnia (art.138, CP) e difamação (art.139, CP). O mesmo código salvaguarda os sujeitos contra infrações penais que afligem a dignidade sexual (art. 216-B, parágrafo único, CP) e o Estatuto da Criança e o Adolescente defende os inimputáveis (art. 241-C, ECA).

A maior ameaça dos Deepfakes no espalhamento de desinformações não é o compartilhamento de inverdades que faz com que pessoas vivam em um mundo irreal e tomem decisões baseado em ilusões, acreditando estar decidindo livremente, mas criar uma erosão de confiança, despertando um pérfido cinismo. Indivíduos que sabiam com antecedência que uma mídia que estavam assistindo era falsa, começavam a desconfiar de qualquer vídeo e não conseguiam distinguir vídeos falsos e verdadeiros (ARONOW, KALLA, 2021).

A dificuldade de detecção dos Deepfakes de discernir quando se está sendo enganado torna a tecnologia mais difícil de ser combatida. A margem de acerto de identificação das falsificações em um estudo foi de aproximadamente 58%, quase o mesmo de adivinhação aleatória (COZZOLINO, NIEBNER, RIESS, ROSSLER, THIES, VERDOLIVA, 2019). Outra pesquisa analisou o nível de acerto de detecção de Deepfakes entre humanos e máquinas. O resultado demonstrou que a criação de Deepfakes já tem um enorme grau de realismo e confundiria a maioria do público (MARCEL, KORSHUNOV, 2020, p. 3-5, tradução nossa).

Quando a convicção do que é visto não se traduz na realidade, nem o mais legitimo discurso poderá ser acreditado. Pode ser cultivado dentre os cidadãos a suposição que não seja

possível ter uma base em que a verdade possa ser estabelecida (BOUCHER, 2021; CHADWICK, VACCARI, 2020).

Sobre o exercício de direitos políticos, os Estados devem garantir que os cidadãos exerçam sem interferências a liberdade de opinião e pensamento sem discursos de ódio. Assim, indivíduos são capazes de ter opiniões sem ser induzido, podendo ser independentes e livres de violências ou ameaças. A proteção destes direitos contra a desinformação é devida, pois as interposições ilegítimas e não justificáveis de informações podem influenciar a mente das pessoas e suas futuras decisões (COLOMINA, MARGALEF, YOUGS, 2021).

A inexistência de uma bala de prata

Uma das formas de combate ao Deepfake é o "mercado de ideias", que concepções diferentes poderiam ser requintadas, sendo concebidas e circular sem restrições pelo espaço público em que irá confronta-se com a verdade (ABBOUD, CAMPOS, NERY JR, 2022). Malgrado a percepção de que a verdade irá aflorar por meio do mercado, por certo, tem sido demonstrado que a busca da verdade é prejudicada com discursos manifestamente dissimulados. O causador disto é o fenômeno conhecido como "verdade ilusória", em que reiteradas declarações são mais fáceis de serem processadas do que novas declarações, podendo levar pessoas a conclusões errôneas de que aquelas são mais verdadeiras. Um estudo realizado demonstrou os efeitos da repetição de pensamentos plausíveis aumenta a probabilidade de estes pensamentos serem creditados como válidos ou verdadeiros (GOLDSTEIN, HASER, TOPPINO, 1977).

Destarte, a exibição contínua de informações falsas a indivíduos, independente da motivação ou finalidade de demonstrar o erro ou contraditar o apresentado, torna maior as chances de a informação ser lembrada como verdadeira e o conhecimento prévio em determinado tópico pode prejudicar ao invés de auxiliar na sua compreensão (ARKES, HACKETT, BOEHM, 1989; BRASHIER, FAZIO, MASH, PAYNE, 2015).

O maior problema do mercado de ideias é a existência de ideias falsas que são irrespondíveis, não sendo possível confronta-las com ideias melhores ou diferentes, como por exemplo, é feito por uma representação, altamente verossímil com a realidade (Deepfake), de uma pessoa em atividade íntima (FRANKS, WALDMAN, 2019).

Desse modo, embora o mercado de ideias falhe em suscitar soluções ao Deepfake, ainda existem formas de intervenção externas capazes de combatê-lo, sendo por meio da própria tecnologia uma opção. Hodiernamente, subsistem vários métodos capazes de identifica-los, sendo a maioria deles através do *deep learning* (MAHMUD, SHARMIN, 2020). O sistema de detecção poderia ser dividido em um extrator e um classificador de faces, este que é baseado em rede neural (CAO, GONG, 2021). Buscando lidar com a ameaça tecnológica a *Defense Advanced Research Projects Agency* (DARPA), em 2019, iniciou pesquisas sobre o programa MediFor com a finalidade de alcançar uma detecção automatizada da integridade de uma mídia digital (CORVEY, 2019).

Na procura de alcançar efetivas respostas à tecnologia, operadores do direito debatem a adoção de medidas legais específicas. Em diversos países, já há a discussão e propostas legislativas de elaborações de diretrizes para regular e proteger os direitos fundamentais, restringindo o acesso prejudiciais e tentando não fazer uma proibição geral dos Deepfakes, como algumas das recentes reformas legislativas copiladas por PAVIS (2021) - Lei de 2018, S.3805 - *Malicious Deep Fake Prohibition*; lei de 2017, n° 29 - *Crimes Amendment* (*Intimate Images*); lei de 2019, n° 004 - *Criminal Law Amendment* (*Intimate Images*) e o *Deepfake Accountability Act* - H.R.3230.

Um dos projetos mais relevantes nos últimos anos sobre o tema foi o da Comissão Europeia. O projeto *Artificial Intelligence Act*, proposto no dia 02 de abril de 2021, classifica os Deepfakes como não sendo de alto risco e deixa vago sobre sua possível classificação como 'proibida'. Por outro lado, os softwares de detecção dos Deepfakes se enquadram como de alto risco.

> A lei possibilita os usos do *Deepfake* se obtiver os requisitos mínimos de transparência (...) como o art. 52, n. º 3, que obriga criadores rotular os conteúdos modificados para o público saber que a mídia foi previamente alterada (...) Contudo, a proposta ainda não possui diretrizes para o controle da divulgação dos *Deepfakes*, não especificando como a rotulagem das mídias deve ser feita, como os provedores online iram gerir as suas postagens e seu papel na visualização do rótulo. Carece também de tipificar as sanções que serão aplicadas aos agentes que violarem a lei (...) o art. 71 não especifica como será aplicada as punições (BOUCHER, 2021, p. 38 tradução nossa).

O risco intolerável aos direitos fundamentais, a segurança social e política de um país, deveriam servir de base para a classificação dos Deepfakes como de alto risco, sendo necessário

futuras mudanças e esclarecimentos sobre quais seriam as práticas realizadas por meio da IA que deveriam ser regulamentadas no projeto de lei.

Atualmente, os projetos de leis que tratam especificamente da inteligência artificial e os Deepfakes no Brasil são incipientes, não existindo as adequações e modificações especificas no ordenamento jurídico brasileiro, contudo o compêndio de leis existente já é capaz de assegurar os direitos dos indivíduos, como por exemplo, a Constituição Federal, o Código Penal e o Estatuto da Criança e do Adolescente.

O combate ao Deepfake deve ser feito através de diversos campos da sociedade (principalmente por meio da tecnologia e a lei), pois não existe 'bala de prata' para por fim aos malefícios das mídias criadas pela inteligência artificial.

CONSIDERAÇÕES FINAIS

Nesta pesquisa é possível se obter uma conclusão de um argumento dedutivo através da verdade da conjunção das premissas expostas. O argumento é dedutivo quando é formado com informações obtidas das premissas e sendo verdadeiras as premissas, a conclusão será necessariamente verdadeira, sendo impossível ela ser falsa.

Com base nas hipóteses apresentadas é admissível concluir que em razão dos impactos do Deepfake: o estado poderá não ser baseado em evidências empíricas e será possível ser criado uma legislação para proteger as pessoas contra os usos irrestritos da tecnologia, não existindo há necessidade de sua confecção, todavia ela serviria como um robustecimento dos direitos já protegidos pelo ordenamento jurídico. Ainda, deduz-se das premissas e do artigo que os indivíduos não conseguem identificar eficientemente as mídias sintetizadas e existem formas eficazes para combater os Deepfakes.

FINANCIAMENTO

Esse projeto faz parte do programa de Iniciação Científica do UNICERP (PROIC) 2021/2022, financiado pela Fundação Comunitária, Educacional e Comunitária de Patrocínio.

REFERÊNCIAS

ABBOUD, Georges; CAMPOS, Ricardo; NERY JR, Nelson. **FAKE NEWS E REGULAÇÃO**. 3. ed. Revista dos Tribunais Ltda, 2022. Disponível: https://proview.thomsonreuters.com/. Acesso em 14 de setembro 2021.

AHMED, Saifuddin. **Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism.** Disponível : https://journals.sagepub.com/doi/abs/10.1177/14614448211019198. Acesso em 10 de maio 2022.

AJDER, Henry; CAVALLI, Francesco; GRAPHIC, Cullen; PATRINI, Giorgio. **Deeptrace: The State of Deepfakes Landscape, Threats, and Impact**. Disponível: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.
ARAL, Sinan; ROY, Deb; VOSOUGHI, Soroush.THE SPREAD OF TRUE AND FALSE NEWS ONLINE. Disponível: https://www.science.org/doi/10.1126/science.aap9559. Acesso em 23 de fevereiro 2022.

ARONOW, Peter M ; KALLA, Josua ; Ternovski. **Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments**. Disponível : https://osf.io/dta97/. Acesso em 4 de fevereiro 2022.

ARKES, H. R., Hackett, C., & Boehm, L. (1989). **The generality of the relation between familiarity and judged validity**. http://dx.doi.org/10.1002/bdm.3960020203. Acesso em 7 de Julho 2022.

BRENNER, Stefan; FILKUVOVÁ, Petra; LANGGUH, Johhannes; POGORELOV, Konstantin; SCHROEDER,Daniel Thilo. **Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes**. Disponível: https://www.frontiersin.org/articles/10.3389/fcomm.2021.632317/full. Acesso em 20 de dezembro 2021..

BOUCHER, PHILIP. **Tackling deepfakes in European policy**. Disponível: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039. Acesso em 3 de outubro 2021.

BRASHIER, Nadia M; FAZIO, Lisa K; MARSH, Elizabeth J., PAYNE, B. Keith. **Knowledge Does Not Protect Against Illusory Truth**. Disponível: https://psycnet.apa.org/record/2015-38275-001. Acesso em 5 de julho 2022

CAO, Xiaoyu; GONG, Neil Zhenqiang. **Understanding the Security of Deepfake Detection**. Disponível: https://link.springer.com/chapter/10.1007/978-3-031-06365-7_22. Acesso em 14 de dezembro 2021

CARVALHO, Vanessa Brasil de; CASTELFRANCHIl, Yurij; FAGUNDES, Vanessa Oliveira; MALCHER, Maria Ataide; MASSARANI, Luisa; MENDES, Ione Maria; MIRANDA, Fernanda Chocron; LOPES, Suzana Cunha. **Jovens e sua percepção sobre fake news na ciência**. Dispomível: https://www.scielo.br/j/bgoeldi/a/PqdXRfWRLjpSZLGqvBfzzgF/?lang=pt. Acesso em 14 de dezembro 2021.

CHADWICK, Andrew ; VACCARI,Cristian. **Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News**. Disponível: https://journals.sagepub.com/doi/full/10.1177/2056305120903408. Acesso em 15 de novembro 2021.

CHELSEY, Robert; CILTRON, Danielle K. Deep Fakes: **Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security**. Disponível: https://scholarship.law.bu.edu/faculty_scholarship/640/. Acesso em 8 de dezembro 2021.

COLOMINA, Carme; MARGALEF, Héctor Sánchez; YOUNGS, Richard. **The impact of disinformation on democratic processes and human rights in the world**. Disponível: https://www.europarl.europa.eu/thinktank/en/document/EXPO_STU(2021)653635 . Acesso em 17 de dezembro 2021.

COMISSION, European. **A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation**. Disponível: https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1. Acesso em 20 de fevereiro 2022.

CORVEY, William. **Media Forensics (MediFor) (Archived)**. Disponível: https://www.darpa.mil/program/media-forensics. Acesso em 19 de Julho 2022.

COZZOLINO, D. ; NIEBNER, M. ; RIESS, C. ; ROSSLER, A. ; THIES, J. ; VERDOLIVA, L. **FaceForensics++: Learning to Detect Manipulated Facial Images**. Disponível: https://arxiv.org/abs/1901.08971. Acesso em 5 de junho 2022.

DAMIANI. Jesse. **A Voice Deepfake Was Used To Scam A CEO Out Of $243,000**. Disponível: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/. Acesso em 25 de outubro 2020

DELHI, New. **I was vomiting: Journalist Rana Ayyub reveals horrifying account of deepfake porn plot**. Disponível: https://www.indiatoday.in/trending-news/story/journalist-rana-ayyub-deepfake-porn-1393423-2018-11-21. Acesso em 15 de julho 2022.

DIAKOPOULOS, Nicholas; JOHSON, Deborah G. **Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections**. Disponível: https://journals.sagepub.com/doi/abs/10.1177/1461444820925811. Acesso em 9 de dezembro 2021.

DONOVAN, Joan; PARIS, Britt. **DEEPFAKES AND CHEAP FAKES: The Manipulation of Audio and Visual Evidence**. Disponível: https://www.hks.harvard.edu/publications/deepfakes-and-cheap-fakes. Acesso em 9 de dezembro 2021.

EUROPE, COMISSION. **Joint declaration on freedom of expression and "fake news", disinformation and propaganda**. Disponível: https://www.osce.org/fom/302796. Acesso em 5 de julho 2022.

Fallis, D. **The Epistemic Threat of Deepfakes**. Philos. Technol. 34, 623–643 (2021). https://doi.org/10.1007/s13347-020-00419-2. Disponível: https://link.springer.com/article/10.1007/s13347-020-00419-2. Acesso em 17 de dezembro 2021

FLORIDI, Luciano. **Artificial Intelligence, Deepfakes and a Future of Ectypes**. Disponível : https://link.springer.com/article/10.1007/s13347-018-0325-3. Acesso em 17 de dezembro 2021.

FRANKS, Mary Anne; WALDMAN, Ari Ezra. **Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions**. Disponível: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3445037. Acesso em 11 de dezembro 2021.

FRAZÃO, Ana; MULHOLLAND, Caitlin. **Inteligência Artificial e Direito: Ética, Regulação e Responsabilidade**. 2. ed. Revista dos Tribunais Ltda, 2020. Disponível: https://proview.thomsonreuters.com/. Acesso em 14 de setembro 2021.

GOLDSTEIN, David ; HASHER, Lynn ; TOPPINO, Thomas . **Frequency and the conference of referential validity**. Disponíve : https://psycnet.apa.org/record/1978-02525-001. Acesso em 5 de julho 2022.

GUARDIAN. **Mother charged with deepfake plot against daughter's cheerleading rivals**. Disponível : https://www.theguardian.com/us-news/2021/mar/15/mother-charged-deepfake-plot-cheerleading-rivals. Acesso em 5 de julho 2022.

HARTZOG, Woodrow; Silbey, Jessica. **The Upside of Deep Fakes**. Disponível: https://scholarship.law.bu.edu/faculty_scholarship/1072/. Acesso em 11 de dezembro 2021.

HISTORY SPEAKS. Holmes. **Dissenting in Abrams v. United States, 1919**. Disponível : https://firstamendmentwatch.org/history-speaks-holmes-dissenting-abrams-v-united-states-1919/. Acesso em 4 de julho 2022.

KIETZMANN, Jan; KIETZMANN, Tim; LEE, Linda; MCCARTHY, Ian. **Deepfakes: Trick or treat?**. Disponível: https://www.sciencedirect.com/science/article/abs/pii/S0007681319301600. Acesso em 15 de novembro 2021.

LEE, Wenke; MIRSKY, Yisroel. **The Creation and Detection of Deepfakes: A Survey**. Disponível: https://arxiv.org/abs/2004.11138. Acesso em 15 de novembro 2021.

MAHMUD, Bahar Uddin; SHARMIN, Afsana .**Deep Insights of Deepfake Technology : A Review**. Disponível: https://www.semanticscholar.org/paper/Deep-Insights-of-Deepfake-Technology-%3A-A-Review-Mahmud-Sharmin/44c2ae65d9ac6fe4a10349547bc0d3a7a7bbbf35. Acesso em 15 de novembro 2021.

MARCEL, Sébastien ; KORSHUNOV, Pavel. L. **Deepfake detection: humans vs. Machines**. Disponível: https://arxiv.org/abs/2009.03155. Acesso em 5 de fevereiro 2022.

MCCARTHY, J.; MINSKY, M. L.; ROCHESTER, N.; SHAMNON, C. E. **A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE**. Disponível: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf. Acesso em 26 de dezembro 2021.

PAVIS, Mathilde. **Rebalancing our regulatory response to Deepfakes with performers' rights**. Disponível : https://journals.sagepub.com/doi/full/10.1177/13548565211033418. Acesso em 17 de dezembo 2021.

PEDROSO, Fernando de Almeida. **Crimes contra a honra.** Doutrina essenciais de Direito Penal. São Paulo: RT. Volume 5. 2010.

Tom Dobber1 , Nadia Metoui1, Damian Trilling1 , Natali Helberger1, and Claes de Vreese1
DOBBER, Tom; HERBERGER, Natali; METOUI, Nadia; TRILLING, Damian; VREESE, Claes de. **Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?**. Disponível : https://journals.sagepub.com/doi/full/10.1177/1940161220944364. Aceso em 15 de novembro 2021.

RAIS, Diogo. **Fake news**. 2. ed. Revista dos Tribunais Ltda, 2020. Disponível: https://proview.thomsonreuters.com/. Acesso em 14 de setembro 2021.

RINI, Regina. **Deepfakes and the Epistemic Backstop**. Disponível: https://philpapers.org/rec/RINDAT. Acesso em 11 de dezembro 2021.

RUITER, Adrienne de. **The Distinct Wrong of Deepfakes**. Disponível: https://link.springer.com/article/10.1007/s13347-021-00459-2. Acesso em 15 de novembro 2021.

SPIVAK, Russell. **DEEPFAKES": THE NEWEST WAY TO COMMIT ONE OF THE OLDEST CRIMES**. Disponível: https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=&v=2.1&it=r&id=GALE%7CA592039891&sid=googleScholar&linkaccess=abs&userGroupName=anon%7E866c767e. Acesso em 11 de dezembro 2021.

TANDOC, Edson C.; LIM, Zheng Wei; LING, Richard. **Defining "Fake News": A typology of scholarly definitions**. Disponível:

https://journals.scholarsportal.info/details/21670811/v06i0002/137_dn.xml&sub=all. Acesso em 12 de janeiro 2022.

WARDLE, Claire; DERAKHSHAN, Hossein. **INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making**. Disponível: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html. Acesso em 20 de fevereiro 2022.

WESTERLUND, Mika. **The Emergence of Deepfake Technology: A Review**. Disponível: https://www.semanticscholar.org/paper/The-Emergence-of-Deepfake-Technology%3A-A-Review-Westerlund/17734113f254a64b3bae312713edba3b1e34fb56. Acesso em 15 de novembro 2021.