DIGITAL PEACEBUILDING

2024 Issue 1

Obfuscating Transparency: The Promises and Risks of Defusing Deepfake Disinformation through Blockchain

Annette Y. Lee

Executive Summary

Concerns are mounting over the threat of deepfake disinformation. Though the conversation is largely speculative so far, we have seen examples of deepfakes used for harmful or violent purposes. One solution, touted for its transparency and security, could be authenticating content through blockchain. However, there is still a long way to go before this technology is widespread and well-designed enough to be effective. Furthermore, proof-of-authenticity blockchain may backfire by marginalizing the most at-risk, perpetuating harmful power dynamics, and creating further confusion. I recommend that policymakers, scholars, and tech companies pay close attention to certain design aspects, and continuously research and reflect on potential for harm.

Introduction

Disinformation is evolving, and Americans are concerned: a 2019 Pew Research poll found that 91% of Americans believe that altered or fake videos and images create confusion.[1]

It is important to understand the risks deepfake disinformation presents to peace in order to combat it. To that end, blockchain technology that authenticates media may be a useful tool.

However, I argue that this approach comes with its own risks, which policymakers and tech companies alike should consider.

What are Deepfakes?

Deepfakes are realistic media created with machine-learning algorithms, encompassing videos, images, and audio. Several methods for creating deepfakes exist,[2] though many scholars focus on the use of generative adversarial networks (GANS) in which two algorithms are rapidly trained against each other.[3] Deepfakes, or at least that term, first surfaced in a 2017 Reddit post featuring face-swapped pornography of a female celebrity, and women continue to be disproportionately targeted by deepfakes.[4]

[1] Sara Atske, "Many Americans Say Made-Up News Is a Critical Problem That Needs To Be Fixed," *Pew Research Center's Journalism Project*, June 5, 2019, https://www.pewresearch.org/journalism/2019/06/05/manamericans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.

[2] Jia Wen Seow et al., "A Comprehensive Overview of Deepfake: Generation, Detection, Datasets, and Opportunities," *Neurocomputing* 513 (November 7, 2022): 351–71, https://doi.org/10.1016/j.neucom.2022.09.135.

[3] Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War," *Foreign Affairs* 98, no. 1 (December 11, 2018), https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war; Bryan C. Taylor, "Defending the State from Digital Deceit: The Reflexive Securitization of Deepfake," *Critical Studies in Media Communication* 38, no. 1 (2021): 1–17, https://doi.org/10.1080/15295036.2020.1833058; Kathryn A. Paradis, "Fighting Deep Fakes: The Inadequacy of Current Law for the Future War," *Naval Law Review* 68 (2022): 83–104, https://heinonline.org/HOL/P?h=hein.journals/naval68&i=89; Md Shohel Rana et al., "Deepfake Detection: A Systematic Literature Review," *IEEE Access* 10 (2022): 25494–513, https://doi.org/10.1109/ACCESS.2022.3154404.

[4] Henry Ajder et al., "The State of Deepfakes: Landscape, Threats, and Impact" (Deeptrace, September 2019), https://enough.org/objects/Deeptrace-the-State-of-Deepfakes-2019.pdf; Sophie Maddocks, "'A Deepfake Porn Plot Intended to Silence Me': Exploring Continuities between Pornographic and 'Political' Deep Fakes," *Porn Studies* 7, no. 4 (October 1, 2020): 415–23, https://doi.org/10.1080/23268743.2020.1757499; Sam Gregory, "Deepfakes, Misinformation and Disinformation and Authenticity Infrastructure Responses: Impacts on Frontline Witnessing, Distant Witnessing, and Civic Journalism," *Journalism* 23, no. 3 (March 1, 2022): 708–29, https://doi.

Annette Y. Lee. Brown University, A.B. International and Public Affairs '23 and MPA '24.
Contact: annetteylee@protonmail.com

Editor: Dawn Brancati

Deepfake Disinformation: Challenges to Peace

Current research on deepfakes' connection to violence is largely speculative. Indeed, recent deepfakes of Presidents Zelenskyy and Putin caused little material harm.[5] In the future, however, deepfake disinformation could be potent. Its realistic and evocative nature could make it difficult to debunk, widely spread, and widely believed.[6] As Danielle Citron and Robert Chesney write, "There is no doubt that deep fakes will play a role in future armed conflicts."[7]

World Economic Forum/Jakob Polacsek. 2020. A viewer is transformed into Michelle Obama at Pinscreen's Deep Fake Exhibition at the World Economic Forum Annual Meeting in Davos-Klosters, Switzerland. https://www.flickr.com/photos/worldeconomicforum/49425304693/in/photostream/
This image is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.0 Generic license, and has been cropped.

Deepfakes could lead to violence by:

- Depicting influential people doing something inflammatory, to incite violence or bolster support of violence.[8]
- Capitalizing on existing tensions or discrimination against a specific group.[9]
- Perpetuating conflict by legitimizing uprisings, falsifying orders, discrediting leaders, and dividing allies.[10]
- Being used by terrorist groups to incite violent reactions.[11]

Even the possibility that something is a deepfake could incite violence. In 2019, the allegation that a video of President Bongo of Gabon was a deepfake contributed to an attempted military coup.[12]

However, deepfakes are so new that their potential impact is still up in the air, influenced by several potential factors:

- *Technological advancement*: It is difficult for deepfake detection techniques to keep up with rapidly developing deepfake technology.[13] If detection technology struggles to keep up, this could magnify the harm of deepfake disinformation. Furthermore, as deepfake technology becomes more accessible, there are fewer barriers to using it for violent purposes.
- *Legislation and regulation*: Future regulations on deepfakes could curb their harmful use, but poorly crafted legislation could cause further damage by restricting or criminalizing free expression,[14] thus potentially facilitating further violence.
- *Education and training*: Increased opportunities for learning about deepfakes could raise audience awareness, mitigating the harmful impact of deepfake disinformation.[15]
- *Private sector policies*: Social media platforms' deepfake policies could curb potential harm.[16] Currently, platforms prioritize deepfakes affecting geopolitical conflict over deepfakes impact-

org/10.1177/14648849211060644.

[5] Britt Paris, "Seeing Through the Fog of War: Assessing Epistemic Burden Around Cheapfakes and Deepfakes of Geopolitical Crisis," in *Re-Thinking Mediations of Post-Truth Politics and Trust* (Routledge, 2023).

[6] Alisha Anand and Belen Bianco, "The 2021 Innovations Dialogue Conference Report: Deepfakes, Trust and International Security," *UNIDIR*, December 22, 2021, https://unidir.org/publication/the-2021-innovations-dialogue-conference-report/; Danielle Citron and Robert Chesney, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107, no. 6 (December 2019): 1753, https://doi.org/10.15779/Z38RV0D15J.

[7] Citron and Chesney, "Deep Fakes."

[8] Chesney and Citron, "Deepfakes and the New Disinformation War."

[9] Maria Pawelec, "Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions," *Digital Society* 1, no. 2 (September 8, 2022): 19, https://doi.org/10.1007/s44206-022-00010-6.

[10] Daniel L. Byman et al., "Deepfakes and International Conflict," *The Brookings Institution*, January 2023, https://www.brookings.edu/articles/deepfakes-and-international-conflict/.

[11] Arije Antinori, "Terrorism and Deepfake: From Hybrid-Warfare to Post-Truth Warfare in a Hybrid World," in *ECIAIR 2019 European Conference on the Impact of Artificial Intelligence and Robotics* (Academic Conferences and Publishing Limited, 2019).

[12] Ajder et al., "The State of Deepfakes: Landscape, Threats, and Impact."

[13] Rana et al., "Deepfake Detection."

[14] Tyrone Kirchengast, "Deepfakes and Image Manipulation: Criminalisation and Control," *Information & Communications Technology Law* 29, no. 3 (September 1, 2020): 308–23, https://doi.org/10.1080/13600834.2020.1794615.

[15] Mika Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review* 9, no. 11 (November 2019): 40–53, https://doi.org/10.22215/timreview/1282; Byman et al., "Deepfakes and International Conflict."

[16] Ibid.

ing individuals, thus failing to address the root problem of truth decay and disproportionately neglecting women.[17]

- Political polarization: Deepfakes benefit from “a heated political context where false narratives are easily spread and easily believed online.”[18] Decreased polarization could mitigate their potential violent impacts.

At present, it is still quite a leap from deepfake disinformation to actual kinetic violence.[19] Deepfakes could also indirectly contribute to violence by:

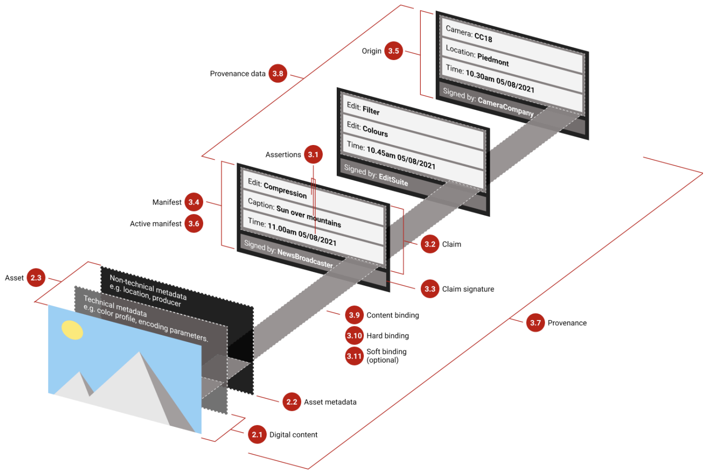
- Being used for sabotage, blackmail, or weakening diplomatic ties,[20] which could in turn help justify or incite violence.
- Contributing to lower trust and greater polarization, which can destabilize regimes, erode democracy,[21] and ultimately increase the likelihood of violence.
- Leading an increasingly skeptical audience to believe truths are falsehoods, in a phenomenon called the “liar’s dividend.”[22] Similarly, by casting doubt on anything that cannot be perfectly verified, deepfakes could harm vulnerable or prosecuted voices, which by their nature cannot disclose everything.[23]

How Blockchain Could Help or Hurt

Most strategies for fighting deepfake disinformation focus on detection, i.e., proving that media is fake. However, considering how rapidly detection must evolve in order to keep pace with deepfakes, proving that media is authentic may be more effective.[24]

Blockchain records data in a decentralized shared database, also known as a Distributed Ledger Technology (DLT). This data is stored in a fixed order in “blocks” that are functionally impossible to modify. In this way, blockchain allows information to be stored free of tampering. Many describe it as inherently transparent, equalizing, and secure. Blockchain could indicate proof of authenticity in multiple ways, which could be used in concern with one another:

- *Provenance and Origin*: Blockchain could store information about a piece of media’s origin (i.e., metadata), helping people identify credible content.[25]
- *Traceability and History*: Blockchain could help trace media over time and across different sources. Users could see editing history and how content was used in various contexts.[26]
- *Watermarking*: Blockchain could pair with advanced watermarking, which companies like Google are developing.[27] Watermarks could embed unique identifiers in media, allowing one to look up corresponding metadata in blockchain.[28]



Coalition for Content Provenance and Authenticity (C2PA). 2022. Elements of metadata when following C2PA standards for storing and accessing cryptographically verifiable information about a piece of media. https://github.com/c2pa-org/public-draft/blob/2030be1978332cd8a8472f6b2f-c831d94b7a79cd/docs/images/c2pa_visualglossary.png
This image is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

[17] Paris, “Seeing Through the Fog of War.”

[18] John Fletcher, “Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance,” *Theatre Journal* 70, no. 4 (2018): 455–71, https://doi.org/10.1353/tj.2018.0097.

[19] Paradis, “Fighting Deep Fakes.”

[20] Citron and Chesney, “Deep Fakes.”

[21] Lindsey Wilkerson, “Still Waters Run Deep(Fakes): The Rising Concerns of ‘Deepfake’ Technology and Its Influence on Democracy and the First Amendment,” *Missouri Law Review* 86, no. 1 (January 1, 2021), https://scholarship.law.missouri.edu/mlr/vol86/iss1/12.

[22] Citron and Chesney, “Deep Fakes.”

[23] Gregory, “Deepfakes, Misinformation and Disinformation and Authenticity Infrastructure Responses.”

[24] For more information on the private sector’s work to this end, see the Coalition for Content Provenance and Authenticity (C2PA) and Content Authenticity Initiative.

[25] Haya R. Hasan and Khaled Salah, “Combating Deepfake Videos Using Blockchain and Smart Contracts,” *IEEE Access* 7 (2019): 41596–606, https://doi.org/10.1109/ACCESS.2019.2905689.

[26] Hasan and Salah; Christopher Chun Ki Chan et al., “Combating Deepfakes: Multi-LSTM and Blockchain as Proof of Authenticity for Digital Media,” *IEEE*, 2020, 55–62, https://doi.org/10.1109/AI4G50087.2020.9311067; Abbas Yazdinejad et al., “Making Sense of Blockchain for AI Deepfakes Technology,” *IEEE*, 2020, 1–6, https://doi.org/10.1109/GCWkshps50303.2020.9367545.

[27] Josh A. Goldstein and Andrew Lohn, “Deepfakes, Elections, and Shrinking the Liar’s Dividend,” AI and Democracy, *Brennan Center for Justice*, January 23, 2024), https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend; Ryan Heath, “Google Joins Coalition to Label AI-Generated Content,” *Axios*, February 8, 2024, sec. Technology, https://www.axios.com/2024/02/08/google-adobe-label-artificial-intelligence-deepfakes.

[28] Adnan Alattar et al., “A System for Mitigating the Problem of Deepfake News Videos Using Watermarking,” *Electronic Imaging* 32 (January 26, 2020): 1–10, https://doi.org/10.2352/ISSN.2470-1173.2020.4.MWSF-117.

Case Study: Reuters, Canon, and the Starling Lab for Data Integrity

Photo metadata today is easily stripped away or made inaccessible.[1] As part of the Content Authenticity Initiative (CAI), Reuters, Canon, and the Starling Lab for Data Integrity developed a system for storing information about photos in blockchain.[2] Their proof-of-concept in 2023 allowed metadata for a photograph to be digitally signed by a Canon camera at the point of capture, attaching authenticating information directly to the image. This information and the photo were registered onto a public blockchain.

Each subsequent modification from Reuters editors was logged until publication of the final photo, at which point the photo's metadata and blockchain registration information was embedded in the image file. This way, one could look up a photo in a public ledger to access metadata information and to check if the hash values match, i.e., that the image is authentic.

While this project shows the feasibility of using blockchain for proof of authenticity, it also demonstrates potential problems. Two main technical issues arose out of this project: 1) excessive processing time on the camera hindered photographing quality, and 2) difficulty permitting and capturing minor automated edits to photo metadata. Also, as the Starling Lab has observed, this technology carries risk if it "becomes an obligation, not a choice."[3] The lab has warned about the potential for image authentication to surveil or gatekeep journalists, photographers, and their sources. Much more work is necessary to make this technology feasible.

[1] "Image Authentication," Starling Lab, 2021, https://www.starlinglab.org/image-authentication/.

[2] "Preserving Trust in Photojournalism through Authentication Technology," Reuters, 2023, https://www.reutersagency.com/authenticity-poc.

[3] "Image Authentication."

Proponents say such technology could usher in a "new era of digital trust."[29] Yet there's still a long way to go. A critical mass must adopt this technology for it to be effective; otherwise, content without proof of authenticity is just as likely to be authentic as not.[30] Currently, there is a huge amount of content without any provenance,[31] meaning bad actors do not stand out from the crowd.[32]

There are also potential technical limitations:

- *Scaleability*: Constraints around storage size limits, energy, and hardware costs—as well as environmental unsustainability—could affect scalability and uptake of this technology. Further technological advancement and well-designed data-storing frameworks could help.[33]
- *Lack of Transparency*: Many companies deploying this technology lack transparency on their processes, leaving room to doubt their effectiveness or trustworthiness.[34] Accountability structures are especially important for a technology touted for its apparently inherent transparency.
- *Lack of Confidentiality*: At the same time, blockchain's transparency with certain information may limit adoption among people uncomfortable with that information being public.[35] Careful design could permit anonymization of certain information.

Proof-of-authenticity blockchain could even make deepfake disinformation more harmful:

- *Marginalizing the Most At Risk*: Blockchain's extreme transparency could endanger voices that cannot afford to be so transparent. Dissenting activists who fear persecution, for instance, may be reluctant to use blockchain.[36] If they still take the risk of sharing information, they could

[29] Agathe Laurent, "Blockchain Takes on Deepfakes: Ushering in an Era of Digital Veracity," *InCyber News*, January 22, 2024, https://incyber.org/en/article/blockchain-takes-on-deepfakes-ushering-in-an-era-of-digital-veracity/.

[30] Goldstein and Lohn, "Deepfakes, Elections, and Shrinking the Liar's Dividend."

[31] Brandon Khoo, Raphaël C.-W. Phan, and Chern-Hong Lim, "Deepfake Attribution: On the Source Identification of Artificially Generated Images," *WIREs Data Mining and Knowledge Discovery* 12, no. 3 (2022), https://doi.org/10.1002/widm.1438.

[32] David Evan Harris and Lawrence Norden, "Meta's AI Watermarking Plan Is Flimsy, at Best," *IEEE Spectrum*, March 4, 2024, https://spectrum.ieee.org/meta-ai-watermarks.

[33] Alattar et al., "A System for Mitigating the Problem of Deepfake News Videos Using Watermarking."

[34] Hasan and Salah, "Combating Deepfake Videos Using Blockchain and Smart Contracts."

[35] Laurent, "Blockchain Takes on Deepfakes."

[36] Ibid.

Case Study: Verify from Fox Media and Polygon Labs

In August 2023, Fox Corporation and Polygon Labs first launched an open-source protocol called Verify, which would allow news outlets to register their original content on the blockchain.[1] It was publicly released at the beginning of 2024. Currently, the tool allows users to look up image files or article links, in order to check if there is a match with content registered in Verify.[2]

The site's FAQ states that this allows users to "verify the origin and provide traceability for digital content," allowing users to be "confident" that "the content they see attributed to a source that they trust actually was published by that source."[3]

As Variety noted, Fox's focus on authenticity can seem ironic, considering allegations that Fox News knowingly aired falsehoods about Dominion Voting Systems and Smartmatic during the 2020 U.S. presidential election.[4] But Verify's potential for trust-building is not the only promise; other coverage discusses its usefulness for business and licensing negotiations with artificial intelligence companies.[5]

Chief Technology Officer Melody Hildebrandt told Axios that they intend to have all Fox content—including news, sports coverage, and entertainment—go through the Verify protocol eventually.[6]

Since January, there haven't been any concrete updates on the progress Fox has made towards their several goals. It's unclear what their audience thinks of Verify, if they've noticed it at all. Nor have any other large media companies followed suit, though it is still early days. Overall, it remains to be seen whether this initiative is prescient, or a misstep.

[1] Kyle Wiggers, "Fox partners with Polygon Labs to tackle deepfake distrust," *Techcrunch*, January 9, 2024, https://techcrunch.com/2024/01/09/2648953/.

[2] "How to Use," Verify, 2024, https://www.verify.fox/how-to-use..

[3] "FAQ," Verify, 2024, https://www.verify.fox/faqs.

[4] "Fox Launches Tool to Verify Online Content as Authentic — and Not AI-Generated Fakes or Misinformation," Variety, January 9, 2024, https://variety.com/2024/digital/news/fox-verify-authentic-content-ai-misinformation-1235865243/.

[5] Ibid.

[6] Sara Fischer, "Exclusive: Fox Corp. launches blockchain platform to negotiate with AI firms," *Axios*, January 9, 2024, https://www.axios.com/2024/01/09/fox-corp-blockchain-platform-ai-licensing-verify.

suffer harsher retribution because they are forced to disclose information that is potentially identifying or sensitive;[37] if they share information outside the blockchain system, their testimony may seem less credible.

- *Exacerbating Oppression*: Technology is not neutral. Despite what some proponents say, blockchain's decentralization doesn't mean equality. As an example, evangelizing of blockchain in the Global South is often paternalistic and disregards potentially different values and interests. Thus, blockchain could perpetuate systemic oppression and harm.[38]
- *Creating Further Distrust*: Blockchain is technically opaque. Most people would have to take the word of others that it is trustworthy. Regardless of blockchain's robustness, a "leap of faith" is still required to gain trust.[39] This creates room for bad actors to casts doubt on blockchain's effectiveness.
- *Obscuring Weaknesses*: Blockchain's reputation for iron-clad credibility could obscure awareness of serious vulnerabilities, such as quantum computing attacks,[40] or security fixes that are slow to implement due to blockchain's decentralization.[41] This means that information verified through blockchain could be compromised, while still having the veneer of legitimacy.

Recommendations

Blockchain is not a silver bullet against deepfake disinformation and violence. Rather, it is one potential tool in critical engagement with information and media. Whether it will be a success depends on how future developers and purveyors of this technology approach their work.

Furthermore, emphasis on blockchain's technical

[37] Johannes Sedlmeir et al., "The Transparency Challenge of Blockchain in Organizations," *Electronic Markets* 32, no. 3 (September 1, 2022): 1779–94, https://doi.org/10.1007/s12525-022-00536-0.

[38] Syed Omer Husain, Alex Franklin, and Dirk Roep, "The Political Imaginaries of Blockchain Projects: Discerning the Expressions of an Emerging Ecosystem," *Sustainability Science* 15, no. 2 (March 1, 2020): 379–94, https://doi.org/10.1007/s11625-020-00786-x.

[39] Johannes Bennke, "Media of Verification: An Epistemological Framework for Trust in a Digital Society," *Communication +1* 10, no. 1 (December 15, 2023), https://doi.org/10.7275/cpo.1878.

[40] Paula Fraga-Lamas and Tiago M. Fernández-Caramés, "Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality," *IT Professional* 22, no. 2 (March 2020): 53–59, https://doi.org/10.1109/MITP.2020.2977589.

[41] Sunoo Park et al., "Going from Bad to Worse: From Internet Voting to Blockchain Voting," *Journal of Cybersecurity* 7, no. 1 (January 1, 2021), https://doi.org/10.1093/cybsec/tyaa025.

bona fides can overshadow the fact that technology and truth are not neutral. As scholars, governments, and tech companies alike continue to explore ways to deploy blockchain to fight deepfake disinformation, it must be accompanied by research of its social impact in the context of existing systems of power.

> *"With each project, we must question which values are encoded in the system, who controls it, which organizational structure is present to form consensus, and what are the political visions of the code" (Husain et al. 2020).*

In order to make proof-of-authenticity blockchain more useful, I recommend that these parties:

- Adopt high standards of transparency for their methods and processes, to build credibility
- Increase public awareness of blockchain's function, as well as its vulnerabilities, to reduce the risk of confusion
- Accommodate authentic modifications like logos or closed captions to fulfill practical needs
- Expand to different types of media, not just images, to build a more comprehensive and practical system for authenticating content
- Explore ways to limit access to certain information to accommodate those for whom it would be a risk otherwise (see Sedlmeir et al. 2022 for further discussion)[42]
- Research and reflect on power dynamics, cultural context, and potential for harm
- Move away from claims to neutrality, which are rarely true in practice and may backfire, and instead make one's values and blind spots explicit

Lastly, research on deepfakes is rich but also new and scattered. Evidence relating to deepfakes is largely hypothetical or highly technical, lacking empirical grounding and diversity in topics investigated.[43] As this nascent field continues to grow, it is imperative that policymakers work closely with a range of interdisciplinary scholars and tech companies to design effective legislation for combatting deepfake disinformation.

The problem cannot be addressed by proof-of-authenticity blockchain alone. As with many difficult policy problems, a multi-pronged approach across sectors will likely be the most effective.

[42] Johannes Sedlmeir et al., "The Transparency Challenge of Blockchain in Organizations."

[43] Pramukh Nanjundaswamy Vasist and Satish Krishnan, "Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research," *Communications of the Association for Information Systems* 51, no. 1 (November 16, 2022), https://doi.org/10.17705/1CAIS.05126.