2024

Cracking Political Deepfakes: An Exploratory Study to Unveil Cues for Detection

Animesh Kumar

Christoph Burtscher

Andreas Eckhardt

Cracking Political Deepfakes: An Exploratory Study to Unveil Cues for Detection

Extended Abstract

Animesh Kumar
Department of Information Systems, Production and Logistics Management, University of Innsbruck, Innsbruck, Austria

Christoph Burtscher
Department of Information Systems, Production and Logistics Management, University of Innsbruck, Innsbruck, Austria

Andreas Eckhardt
Department of Information Systems, Production and Logistics Management, University of Innsbruck, Innsbruck, Austria

Deepfakes are computer-generated audio-visual media that appear deceptively real or are photorealistic fake representations of real people (Eiserbeck et al., 2023). These images, videos, and audio are typically synthesized using algorithmic models that are a subset of generative adversarial networks (Durall et al., 2020). Unsurprisingly, they can be leveraged for nefarious ends in politics.

Deepfakes have been found to lead to generalized indeterminacy and cynicism toward politics (Vaccari and Chadwick, 2020) and impede citizens' inclusion in debates and decisions (Pawelec, 2022). Disinformation conveyed via deepfakes in elections has been found to pose misrepresentation challenges and may even undermine the legitimacy of the democratic process (Bennett and Livingston, 2018; Dobber et al., 2021).

To date, mitigation measures against deepfakes have focused mainly on computational approaches. The success rate of these computational detection models depends heavily on the availability of the requisite training data (Durall et al., 2020; Silva et al., 2022). However, Groh et al. (2022) found a system that integrates human and computational model predictions is more accurate than humans or the model alone.

To this effect, we embraced an integrative deepfake detection approach that opens the door for intelligence augmentation (Zhou et al., 2021) in the future and creates further opportunities for integrative learning and cross-pollination between human and computational approaches in deepfake detection. In the present study, we focused on deepfake detection by humans and adopted an explorative design rooted in the theory of perception of visual incongruity by Bruner and Postman (1949).

We conducted a preliminary exploratory survey study (Malhotra and Grover, 1998). Studies that focus on examining a given issue through open-ended questions or finding what is out there in the field are considered exploratory. To this end, we surveyed 15 pilot subjects drawn at random from a Prolific-generated sample group and a Qualtrics survey. Subjects were shown four videos, which we drew randomly from the presidential deepfakes dataset (Sankaranarayanan et al., 2021), two for each

19th International Conference on Wirtschaftsinformatik,
September 2024, Würzburg, Germany

presidential candidate of Joseph Biden and Donald Trump, one fake and the other authentic. To analyze the data and identify the utilized cues, we deployed the grounded theory method coding techniques (Glaser, 1992; Urquhart et al., 2010) of open, axial, and selective coding to achieve a detailed analysis and theorizing.

Our preliminary findings point to three key factors that drive humans' detection of deepfake videos. First, a significant majority (64%) of cues deployed to detect deepfakes are of the type tangential (both audio and visual). This corroborates with the theory of Perception of Visual Incongruity, which attributes the 'sense of wrongness' a subject feels when faced with an incongruous stimulus to the subject's focus on a rather tangential but correct aspect of the incongruous stimulus. We also found the corollary to be evident. When asked to articulate, subjects also attributed their 'sense of correctness' about an authentic video to the correctness of the somewhat tangential aspects of the congruous stimulus.

Second, cues that were considered to be of greater social value by some subjects, such as xenophobia and racism, were perceptually accentuated. Such cues were in the minority and accounted for only 21.6% of all listed cues to deepfake videos. Lastly, humans deploy diverse cues to detect deepfake and authentic videos: each subject utilized an average of 3.8 cues per video. This was seen as analogous to the cyclical trial-and-checks process leading to veridical recognition (Bruner and Postman, 1949).

This preliminary study shows three essential contributions. First, the study extends the limited research on understanding the processes of human detection of deepfakes, specifically on the driving determinants of the deepfake perceptual process. Second, these findings extend the application of the theory of perception of incongruity to audio-visual stimuli and broaden its usefulness to generative artificial intelligence. Third, these findings underscore the importance of both audio and visual tangential cues in the detection process and highlight the need to consider a greater diversity of cues in any deepfake detection process. Understanding such cues helps people engaged in political process to identify deepfakes better leading to use of more accurate and correct information in the decision making during the election.

Keywords: Deepfake videos, political deepfakes, generative AI, theory of perception of visual incongruity, grounded theory coding.

References

BENNETT, W. L. & LIVINGSTON, S. 2018. The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication,* 33**,** 122-139.

BRUNER, J. S. & POSTMAN, L. 1949. On the perception of incongruity: A paradigm. *Journal of personality,* 18**,** 206-223.

DOBBER, T., METOUI, N., TRILLING, D., HELBERGER, N. & DE VREESE, C. 2021. Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *The International Journal of Press/Politics,* 26**,** 69-91.

DURALL, R., KEUPER, M., PFREUNDT, F.-J. & KEUPER, J. 2020. Unmasking deepfakes with simple features. *arXiv preprint,* arXiv:1911.00686.

EISERBECK, A., MAIER, M., BAUM, J. & ABDEL RAHMAN, R. 2023. Deepfake smiles matter less—the psychological and neural impact of presumed AI-generated faces. *Scientific Reports,* 13.

GLASER, B. G. 1992. *Emergence vs forcing: Basics of grounded theory analysis*, Sociology Press.

GROH, M., EPSTEIN, Z., FIRESTONE, C. & PICARD, R. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences,* 119**,** e2110013119.

MALHOTRA, M. K. & GROVER, V. 1998. An assessment of survey research in POM: from constructs to theory. *Journal of Operations Management,* 16**,** 407-425.

MITTAL, T., SINHA, R., SWAMINATHAN, V., COLLOMOSSE, J. & MANOCHA, D. Video manipulations beyond faces: A dataset with human-machine analysis. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023. 643-652.

PAWELEC, M. 2022. Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *Digital Society,* 1.

SANKARANARAYANAN, A., GROH, M., PICARD, R. & LIPPMAN, A. The presidential deepfakes dataset. workshop 'AIofAI: 1st workshop on adverse impacts and collateral effects of artificial intelligence technologies'. http://ceur-ws. org, 2021.

SILVA, S. H., BETHANY, M., VOTTO, A. M., SCARFF, I. H., BEEBE, N. & NAJAFIRAD, P. 2022. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy,* 4**,** 1-14.

URQUHART, C., LEHMANN, H. & MYERS, M. D. 2010. Putting the 'theory' back into grounded theory: guidelines for grounded theory studies in information systems. *Information Systems Journal,* 20**,** 357-381.

VACCARI, C. & CHADWICK, A. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society,* 6**,** 205630512090340.

ZHOU, L., PAUL, S., DEMIRKAN, H., YUAN, L., SPOHRER, J., ZHOU, M. & BASU, J. 2021. Intelligence Augmentation: Towards Building Human-machine Symbiotic Relationship. *AIS Transactions on Human-Computer Interaction,* 13**,** 243-264.