Griot : Revista de Filosofia, Amargosa - BA, v.25, n.2, p.107-122, junho, 2025 ISSN 2178-1036
https://doi.org/10.31977/grirfi.v25i2.5296
Recebido: 13/02/2025 | Aprovado: 29/04/2025
Received: 02/13/2025 | Approved: 04/29/2025

TODA FONTE PODE ESTAR ENVENENADA: SOBRE DEEPFAKES E A AMEAÇA DE APOCALIPSE EPISTÊMICO

Bismarck Bório de Medeiros[1]
Universidade Federal de Santa Maria (UFSM)
https://orcid.org/0000-0002-4473-9849
E-mail: bismarckborio@gmail.com

RESUMO:
Este artigo visa argumentar sobre o problema das denominadas *deepfakes* em ambientes epistêmicos e como tal fenômeno é uma ameaça às estruturas de obtenção, armazenamento, divulgação de informação e aquisição de conhecimento. Para isso, faremos uma análise das principais semelhanças e distinções ao fenômeno das *fake news* e como ela é um tipo de desinformação substancialmente perniciosa, fazendo com que haja perda generalizada tanto de confiabilidade em qualquer ambiente virtual em que possa incidir tais falsificações, quanto de credibilidade dos agentes epistêmicos envolvidos na *deepfake*. Desta forma, buscamos embasar nossa posição às críticas sobre a possibilidade deste apocalipse epistêmico com investigações históricas sobre casos diversos de enganação e falsificação de moeda corrente com consequências sociais, econômicas e políticas. Pontua-se com base nesta análise que procedimentos de autenticação estão sempre presentes e podem ser implementados para mitigar o espalhamento de desinformação nas redes, adotando uma posição moderada com relação a esta ameaça.

PALAVRAS-CHAVE: Apocalipse Epistêmico; Deepfake; Epistemologia Social; Fake News; História da Falsificação.

EVERY WELL (SOURCE) CAN BE POISONED: ON DEEPFAKES AND THE THREAT OF EPISTEMIC APOCALYPSE

ABSTRACT:
This paper intends to argue about the problem of deepfakes in epistemic environments and how this phenomenon is a threat to the structures for obtaining, storing, disseminating information and acquisition of knowledge. To do this, we make an analysis of the main similarities and distinctions to the phenomenon of fake news and how it is a type of disinformation substantially harmful, causing a widespread loss of reliability in any virtual environment in which such incidents may occur falsifications and loss of the epistemic credibility involved in the deepfake. In this way, we seek to support our position to criticisms about the possibility of this epistemic apocalypse, with historical investigations into various cases of deception and falsification of currency with social, economic and political consequences, highlighting based on this analysis in authentication procedures can be implemented to mitigate the spread of disinformation on the networks and adopting a moderate position regarding this threat.

KEYWORDS: Counterfeit History; Deepfake; Epistemic Apocalypse; Fake News; Social Epistemology.

[1]Doutorando(a) em Filosofia na Universidade Federal de Santa Maria (UFSM), Santa Maria – RS, Brasil.

Introdução

O desenvolvimento de tecnologias de *deep learning* voltadas a criação de imagens, textos, áudios e vídeos a partir de uma robusta base de dados está em amplo crescimento nestes últimos anos[2]. Podemos vê-las como ferramentas com usos potenciais positivos, como auxiliar pessoas que perderam a fala se expressarem através de sintetizadores de voz (Negi et al., 2021); para apontar vulnerabilidades em sistemas informacionais – como checadores de fatos – e em práticas institucionais, como no sistema educacional (Citron e Chesney, 2019). Há um subgrupo destas tecnologias que têm o potencial de gerarem o que está sendo chamado de *deepfake* (termo derivado das palavras "*deep learning*" + "*fake*"), abrangendo uma variedade de conteúdos informacionais apresentados como dados sensórios (visual e auditivo) manipulados ou construídos para simular uma gravação – i.e, uma mídia sintética.

Frequentemente usada para entretenimento, as *deepfakes* também são indevidamente aplicadas para fins maliciosos como fraude, falsificação de identidade e produzir consenso de determinado público, sendo utilizada como arma (weaponization)[3] contra determinada pessoa ou grupo social. Neste artigo, defendemos que as *deepfakes* são relativamente distintas das *fake news* e precisam ser levadas a sério no debate político e filosófico por corromper ambientes epistêmicos[4] permissivos a sua circulação em níveis que podem tornar sua confiabilidade inviável, descredibilizando o ambiente e os agentes nele, bem como pondo em dúvida qualquer informação ali presente[5].

Desde o início e expansão da Internet 2.0, o receio da perda de credibilidade das fontes de informação devido a criação e compartilhamento de conteúdo pelo usuário foram ficando cada vez mais latentes, e as discussões sobre o tema, frequentes[6]. Com o aumento da tecnologia embarcada em dispositivos portáteis com acesso a internet – como os smartphones – a ascensão das redes sociais e o desenvolvimento de estratégias como disparos em massa e automático de mensagens e com direcionamento de informações aos interesses de públicos-alvo baseados no histórico de seus acessos (*microtargetting*), a enxurrada de desinformação tomou conta das redes. Pesquisas mostram que, devido a fatores emocionais e de novidade associada a propagação de desinformação, notícias falsas atingem com mais rapidez agentes epistêmicos destes ambientes do que as verdadeiras (Vosougui, Roy e Aral, 2020), bem como *bots* sociais alavancam notícias de baixa credibilidade nas redes, manipulando o algoritmo para deixá-las com mais visibilidade[7]. O combate às *fake news* tomou uma nova proporção. Porém, até pouco tempo, a elaboração noticiada de desinformação era por meio de texto ou apenas comunicada. Agora, estamos entrando em uma nova época onde, com as *deepfakes*, a barreira virtual entre o fato e a ficção está tomando proporções fortemente distintas das *fake news*.

2Para uma noção geral da produção de *deepfakes*, suas variações e aplicações, ver (Farid, 2022). Sobre desenvolvimento e detecção, ver (Mirsky e Lee 2020).

3Este é um conceito que já está sendo abordado atentamente, principalmente por órgãos governamentais de segurança, produzindo protocolos e folhetos como o da Polícia Europeia em Europol (2022) e do Centro Internacional de Contra-Terrorismo em Busch e Ware (2023), com uma ampla quantidade informações acerca da ameaça não só virtual, mas social, política e econômica das deepfakes.

4Nossas noções são semelhantes às de (Anderau, 2023). Definimos aqui por ambiente epistêmico como qualquer ambiente constituído por fatos que nos auxiliem devido a sua estrutura a obter certo valor epistêmico – como crença, conhecimento ou compreensão – nele, bem como um agente epistêmico como qualquer sujeito que participe passiva ou ativamente de algum ambiente epistêmico, com capacidades cognitivas para exercer qualquer atividade de cunho epistêmico que tal ambiente o permita (obter, transmitir ou compartilhar informações, por exemplo).

5Todo o mote deste artigo gira em torno deste tema e como abordá-lo. Recomendo inicialmente para rápida leitura a matéria de (Hao, 2019).

6 Por exemplo, o protocolo já elaborado em (Fogg, 2002) e as discussões em (Lankes, 2006) e (Rieh, 2010).

7Para mais detalhes sobre estas manipulações algorítmicas ver (Himelein-Wachowiak et al., 2020), (Shao et al., 2018) e (Ferrara et al., 2016).

Iniciaremos a próxima seção apontando distinções acerca do fenômeno das *fake news* e das *deepfakes*. Estabelecido algumas características que as diferenciam, na seção três mostraremos porque as *deepfakes* seriam mais nocivas ao afetar a confiabilidade da aquisição de conhecimento em nossos ambientes epistêmicos, apresentando a tese do poço envenenado generalizado. Na seção quatro, abordaremos possíveis soluções e críticas a esta ameaça, trazendo na seção seguinte uma abordagem distinta cultivada na literatura – voltada a epistemologia de instrumentos e gravações – sobre a história da falsificação, onde trazemos exemplos visando uma análise comparada aos efeitos perniciosos das *deepfakes* na sociedade. Baseado nesta análise, finalizamos com observações voltadas ao estabelecimento de protocolos de autenticação de informação como uma solução possível e de fato efetiva a esta ameaça.

Deepfake e fake news: distinções fundamentais e problemas

A noção de *fake news* tem origem humorística, destacada em esquetes americanas de tabloides revistas fictícias, porém devido a sua expressão e uso dentro da esfera política[8], foi significada informalmente como a imprensa ou as próprias notícias propagadas como literalmente falsas ou que transmitem falsa informação. Há uma gama de definições disponíveis de *fake news*, obtendo na literatura uma separação definicional das *fake news* em três grupos de abordagens, que são destacadas pelo filósofo Sven Bernecker em (2021, pp. 6-7) como:

a) *híbrida*: não há verdade nem veracidade na notícia que está sendo divulgada, assim transmite-se informação falsa com intenção em enganar ou sem qualquer preocupação com a verdade;
b) *privativa*: as notícias falham em ser genuínas devido a falta de expertise jornalística padrão envolvida no processo;
c) *centrada no consumidor*: abordagem definida pela disposição ao engano de seus consumidores.

Aqui, resolvemos não restringir nossa abordagem a termos que se comprometam com apenas a função exercida por certos agentes epistêmicos (jornalistas) ou a intencionalidade do agente como diretriz determinante. Considerarmos uma definição mais ampla e informacional, i.e., a abordagem híbrida (Jaster e Lanius, 2021), substituindo apenas o que é tomado como informação falsa por desinformação. Temos, inclusive, uma descrição bem definida do termo para considermos – sendo a mesma usada pela mídia social Facebook:

> Conteúdo de informação impreciso ou manipulado que é espalhado intencionalmente. Isto pode incluir notícias falsas [fake news] ou envolver métodos mais sutis, como operações de *false flag*, alimentar citações imprecisas ou histórias a intermediários inocentes ou amplificar conscientemente informações enviesadas ou enganosas. A desinformação é distinta da má-informação, que é a disseminação inadvertida ou não intencional de informações imprecisas sem intenção maliciosa (Weedon, Nuland e Stamos, 2017, p. 5, tradução nossa).

Mesmo a noção de desinformação sendo norteada pela intencionalidade do agente epistêmico, vimos mais acima que a abordagem híbrida pode ser intencional ou não, envolvendo em ambos os casos a falta de comprometimento com a verdade, e não exatamente se o que é

8A primeira vez que o termo foi tomado publicamente com este sentido e propagou-se foi em uma conferência de imprensa onde o ex-presidente dos EUA, Donald Trump, acusou a CNN de "serem *fake news*". Disponível em https://www.theguardian.com/us-news/2017/jan/11/trump-attacks-cnn-buzzfeed-at-press-conference.

transmitido é falso em sentido predicativo. *Fake news* podem ser utilizadas para obter vantagens políticas, econômicas ou simplesmente para falar besteira de forma descompromissada (*bullshitting*)[9]. Pesquisas mostram que boa parte das *deepfakes* acabam utilizadas em contextos que se enquadram fundamentalmente nesta mesma postura descompromissada, pois mesmo que sejam feitos vídeos com a intenção de retratar fielmente algum fato ou que seja feita a sintetização da voz de alguém com perda da fala para o próprio se comunicar, são propriamente material gerado ou manipulado por aprendizado de máquina para cumprir certa função[10].

Ora, é justamente nesse cumprimento que as *deepfakes* inicialmente se diferenciam das *fake news*. Esta última envolve a esfera epistêmica do que seria predicativo dado certo relato acerca de algo – tendo significação cognitiva, contexto linguístico e caráter mais testemunhal com maior formação de crença relativa às comunidades nas quais ela é estabelecida. Já as *deepfakes* têm um potencial danoso, aproximando-se em aparência com evidências e justificativas reconhecíveis perceptual e sensorialmente em um simulacro de gravação. São feitas por meio da tecnologia de aprendizado de máquina para trazer impressão ou pretensão de autenticidade – podendo devido a isso, a partir ou com base nelas, noticiar e propagar falsidades – distinguindo-se epistemicamente (ao menos em graus como fonte de justificação) uma da outra.

Podemos trazer uma analogia que pode caracterizar esta distinção melhor para, mais a frente no artigo, nos voltarmos a uma análise preventiva de como lidar com ambas: a análise dos usos atributivo e predicativo do termo 'falso', conhecida pelos lógicos medievais e retomada por (Geach, 1956) e (Floridi, 2011). Porém, diferente de Floridi, quando falamos que uma certa notícia F é falsa e a caracterizamos como *fake news*, o uso da falsidade para nós é predicativo[11]. Já no caso das *deepfakes*, havendo uma imagem, áudio ou vídeo D, quando dizemos que D é falso, o uso do termo seria atributivo. Assim, também temos com tal distinção que este uso atributivo pode ser positivo ou negativo. Em caso positivo, o atributo qualifica acertadamente o objeto. Quando negativo, ele não apenas nega a atribuição, mas desqualifica o objeto. As consequências, portanto, não seriam só epistêmicas, mas ontológicas. Floridi consegue ilustrar bem esta ideia, e em seu final podemos relacionar diretamente a características das *deepfakes*:

> Por exemplo, um falso policial (uso atributivo) claramente não é um tipo específico de policial, mas de forma alguma é um policial (uso negativo), embora a pessoa que finge ser um policial pode desempenhar com sucesso todos os deveres de um policial genuíno [...] O mesmo vale para outros exemplos, como "notas forjadas", "assinatura falsa", "alarme falso" e assim por diante. São todos exemplos de resposta correta para 'não, é um F(x)' para a pergunta tipo 'este é um x genuíno?'. (Floridi, 2011, p. 97, tradução nossa)

Dado esta característica, podemos destacar outra distinção entre ambas: há uma tendência

9 O ensaio de Harry Frankfurt sobre falar merda (2025) toca justamente neste ponto.

10A utilização de *deepfakes* como arma, ou armificação (*weaponization*) contra mulheres – principalmente com a criação de vídeos pornográficos – vem tornando-se uma prática assustadoramente corriqueira, como mostrado em uma análise feita pelo site Home Security Heroes: foi relatado que mais de 98% das deepfakes encontradas online em 2023 são pornográficas (disponível em https://www.homesecurityheroes.com/state-of-deepfakes/) promovendo um imaginário falsificado e de vilipêndio às mulheres (Martínez e Padilla-Castillo, 2019) e sua objetificação (Rini e Cohen, 2021), colocando-as em situação de fragilidade psicológica e social. Este crime ainda encontra dificuldades em sua tipificação, julgamento e punição, como destacado por (Harris, 2019), sendo erroneamente descrito e relatado pelas mídias jornalísticas, não auxiliando na conscientização e prejuízos que esta prática misógina pode causar como dito por (Gosse e Burkell, 2020).

11O que isto traz como consequência para nossa análise? Que as noções de má-informação e desinformação também são um tipo de informação, diferentemente da posição de Floridi. Lembramos aqui que a argumentação de Floridi é justamente buscando defender que o conceito de informação semântica não possui neutralidade alética, i.e., informação por definição é verdadeira. Desta forma, para Floridi, quando falamos sobre falsa informação, estamos a falar sobre outra coisa que não seria informação de fato, tendo neste contexto o termo 'falso' uso atributivo. Porém, não defendemos esta posição, advogando em favor de uma versão mais fraca do conceito de informação semântica como dados bem formados significativos. Portanto, mesmo considerando um bom argumento, não achamos válido que, para nos referirmos a informação semântica, o termo 'falso' tenha uso atributivo.

de considerarmos evidências sensoriais e perceptivas obtidas por meio de instrumentos de gravação sobre estados-de-coisas como fontes de justificação básica que tornam confiáveis determinadas justificações por testemunho[12]. O que já não ocorreria tanto substancialmente quanto em graus – ao menos em parte das circunstâncias[13] – com as *fake news* devido a sua própria natureza proposicional. Sem este suporte pela evidência, estes testemunhos perdem confiança (Leonard, 2021). Em seu artigo, Don Fallis (2020) aponta que o não reconhecimento de uma *deepfake* como tal acarreta em falsas crenças acerca de algum estado-de-coisas e/ou sobre outro agente epistêmico, gerando inclusive incerteza acerca de gravações ou testemunhos legítimos (Fallis, 2020, p. 625-26). Desta forma, gravações de forma geral passam a portar menos informação e solapa-se a confiabilidade de nossos processos de formação de crenças, bem como a credibilidade dos agentes epistêmicos e nos testemunhos que estiverem contidos no *deepfake* (seja em vídeo ou áudio). Isso impulsiona agentes epistêmicos virtuosos a realizarem uma suspensão global de juízo dentro destes ambientes e prejudicando a utilização legítima de gravações como salvaguarda epistêmica (*epistemic backstop*)[14]. Assim, como ponto complementar, podemos tomar como tese de que um agente epistêmico vicioso, intencionalmente ou não, pode ter uma crença pungente em um argumento e, contrariamente a uma virtuosa suspensão de juízo, usar instrumentalmente *deepfakes* como evidências para reforçá-la ou como salvaguarda epistêmica.

Dito isso, vamos trazer aqui um exemplo que em geral aponta como ferramentas de aprendizagem de máquina, mesmo que para entretenimento, tem o potencial de auxiliar na propagação de *fake news* (independente se for intencional ou não). Ano passado, o pesquisador e filósofo brasileiro Giovanni Rolla postou em seu perfil pessoal no Twitter uma foto gerada pelo aplicativo Midjourney do ator Brad Pitt e a atriz Eva Green caracterizados respectivamente como Jean-Paul Sartre e Simone de Beauvoir[15]. Tudo não passava de uma brincadeira, porém em alguns dias houveram milhares de compartilhamentos com as mais diversas opiniões, já com uma *fake news* intrínseca de que o filme estaria em fase de produção em Hollywood. Percebam que há uma lacuna, sendo esta preenchida quando considera-se a imagem (*deepfake*) como justificativa às afirmações feitas por certos agentes epistêmicos na rede social que o filme estaria em produção em Hollywood (*fake news*).

Acerca do ponto complementar no parágrafo anterior, muitos usuários da rede momentaneamente discutem e problematizam como Hollywood invisibiliza e descaracteriza aspectos das personalidades sobre as quais propõe-se filmar uma biografia (como a famigerada feiura de Sartre). Os usuários (aqui, reconhecidos por nós como agentes epistêmicos) utilizaram como evidência ou salvaguarda epistêmica para este argumento a própria *deepfake*. Desta maneira, entra o primeiro potencial substancialmente pernicioso a ser discutido: sua produção e utilização leva a uma situação em que não há qualquer garantia segura de que, nos sistemas de informação existentes, a evidência que eu posso apresentar para defender ou contrapor uma afirmação como verdadeira não seja uma *deepfake* (Fallis, 2020, p. 625-26), minando a confiança

12Há pouca literatura acerca do tema. A abordagem mais proeminente encontra-se em (Audi, 2003).

13Quando falamos aqui em graus, nos referimos a certas situações em que mesmo que haja certo testemunho envolvido que possa ser considerado por alguns fonte de justificação primária, para outros demandaria-se outra fonte primária que não a testemunhal, não havendo consenso epistêmico. Não entrarei aqui nas discussões sobre as posições reducionistas ou antirreducionistas sobre a consideração do testemunho como fonte básica de justificação ou conhecimento. Partimos de uma ideia mais flexibilizada de que podemos admitir parte do conhecimento por testemunho precisam como fonte básica a evidência sensorial ou perceptual. Para uma exposição e revisão desta discussão ver (Lackey, 2006) e (Leonard, 2021).

14A noção de uma gravação ter função de salvaguarda epistêmica trata-se do seu potencial de regulagem de nossas práticas de testemunho, i.e., áudios, vídeos e imagens tem a função de modular as virtudes exercidas no testemunho, como a sinceridade e qualidade do relato, sob pena de ter sua confiabilidade abalada quando o testemunho não for condizente com as gravações existentes. Os *deepfakes*, por motivos óbvios, põem esta salvaguarda em crise. Para mais detalhes ver (Rini, 2020).

15Podemos ver a repercussão internacional pela matéria exibida no jornal *La República*, está disponível em https://larepublica.pe/verificador/2023/03/21/brad-pitt-y-eva-green-seran-jean-paul-sartre-y-simone-de-beauvoir-en-una-proxima-pelicula-1592850.

no compartilhamento de informações. Se tivermos prudência, devemos ter ao menos uma posição cética com relação à circulação de informações presente no ambiente epistêmico em que nos encontramos nas redes. Porém, como veremos a seguir, o problema é mais profundo.

A tese da fonte envenenada generalizada

Observadas certas distinções entre as *fake news* e *deepfakes*, a partir daqui começaremos a dialogar com uma alarmada visão acerca do perigo iminente pelo qual nossos ambientes epistêmicos e seus agentes estão correndo: o solapamento da confiabilidade epistêmica que atribuímos às mídias sintéticas. Autores intitulam este risco de várias formas, sendo o termo usado neste artigo o mesmo de *Deepfakes and the epistemic apocalypse* – artigo este que se fará bem presente em nossa discussão: o de apocalipse epistêmico. Em (Habgood-Coote, 2022, p. 102) são elencadas todas estas denominações, bem como três afirmações que fazem parte desta narrativa que envolve a desconfiança geral dentro de sistemas de informação nos quais circulam tais mídias:

H1. *Deepfakes terão efeitos terríveis em nossas práticas sócio-epistêmicas*;
H2. *Deepfakes são historicamente sem precedentes*;
H3. *As soluções para as deepfakes são tecnológicas*.

Habgood-Coote considera H1 como uma visão plausível, porém incorreta, trazendo como exemplo elementos da epistemologia de instrumentos, bem como usa tal exemplo para contrapor H2; e ao final, argumenta contra H3 por caracterizar que há uma valorização indevida dos aspectos tecnológicos do problema em detrimento dos seus aspectos sociais. Neste texto, discutiremos sobre estas afirmações buscando concordar plenamente com H1, concordar em parte com H2 e discordar em parte de H3. Todavia, seria bom atestar antes o quê estaria em jogo.

Quando falamos sobre a ameaça das *deepfakes* em termos de um apocalipse epistêmico, temos exemplos na literatura que explicitam na verdade as consequências deste problema. No artigo de Fallis (2020), temos que os processos de formação de crenças tornam-se insuficientemente confiáveis, afirmando que "[...] quando vídeos falsos são difundidos, pessoas ficam menos propensas a acreditar no que está sendo retratado em um vídeo realmente ocorreu" (Fallis, 2020, p. 625, tradução nossa). Já nos termos de (Rini, 2020, p. 11-13), *deepfakes* geram crises de salvaguarda. Para ela, a concepção da gravação como fonte de conhecimento que assegura a confiabilidade de testemunhos é posta em xeque, fazendo com que ela deixe de ser uma fonte evidencial mais básica e passe a ser semelhante à própria fonte testemunhal à qual garante-se salvaguarda epistêmica.

Como Habgood-Coote destaca, há uma diferença quantitativa no primeiro caso acima – e no segundo, qualitativa – devido às *deepfakes*. Por último, Floridi afirma que "as tecnologias digitais parecem minar a nossa confiança na natureza original, genuína e autêntica daquilo que vemos e ouvimos" (Floridi, 2018, p. 320, tradução nossa). Porém, há dois aspectos – um explícito e outro, implícito – que se fazem presentes nestas análises: o primeiro trata sobre a confiabilidade do ambiente epistêmico em que estas gravações circulam; o segundo, sobre a confiabilidade dos agentes e grupos epistêmicos integrantes destes ambientes.

Comecemos aqui com um exemplo fictício baseado em situações que já ocorrem para ilustrarmos o problema: a utilização de informações acerca de uma pessoa pública para fazer um *deepfake* com conteúdo pornográfico e uso de drogas. Supondo que se faça um vídeo propositadamente com aparência de vídeo caseiro – com, por exemplo, um político realizando ações que prejudicam seu desempenho eleitoral e sua imagem. Vamos supor que a plataforma na qual o vídeo foi exposto é uma rede social permissiva a este tipo de material e proporciona uma

facilidade de circulação do mesmo[16]. O político pode vir a público e desmentir tal calúnia, porém outro vídeo é publicado contendo um pedido de desculpas e o mesmo político dizendo que nunca fez aquilo e que fazia questão de esclarecer isso a suas eleitoras e eleitores. Porém, este vídeo também é *deepfake*, e logo em seguida outro vídeo vem à tona, com uma melhor qualidade e menos características de ter sido forjado, com o político sendo filmado secretamente confessando para algumas pessoas em uma conversa informal que fez o que estava no vídeo e ninguém tinha nada a ver com isso. Este último vídeo, claro, sendo também produto de aprendizado de máquina e intencionalmente lançado para confundir as pessoas e deixar em suspenso a possibilidade do ocorrido e a sugestão que os vídeos que sugerem o mesmo seriam mais legítimos. O que fazer?

Acrescentemos uma segunda parte, completando o quadro. Há um conjunto de pessoas que têm o potencial de defendê-lo nas redes, entre elas algumas socialmente influentes. Em pouco tempo, é "vazado" o vídeo de uma destas pessoas no mesmo local que o político, realizando atos semelhantes, assim como há uma enxurrada de comentários e críticas pessoais aos agentes epistêmicos que buscam intervir, seja de maneira inquisitiva ou apenas cética, a veracidade dos conteúdos divulgados. Dadas as proporções sociais de cada um, todos teriam o potencial de ser atacados e expostos através de *fake news* reforçadas por *deepfakes*. Mesmo utilizando tecnologias de rede gerativas adversariais para identificar paulatinamente a desinformação contida em cada uma destas etapas, o processo de desconfiança tanto nas informações que circulam nesta rede quanto de seus agentes epistêmicos integrantes já é claro, e os instrumentos regulatórios de práticas testimoniais estabelecidos como adequados (como evidências de áudio e vídeo) não são dignas de credibilidade[17]. Entramos em um tipo de crise epistêmica que envenena o ambiente virtual como um todo e transborda para o cotidiano material, pois também compromete a confiabilidade da obtenção de evidências que o aprendizado de máquina possa simular.

Além do ambiente epistêmico ser envenenado pela propagação de *deepfakes*, seus agentes epistêmicos também não ficam incólumes. Pois, como vimos acima, a descredibilização de qualquer agente epistêmico, seja ele a motivação da *deepfake* ou apenas alguém que não esteja alinhado com as intenções escusas dos seus idealizadores, pode ser promovida e atualizada sem grandes esforços[18], obrigando desde o seu afastamento público até a cooptação da atenção e demanda de tempo deste agente para justificar-se e explicar-se, ocasionando o que vem a ser denominado de *dano ilocucionário* (Schiller, 2021) e (Rini e Cohen, 2021). Aqui, destacamos que, posto o devido potencial danoso das *deepfakes*, pode se instalar uma cadeia de desconfiança entre os agentes epistêmicos envolvidos nesta propagação de desinformação. Haveria uma distorção de quem é passível de confiança e quem não é pelo próprio ambiente, tornando-o insustentável para interação e troca de informações adequadas. Assim, mesmo que muitos sejam confiáveis, nunca ficaria claro para estes agentes como discriminar informações devido a falta de credibilidade do ambiente pernicioso, permissivo à falsidades e enganações.

Desta forma, podemos denominar este fenômeno de *tese da fonte envenenada generalizada*, inspirado pela sua semelhança com a falácia do poço envenenado, um subgrupo da falácia *ad hominem*. Há nesta tese um duplo aspecto: as *deepfakes* em potencial tornam qualquer ambiente epistemicamente pernicioso (o poço envenenado, de fato). Todavia, dada sua incidência, os

16As mídias sociais tornaram-se veículos de propagação de desinformação e um sítio onde formam-se comunidades com fenômenos epistêmicos bem peculiares relacionados a este nicho com grupos coesos que promovem compartilhamento de notícias, comentários e reforços de determinados comportamentos perniciosos. Para uma discussão maior sobre o tema ver Medeiros (2023).

17Aqui podemos notar que a crise da salvaguarda epistêmica destacada em (Rini, 2020) vem como consequência do potencial envenenamento do ambiente epistêmico. Daí a necessidade de instrumentos que garantam a confiabilidade deste ambiente.

18Se estivermos falando de um ambiente epistemicamente viciado que propicia a existência de bolhas epistêmicas ou câmaras de eco, este quadro piora devido ao fenômeno de corroboração retroalimentativa (*bootstrapping corroboration*), onde os agentes da bolha reforçam a confiança nas crenças um do outro por considerarem os integrantes do próprio grupo como fontes testemunhais confiáveis. Para definições e aprofundamentos ver (Nguyen, 2020) e (Boyd, 2018).

agentes epistêmicos que formam a crença falsa podem direcionar sua desconfiança aos agentes epistêmicos que se encontram prejudicados pelas enganações ou até a quem está incluso nesses ambientes e faça oposição a tal crença (poço envenenado falacioso). Para reforçarmos o quão este problema pode ser profundo, há certos fatores a serem considerados que apenas reforçam a ideia de que *deepfakes* podem ser uma forte ameaça à convivência social estabelecida pelas redes por potencializar problemas epistêmicos já identificados na literatura. Como exemplos, temos:

i) agentes com uma maior exposição a informações falsas tendem a acreditar nelas (Pillai e Fazio, 2021);

ii) a exposição de um agente epistêmico a imagens que possam lhe emocionar, ou que não trazem qualquer evidência para crer em alguma alegação, mas a sugerem emotivamente ou indicativamente através das imagens, enviesam sua formação de crença – o primeiro caso chama-se *efeito de amplificação da crença emocionalmente induzida* (Vlasceanu, Goebel e Coanu, 2020). No segundo, recorrendo a um neologismo, a imagem traria algo como veracidez (*truthiness*) a alegação, como exposto em (Newman e Zhang, 2020)[19];

iii) ambientes de mídias sociais que são enganosos, tornando a cooperação cada vez menos viável devido a mentiras que não visam o bem social, teoricamente tendem a se fragmentar (Iñiguez et al., 2014), bem como há correlações entre o uso de mídias sociais e a crença em más-informações ou informações consideradas conspiratórias (Enders et al., 2021), e;

iv) mesmo que o agente epistêmico que deteve a crença falsa com base em alguma *deepfake* tenha acesso a fontes de conhecimento que ele considere legítimas, há o que se caracteriza como ecos de crença (*belief echoes*), observando-se atitudes do agente que baseia-se em tal crença falsa, havendo uma perseverança de crença (*belief perseverance*) (Thorson, 2015).

Estas pesquisas acima indicam que a utilização tanto de *fake news* quanto *deepfakes* se complementam na propagação de desinformação e formação de crenças epistemicamente perniciosas. Esta conexão corrobora nossa afirmação que a última pode funcionar como fonte de justificação da primeira em certos casos. Portanto, temos que nestes ambientes epistêmicos, potencialmente, toda fonte pode ser (ou já estar) envenenada. Desta forma não teríamos como ter confiança nas informações nem nos agentes dentro de mídias sociais. Mantêm-se a pergunta: o que fazer? A resposta não é tão fácil, e nem completa. Ela torna-se dependente de como enxergamos o problema.

Entendendo soluções e críticas a um possível apocalipse epistêmico

Inicialmente, para acrescentar junto das já existentes medidas contra as *deepfakes*, poderíamos trazer as mesmas soluções e prevenções que estão sendo implementadas para evitar a propagação de *fake news*. Ou seja, adaptarmos tais medidas ao combate da indevida utilização da mídia sintética, bem como para o desmantelamento de ambientes com grupos epistemicamente perniciosos (Boyd, 2018) por onde tais notícias circulam. Como exemplos da literatura de *deepfake* temos a identificação e seu combate também por aprendizado de máquina e outras tecnologias *anti-deepfake*, legislação e regulação, conscientização educacional e por políticas públicas e treinamento de especialistas e dos usuários, como destacado em (Westerlund, 2019), (Mirsky e Lee, 2020) e (Farid, 2022). Na literatura acerca de *fake news*, podemos aproveitar a análise de rede e checagem de fatos com anexação de avisos e correção factual (Vosougui, Roy e Aral, 2018), (Pennycook e Rand, 2021) e (Achimescu e Chachev, 2021), bem como utilização de aprendizado de máquina junto com o princípio de "sabedoria das multidões" para capturar avaliações de confiabilidade do público leigo (Allen et al., 2020).

19 Veracidez seria a característica de uma mídia (seja ela sintética ou não) passar, apenas na exposição do agente à mesma, uma impressão ou sensação de que o que está sendo mostrado por meio desta mídia é verdadeiro.

Porém, algumas destas soluções dependem de avanços tecnológicos para identificação e detecção que podem, a qualquer momento, encontrar redes generativas adversariais superadoras geradoras de *deepfakes* que passariam por seu escrutínio. Esta disputa avançaria tal tecnologia sem termos qualquer garantia que há aspectos fulcrais que fizessem os programas sempre diferenciarem uma mídia digital gerada convencionalmente da mídia sintética. Outro ponto é, em certas soluções, a demanda pela participação voluntária dos agentes epistêmicos envolvidos nas redes. Pois justamente o problema do poço envenenado generalizado impede que haja confiança no próprio processo de produzir informação de qualidade em redes já comprometidas pelo fluxo de desinformação, descredibilidade entre seus agentes e o que podemos denominar de tese do viés cético, ou viés de valoração carregada (*value-laden bias*), onde:

> [...] valores podem ser considerados politicamente enviesados por pessoas que integram ambientes perniciosos – como as bolhas [epistêmicas] e câmaras [de eco] já citadas – sendo justamente os agentes que se encontram nesses ambientes parte relevante do público – alvo dessas medidas. [...] agentes epistêmicos que detêm um sistema de crenças baseado na negação e distanciamento de certos valores que embasam a implementação de certos métodos ou políticas dificilmente irão aceitar esses métodos ou políticas como legítimos. (Medeiros, 2023, pp. 511-12, chaves nossas)

Como exemplo, podemos destacar soluções que são direcionadas a criação de ambientes virtuais de *fact-checking*, seja considerando o juízo do que seria *fake news* ou *deepfake* verificado e legitimado por um corpo jornalístico tradicional ou por algoritmos que consideram o princípio de sabedoria das multidões – como já destacado na página acima. Temos aqui três situações:

I) Na primeira consideração acima, se temos bolhas epistêmicas ou câmaras de eco que rechaçam os valores que o jornalismo tradicional considera para a realização de um bom jornalismo, sua iniciativa epistemicamente não surtirá efeito nestes círculos;

II) Na segunda consideração, a aplicação desta solução pode considerar certo nicho populacional com valores que são negados por certos grupos, incidindo no mesmo problema de (I), ou;

III) Ainda que o algoritmo seja inserido de maneira indiscriminada, pode não haver diferenciação dos valores considerados para um bom julgamento dentro do princípio de sabedoria das multidões (coerência e fidedignidade, por exemplo) de outros valores característicos do fenômeno da loucura das massas (parcialidade, expectativas superdimensionadas), tornando a análise sobre desinformação valorativamente carregada justamente a considerar verdadeiro ou legítimo o que é falso.

Assim, somando-se a estas as dificuldades na listagem da seção anterior, os problemas incontornáveis ao menos nesta perspectiva demonstraram que devemos genuinamente nos preocupar diante da ameaça das *deepfakes* e de uma séria crise em nossas práticas sócio-epistêmicas que elas podem desencadear, corroborando com H1. Reforço aqui que a questão discutida neste artigo não é de crise moral – mesmo que tal consequência possa estar intrínseca a este tópico[20] – mas propriamente epistêmica. Leva-se à crise outras áreas dependentes dos atuais sistemas de informação que necessitam de informações autênticas, legítimas[21], que sirvam de fonte de justificação confiável[22].

Contudo, há quem defenda que o problema das *deepfakes* não é tecnológico, mas

20Em de Ruiter (2021) há uma síntese dos problemas éticos envolvidos, tendo por argumento que as *deepfakes* não são intrinsecamente moralmente condenáveis, mas são moralmente suspeitas.

21Nossa breve análise anterior sobre a característica atributiva da alcunha de 'falsidade' às *deepfakes* tem justamente esta intenção de destacar que há uma noção de autenticidade que deve ser melhor observada dentro dos sistemas de informação em distinção às *fake news*.

22Para termos noção da abrangência e amplo impacto social de seu mau-uso, ver (Citron e Chesney, 2019) e (Farid, 2022).

intrinsecamente social. No artigo (Habgood-Coote, 2023) argumenta-se contra uma noção de um apocalipse epistêmico gerado pela ameaça das *deepfakes* minarem qualquer garantia de informação confiável acerca da realidade que possamos obter pelas redes. Ele foca sua crítica nas propostas de solução focadas no desenvolvimento de tecnologia, bem como na afirmação de que não haveria precedentes deste tipo na história. Assim, aponta pelos mesmos recursos da literatura abordados por Fallis e Rini – i.e., através da epistemologia social e da história de manipulação das gravações – como este fenômeno de falsificação de mídia, ou contrário do que é advogado, não é sem precedentes, além de termos tais crises epistêmicas solucionadas dentro das nossas próprias práticas sociais:

> Ao invés de vermos os deepfakes como um problema sobre uma forma de tecnologia exclusivamente perigosa, indicativa da queda de um estado puro de graça epistêmico, deveríamos ver a existência de *deepfakes* como um sintoma de problemas de longa data em torno da gestão das normas de produção e divulgação de gravações. (Habgood-Coote, 2023 p. 102, tradução nossa)

Ele coloca que o tecnochauvinismo[23] e narrativas de pós-verdade[24] são posturas inadequadas para lidar com este problema, existindo soluções mais práticas e economicamente viáveis para seu combate do que mais investimento em tecnologia e nutrir um discurso catastrofista contemporâneo de que há um abandono dos valores epistêmicos iluministas. Algumas delas são o banimento de fóruns, privar o livre acesso a produção de mídia sintética, remover financiamento para produção de pornografia com *deepfake*, entre outros.

Ora, tendemos a concordar com alguns pontos dos argumentos fornecidos por Habgood-Coote, como a falha no estabelecimento de normas adequadas para controle da criação e propagação de *deepfakes* e colocar o encargo da solução no desenvolvimento tecnológico (este último apontado acima). Porém, aqui há uma discordância em parte ao artigo de Habgood-Coote da alegação que este é um fenômeno com precedentes (discordância de H2). Tanto no que concerne os exemplos utilizados dentro da epistemologia de instrumentos, quanto na história de manipulação de gravações conduzidos para defesa deste argumento, analisando o fenômeno das mídias sintéticas à luz da manipulação de fotografias. Esta discordância será explicitada abaixo.

Atualmente, vivemos em um período em que as atividades econômicas e a influência política na Internet passaram a ter grande valor. Em 2021, um estudo mostrou que apenas nos EUA a economia digital é responsável por 12% do seu PIB, subindo em um período de quatro anos sete vezes mais do que sua contraparte física (Deighton e Kornfeld, 2021). O Banco Mundial estimou que a economia digital corresponde a 15% do PIB global, com a expectativa de dobrar este valor até 2030, sendo estes dados fatores preponderantes na avaliação econômica da importância de se manter a confiança digital dos sistemas de informação[25].

Devemos observar que a informação não tem apenas dimensão epistêmica, mas também é uma *commodity* com dimensões de valor social, econômico e político agregados que são extremamente influentes e onipresentes em nosso panorama contemporâneo. E o que, para nós, isso significa? Que não devemos comparar *deepfakes* apenas com manipulação de fotografias –

23"[...] é uma atitude intelectual, envolvendo três tendências: reembalar os problemas sociais como problemas tecnológicos (tecno-solucionismo), acreditar que os sistemas tecnológicos podem realizar tarefas complexas (tecno-otimismo) e ignorar ou subestimar a importância dos projetistas, operadores e mantenedores de sistemas tecnológicos (tecno-fixação)." (Habgood-Coote, 2023, p.103, tradução nossa).

24Não entraremos nesta crítica de Habgood-Coote, pois fugiria do escopo deste artigo. Porém, ela nos parece a princípio infundada devido a pós-verdade não ser apenas um termo guarda-chuva, mas um fenômeno social com aspectos observáveis ao menos desde o desenvolvimento contemporâneo da Propaganda (que se reformulou através de Edward Bernays como Relações Públicas). Mais detalhes em (Medeiros, 2023).

25Notícia disponível em Digital trust: How to unleash the trillion-dollar opportunity for our global economy.

estaríamos fazendo um recorte impreciso e seletivo – mas também com a falsificação de artefatos e identidade com valor agregado – como pinturas, moedas, títulos de terra e de nobreza, peças originais de reposição e até a figura histórica do impostor. Vamos nos ater aqui a falsificação de moeda corrente.

Breve esboço de uma história da falsificação

A existência da falsificação de objetos ou da própria *persona* que poderiam trazer ao falsificador vantagens sociais têm registros desde antes da fabricação de dinheiro, quando já entre 3300-2000 a.c usavam-se conchas de búzios como moeda corrente na Ásia e na África – justamente por serem difíceis de se simular – e forjavam-se falsificações usando marfim, osso, concha e pedra (Peng e Zhu, 1995). As primeiras notas de valor (*jiaozi*) apareceram no fim do século X na China, sendo cedido a dezesseis casas mercantis o direito de impressão. Mesmo com cores e designs distintos, selos e estampas, foram fabricadas tantas falsificações que houve inflação da moeda em menos de vinte anos, obrigando o governo da dinastia Song restringir e centralizar a produção (Von Glahn, 2006). A raspagem e o recorte de moedas para derretimento e cunhagem de outras foi um procedimento feito por falsificadores na Inglaterra do século XVII, fazendo com que o peso das moedas caísse pela metade do valor estipulado e uma a cada dez fosse falsa, havendo um recolhimento das moedas e nova cunhagem (Levenson, 2010).

Há passagens históricas que mostram a capacidade que a falsificação de coisas com valor agregado tem de distorcer e minar toda a confiança de um sistema de forma mais contundente, com sérias consequências. Como exemplo, devido a situação econômica de desigualdade e pobreza da região, no século XVII houve uma produção razoável de moedas falsificadas em conventos castelhanos da Córdoba no reinado de Filipe IV. A confiabilidade entre as trocas mercantis e financeiras foram abaladas de tal forma – por negociantes, mercadores e produtores de todas as escalas não saberem distinguir o numerário autêntico do falso – que como consequência houve completa paralisação mercantil de várias cidades e a medida da coroa de retirar todo a moeda corrente de circulação (Fernández, 1997).

Podemos citar aqui o caso das notas bancárias portuguesas de 1925, relatado em detalhes em (Wigan, 2004). Este é um caso bem icônico e conhecido: Artur Alves Reis forjou contratos de impressão de cédulas e convenceu a *Waterloo and Sons* – um impressor com sede londrina que detinha placas de impressão oficiais do banco de Portugal – que seus arranjos foram autorizados pelo banco central. Utilizando a má fama da governabilidade corrupta portuguesa na época pós-primeira guerra, confiaram na legitimidade de seus contratos e imprimiram uma quantia equivalente a 0,88% do PIB de Portugal na época. Assim, Reis montou um banco de comércio, realizando vários investimentos e através da compra de ações do banco de Portugal, quase obtendo seu controle. O esquema foi descoberto e Alves preso, mas o estrago inflacionário já estava feito, com consequências na confiabilidade nos escudos portugueses, culminando em uma crise, deposição do presidente e ascensão do fascismo português.

Como podemos ver no exemplo anterior, as consequências políticas também estão fortemente presentes neste tipo de falsificação. Desta forma, observamos que historicamente os problemas gerais causados pela falsificação de moeda corrente são de inflação e crise de confiabilidade, gerando desestabilidade econômica, mercantil e social. Devido a estes fenômenos, engenhosamente a falsificação passou a ser utilizada como arma (*weaponization*) política por Estados e governos: há relatos de uso deste artifício pelos ingleses na Guerra de Independência Americana para desestabilizar as colônias. Na Alemanha pós-primeira guerra para enfraquecer belgas e franceses usando francos falsificados para pagar dívidas de guerra, e; os próprios americanos através da CIA, aproveitando a experiência da tentativa de desestabilização do Japão

na segunda guerra, usaram as mesmas técnicas contra países que pelas suas análises estariam sofrendo influência comunista, como Vietnã, Laos, Coreia do Norte e Cuba, entre outros casos relatados em (Cooley, 2008). Nestes fatos há similaridades com as consequências do uso que pode ser (ou já são) feitos das *deepfakes*, não à toa: a dicotomia em um sistema entre artefatos autênticos e falsos em ambientes que dependem justamente da autenticidade destes artefatos para manter sua credibilidade geram crises e desestabilidade – seja em sistemas econômicos, seja em sistemas de informação, levando-o a disfuncionalidade e até ao colapso.

Fornecemos aqui uma pequena amostra de como percebemos o problema das *deepfakes* e de onde devemos partir para pensar em soluções: dentro da dinâmica social da falsificação de objetos com determinado valor. Trazendo novamente aqui as três afirmações de Habgood-Coote, concordamos em parte com H2 pois, mesmo que haja uma semelhança dentro de casos históricos de falsificação, podemos ver que este fenômeno, pelo seu alcance e caráter informacional, é sem precedentes. Para deixar claro, não estamos com este argumento deixando de lado as análises em epistemologia social e de instrumentos[26] – pois consideramos que as soluções encontram-se mesmo dentro de nossas práticas sociais – mas procurando localizar qual o cerne da questão a ser tematizada quando estamos a elucidar os perigos do uso desta tecnologia. Ao final, iremos apontar horizontes de investigação para discussões sobre o futuro das mídias sintéticas.

Considerações finais

Vimos neste artigo as características nocivas, distinções das fake news e possíveis consequências da propagação de desinformação em formato de *deepfake*, minando a credibilidade dos sistemas de informação por onde as mídias sintéticas têm potencial de circulação, gerando uma crise epistêmica de confiabilidade que denominei de tese da fonte envenenada generalizada. Discutimos as soluções oferecidas tanto às *deepfakes* quanto as que, adaptando seu uso às *fake news*, haveria possibilidade de se aplicar a mídia sintética. Desta forma, voltamo-nos a posição mais cética de Habgood-Coote acerca desta crise, alegando por meio da história de manipulação de fotografias que este tipo de fenômeno social de forjar gravações não é recente. Contudo, buscamos demonstrar que, mesmo concordando com a afirmação que a regulação de nossas práticas sociais seja a forma mais adequada de entender a solução do problema das *deepfakes*, discordamos em parte da construção argumentativa e análise de Habgood-Coote, onde indicamos que uma investigação direcionada a história das falsificações seja mais frutífera para compreendermos tal fenômeno social e possíveis medidas para seu combate, expondo e descrevendo alguns casos históricos de falsificação de moeda corrente.

Deste modo, nos vem a questão: ao final, quais medidas consideramos aqui como mais adequadas? Ao longo da história das falsificações, podemos observar duas constantes: punições rígidas e métodos para garantir, identificar e verificar a autenticidade do valor avaliado, seja ele em objetos de arte, moeda, ou documentos e chancelas que referenciam a alguma pessoa considerada de renome. Pegando inicialmente a moeda corrente, há diferenciais físicos e perceptuais como constituição material, tamanho, peso – e outros mais atuais, como marca d'água – bem como número de série, assinaturas e símbolos que, em conjunto, trazem dificuldade à falsificação e maior confiabilidade. Com relação a documentos, na Roma Antiga já existiam uma gama de autenticadores como selos, testemunhas, escrita à mão e símbolos, com retenção dos registros em cartórios públicos – este último visto como influência da província egípcia, que

26Estudos como no desenvolvimento acerca da salvaguarda epistêmica em (Rini, 2020) e o discutido acima, de (Habgood-Coote, 2023) são valiosos pelo entendimento que trazem em incorporar a historicidade social de nossos artefatos e práticas associadas ao seu uso como epistemicamente relevantes no desenvolvimento, preservação e aquisição do conhecimento. Acredito que tais abordagens são herdeiras do que o filósofo canadense Ian Hacking denominava de metaepistemologia em (Hacking, 2009).

mantinham registros de seus faraós, dos persas e da dinastia ptolomaica – existindo um antepassado de tabelião, denominado notário (*notary*), que dava garantia de legalidade aos autenticadores (Haighton, 2010). Na Idade Média, tais procedimentos se mantiveram[27], acrescido de outros recursos como símbolos heráldicos e o que chamam de *quirografia*[28].

Assim, quando relatamos historicamente falsificações, também há uma demanda de compreensão acerca da historicidade dos procedimentos de autenticação relativo ao que está sendo falsificado e seu valor epistêmico[29]. Para nós, o desenvolvimento destas linhas de estudo voltados à falsificação de outros tipos de objetos (como bens e peças de reposição, identidade pessoal) podem auxiliar a pensarmos melhores formas de lidar com as *deepfakes*. Até porque boa parte das nossas atuais gravações produzidas por dispositivos e rodadas em aplicativos, seja em pequena ou grande medida, são alteradas a nível algorítmico, sem qualquer intervenção nem edição intencional por parte do sujeito que realiza a gravação. Portanto, precisamos de pesquisas para avaliar similarmente estes graus de autenticação e legitimidade.

Floridi (2018) destaca a distinção entre o que seria falso (*fake)* e uma réplica apelando a noção de *ectype* em contraste com arquétipo (*archetype*) de um conjunto de artefatos. Extraindo as características mais salientes que definem um tipo de artefato (os quadros de Rembrandt) pode-se criar uma peça que pode ser considerada autêntica, mas não original com relação a sua fonte prototípica – i.e., as características salientes dos quadros de Rembrandt. Portanto, há autenticidade nesta peça que é uma cópia de valor próximo ao de uma peça original, sem ambiguidades nem más-intenções em sua criação ou reprodutibilidade, diferentemente de uma *deepfake* (Floridi, 2018 pp. 319-20). Após tais distinções, Floridi destaca o papel que tecnologias como a de *blockchain* podem ter no auxílio à verificação de autenticidade e originalidade de artefatos.

Aqui, observamos que sua abordagem se enquadra dentro das críticas às soluções pela tecnologia de Habgood-Coote (discordância do ponto H3) e concordamos com tal crítica. Devemos seguir com temperança e ter a clareza de que a produção e aplicação de tecnologias na resolução de problemas é, em si, uma prática social que deve ser exercida de forma virtuosa como um meio – em conjunto com outras práticas. Mitigamos assim a propagação de desinformação e mantemos a confiabilidade de nossos ambientes epistêmicos, garantindo efetividade e evitando outros malefícios sociais. Entretanto, há uma discordância nossa distinta da de Habgood-Coote. As *deepfakes* são artifícios que, devido a grande quantidade de aplicações, vieram para ficar. Traçar investigações históricas[30] e pensar a partir dos mesmos procedimentos de autenticação contemporâneos viáveis para serem implementados nos sistemas de informação indica ser um caminho para a mitigação de sua ameaça epistêmica.

27No caso de registros públicos, havia uma dissonância, pois na Inglaterra e País de Gales a lei consuetudinária autorizava qualquer pessoa autenticar seu próprio documento apenas selando-o. Portanto a Igreja, mantendo a tradição legal romana, tinha seus próprios notários para autenticação documental do clero.

28Quirografia consistia em confeccionar os documentos em duplas em uma folha e cortá-los em certos padrões, dividindo as cópias entre as partes interessadas. Juntando as cópias e os padrões sendo compatíveis era a comprovação da legitimidade do documento. Para maior segurança, entre os padrões poderia ter algo escrito anteriormente, para quando as peças se juntam, os pedaços de palavras também se completam. Poderíamos comparar este procedimento a uma criptografia analógica. Para mais detalhes, ver (Bedos-Rezak, 2010).

29O conjunto de ensaios intitulado *Beyond Disinformation* em (Rodhy, Miller e Weber, 2021) volta-se às fake news, tratando-se de um início de abordagem acerca desta temática relacionando pesquisas sobre autenticação e propagação de desinformação.

30Além desta análise, há outras similarmente úteis para entendermos fatores relevantes, como a tensão econômica entre as falsificações e a implementação de autenticações por meio de convenções que podem ter sua qualidade bem estabelecida através de *expertise*. Um início pode ser o estudo de (Bessy e Chateauraynaud, 2019).

Referências

ACHIMESCU, V.; CHACHEV, P. D.. Raising the flag: monitoring using perceiving disinformation on Reddit. *Information* 12(1): 4, 2021.
ALLEN, J.; ARECHAR, A. A.; PENNYCOOK, G.; RAND, D. Scaling up fact - checking using the wisdom of crowds. *PsyArXiv Published online*, 2021.
ANDERAU, G. Fake News and epistemic flooding. *Synthese* 202: 106, 2023.
AUDI, R. The sources of knowledge. In: MOSER, Paul K. (ed.) *The Oxford Handbook of Epistemology*. Oxford University Press, 2002.
BEDOS-REZAK, B. M. Cutting edge: The economy of mediality in twelfth-century chirographic writing. *Das Mittelalter* 15(2): pp. 134-61, 2010.
BERNECKER, S.; FLOWERREE, A. K.; GRUNDMANN, T. (eds.). *The Epistemology of Fake News*. New York, NY: Oxford University Press, 2021.
BESSY, C.; CHATEAUREYNAUD, F. The dynamics of authentication and counterfeits in markets. *Historical social research/Historische Sozialforschung* 44, No. 1(167): Markets, organizations, and law – Perspectives of convention theory on economic practices and structures: 136-59, 2019.
BOYD, K. Epistemically pernicious groups and the groupstrapping problem. *Social Epistemology* 33(1): pp. 61-73, 2018.
BUSCH, E.; WARE, J. The weaponization of deepfakes: digital deception by the far-right. *ICCT Policy Brief*, 2023.
CITRON, D. K.; CHESNEY, R. "Deep fakes: A looming challenge for privacy, democracy, and national security". *California Law Review* 107: pp. 1753–819, 2019.
COOLEY, J. K. *Currency wars: how forged money is the new weapon of mass destruction*. SkyHorse Publisher: New York, 2008.
DEIGHTON, J.; KORNFELD, N. The economic impact of the market-making Internet. *Interactive Advertising Bureau* (IAB): New York, 2021. Disponível em: https://www.iab.com/wp-content/uploads/2021/10/IAB_Economic_Impact_of_the_Market-Making_Internet_Study_2021-10.pdf.
ENDERS, A. M.; USCINSKI, J. E.; SEELING, M. I.; KLOFSTAD, C. A.; WUCHTY, S.; FUNCHION, J. R.; MURTHI, M. N.; PREMARATNE, K.; STOLER, J. The relationship between social media use and beliefs in conspiracy theories and misinformation. *Political Behavior* 45: pp. 781–804, 2021.
EUROPOL. Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab. *Publications Office of saperethe European Union*, Luxembourg, 2022.
FALLIS, D. The epistemic threat of deepfakes. *Philosophy and Technology* 34(4): pp. 623–43, 2020.
FARID, H. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety* 1(4): pp. 1-33, 2022.
FERNÁNDEZ, J. de Santiago. Falsificación de moneda en conventos cordobeses en 1661. *Hispania Sacra* 49(99): 233-50, 1997.
FLORIDI, L. *Philosophy of Information*. Oxford University Press: New York, 2011.
FLORIDI, L. Artificial intelligence, deepfakes and a future of ectypes. *Philosophy and Technology* 31: pp. 317-21, 2018.
FOGG, B. J. Stanford guidelines for web credibility. A research summary from the Stanford Persuasive Technology Lab. *Stanford University*, 2002.
FRANKFURT, H. G. *Sobre falar merda*. Tradução de Ricardo Gomes Quintana. Editora Intrínseca: Rio de Janeiro, 2005.
GEACH, P. Good and Evil. *Analysis* 17(2): pp. 33-42, 1956.
GOSSE, C.; BURKELL, J. Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication* 37(5): pp. 497-511, 2020.

HACKING, I. *Ontologia Histórica*. Traduzido por Leila Mendes. Editora Unisinos: Rio Grande do Sul, 2009.
HABGOOD-COOTE, J. Deepfakes and the Epistemic Apocalypse. *Synthese* 201(103): pp. 1–23, 2023.
HAIGHTON, A. Roman methods of authentication in the first two centuries A.D.. *Journal of the Society of Archivists* 31(1): pp. 29-49, 2010.
HAO, K. The biggest threat of deepfakes isn't the deepfakes themselves. *MIT Technology Review*, 2019. Disponível em
https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/.
HARRIS, D. Deepfakes: false pornography is here and the law cannot protect you". 17 *Duke Law & Technology Review*: pp. 99-127, 2019.
HIMELEIN-WACHOWIAK, M.; GIORGI, S.; DEVOTO, A.; RAHMAN, M.; UNGAR, L.; SCHWARTZ, H. A.; EPSTEIN, D. H.; LEGGIO, L.; CURTIS, B. Bots and misinformation spread on social media: implications for COVID-19. *Journal of Medical Internet Research* 23(5): e26933, 2020.
IÑIGUEZ, G.; GOVEZENSKY, T.; DUNBAR, R.; KASKI, K.; BARRIO, R. A. Effects of deception in social networks. *Proceedings of the Royal Society B* 281: 20141195, 2014.
JASTER, R.; LANIUS, D. Speaking of fake news: definitions and dimensions. In: BERNECKER, S.; FLOWERREE, A. K.; GRUNDMANN, T. (eds.). *The Epistemology of Fake News*. Oxford University Press: pp. 19-45, 2021.
LACKEY, J. It Takes Two to Tango: Beyond Reductionism and Non-Reductionism in the Epistemology of Testimony. In: LACKEY, Jennifer; SOSA, Ernest(eds.). *The Epistemology of Testimony*. Oxford: Oxford University Press: pp. 160−82, 2006.
LANKES, R. D. Credibility on the internet: shifting from authority to reliability. *Journal of Documentation* 64(5): pp. 667-86, 2008.
LEONARD, N. Epistemological problems of testimony. In: Edward .N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, 2021.
Disponível em https://plato.stanford.edu/archives/spr2023/entries/testimony-episprob/
LEVENSON,T. *Newton and the counterfeiter: The Unknown Detective Career of the World's Greatest Scientist*. Houghton Mifflin Harcourt: Boston, 2009.
MARTÍNEZ, V. C.; PADILLA-CASTILLO, G. Historia del fake audiovisual: deepfake y la mujer en un imaginario falsificado y perverso. *Historia y comunicación social* 24(2): 505-20, 2020
MEDEIROS, B. B. de. Pós-verdade e aspectos epistêmicos das mídias sociais: análises, críticas e proposições. *Sapere Aude* 14(28): pp. 492-522, 2023.
MIRSKY, Y.; LEE, W. The creation and detection of deepfakes: a survey. *ACM Computing Surveys*, 54(1): pp. 1–41, 2020.
NEGI, S.; JAYACHANDRAN, M.; UPADHAYAY, S. Deep fake: an understanding of fake images and videos. *JSRCSEIT* 7(3): pp. 183-89, 2021.
NEWMAN, E. J.; ZHANG, L. Truthiness: how nonprobative photos shape belief. In: GREIFENEDER, R.; JAFFÉ, M.; NEWMAN, E. J.; SCHWARZ, N. (eds.). *The psychology of fake news: Accepting, sharing, and correcting misinformation*. London, UK: Routledge, 2020.
PENG, K.; ZHU, Y.. New Research on the Origin of Cowries in Ancient China. *Sino-Platonic Papers* **68**: 1–26, 1995.
PENNYCOOK, G.; RAND, D. G. The psychology of fake news. *Trends in Cognitive Sciences* **25**(5): 388-402, 2021.
NGUYEN, C. T. Echo chambers and epistemic bubbles. *Episteme* 17(2): pp. 141-61, 2018.
PILLAI, R. M.; FAZIO, L. K. The effects of repeating false and misleading information on belief. *WIREs Cognition Science*: e1573, 2021.
RIEH, S. Y. Credibility and cognitive authority of information. *Encyclopedia of Library and Information Sciences* (3º ed.) 1(1): pp. 1337-344, 2010.
RINI, R. Deepfakes and the epistemic backstop. *Philosophers' Imprint* 20(24): pp. 1-16, 2020.

RINI, R.; COHEN, L. Deepfakes, deep harms. *Journal of Ethics and Social Philosophy*, 22(2): 141–161, 2022.

RODHY, J.; MILLER, M.; WEBER, P. (org.). *Beyond Desinformation: How do authenticity and trust function in digital spaces?* Media & Democracy Workshop: Social Science Research Concil, 2021. Disponível em https://items.ssrc.org/category/beyond-disinformation/

SCHILLER, H. I. Illocutionary harm. *Philosophy Studies* 178: pp. 1631–646, 2021.

SHAO, Chengcheng; CIAMPAGLIA, G. L.; VAROL, O.; YANG, Kai-Cheng; FLAMINNI, A.; MENCZER, F. The spread of low-credibility content by social bots. *Nat Commun* 9(1): 4787, 2018.

THORSON, E. A. Belief echoes: The persistent effects of corrected misinformation. *Political Communication* 33(3): 460-80, 2015.

VLASCEANU, M.; GOEBEL, J.; COMAN, A. The emotion-induced belief amplification effect. *42nd Annual Meeting of the Cognitive Science Society: Developing a Mind: Learning in Humans, Animals, and Machines*, CogSci 2020: pp. 417-22, 2020.

VON GLAHN, R. Re-examining the Authenticity of Song Paper Money Specimens. *Journal of Song-Yuan Studies* 36: pp. 79-106, 2006.

VOSOUGUI, S.; ROY, D.; ARAL, S. The spread of true and false news online. *Science* 359, issue 6380: 1146-151, 2018.

WEEDON J.; NULAND, W.; STAMOS, A. Information operations and Facebook. *Facebook*, 27 April, 2017.

WESTERLUND, M. The emergence of deepfake technology: a review. *Technology Innovation Management Review*, 9(11): pp. 39–52, 2019.

WIGAN, H. The effects of the 1925 Portuguese Bank Note Crisis. *Economic History Working Paper No. 82/04*: London School of Economics and Political Science, 2004.

Autor(a) para correspondência / Corresponding author: Bismarck Bório de Medeiros. bismarckborio@gmail.com