Selected Papers of #AoIR2020:
The 21st Annual Conference of the
Association of Internet Researchers
Virtual Event / 27-31 October 2020

DEEPFAKES: A PRELIMINARY SYSTEMATIC REVIEW OF THE LITERATURE

Laura Carvajal
Temple University

Andrew Iliadis
Temple University

Deepfakes are becoming a key topic in debates around politics and misinformation on the internet today. According to Paris and Donovan (2018), “AV manipulation includes both the cutting edge, AI-reliant techniques of deepfakes, as well as “cheap fakes” that use conventional techniques like speeding, slowing, cutting, re-staging, or re-contextualizing footage” (p. 4). While the phenomenon of deepfakes is relatively new with the first documented public appearances recorded in 2017, there is already a growing scholarly literature about deepfakes and the various methods that can be used to help understand and combat them. This paper presents a preliminary systematic review of the academic literature on deepfakes.

Methods

We assessed a representative sample (N=1049) of sources from four popular electronic databases, including Google Scholar, Scopus, Crossref, and Web of Science. These electronic databases were chosen because it is relatively easy to automatically pull reference information from them and because they contain a wide variety of academic literature. This allowed us to glean a snapshot of the type of work that is being done in the academic community on deepfakes.

We used relevant search and key terms "deepfake" OR "deep fake" AND ("machine learning" OR "image generation" OR "video") to generate results and create an archive of the articles related to the development and effects of deepfakes in society. This search yielded potentially relevant journal articles, books, dissertations, and conference papers that were screened for retrieval based on their title and abstract. We then

Suggested Citation (APA): Carvajal, L., & Iliadis, A. (2020, October). Deepfakes: A Preliminary Systematic Review of the Literature. Paper presented at AoIR 2020: The 21th Annual Conference of the Association of Internet Researchers. Virtual Event: AoIR. Retrieved from http://spir.aoir.org.

downloaded CSV files for each of the results from the four databases. Then, all the articles were coded in three areas: academic field (STEM, Humanities and Social Sciences, or Medicine), theoretical approach (prescriptive or descriptive), and themes (automated program, review, state-of-the-art, literacy, best practices, public policy, or critical analysis). These three categories were chosen to determine the field of the published articles, the theoretical approach that academic authors took when addressing deepfakes, and the nature of the output that scholars created to address the possible effects of deep fakes in society.

Initially, we separated the sources with respect to the academic field to which each of them belonged. In “STEM” we included "Science, Technology, Engineering, and Mathematics" (Bybee, 2010), in “Humanities and Social Sciences” we included "Psychology, Economic and Finances, Law, Politics and Public Administration, Sociology, Education, Philosophy, and Linguistics” (Huang & Chang, 2008) as well as Media and Communication, and “Medicine” included all fields related to health. The articles were then classified according to their theoretical approach that could be prescriptive or descriptive. By prescriptive we mean the theories that “are concerned with guidelines that describe what to do in order to achieve specific outcomes” (Ullrich, 2008, p.37) and by descriptive we mean when “theories make statements about how learning occurs and devise models that can be used to explain and predict learning results” (ibid). Since our interest is to understand how academics are approaching the study of deepfakes, knowing whether a study is descriptive or prescriptive allows us to see where research is heading; we want to know if the study of deepfakes is primarily around recommendations or concrete outcomes. Lastly, we wanted to see what types of conversations are happening, whether they are primarily around public knowledge or on practices and technological solutions.

After separating the articles by theoretical approach, we classified the sources into seven themes that defined their purpose. The first one was automated programs to find an optimal technical solution. The second theme was a review paper that tries to succinctly review recent progress in a particular topic, and the state-of-the-art papers reflect the present state of scientific or engineering development. The fourth theme was literacy articles, which focus on the documents that generate awareness about the effects of deepfakes. The fifth category was public policy focusing on articles that specify the rules, guidelines, and regulations that government and non-profit organizations may take with respect to deepfakes. The best practices theme corresponds to papers where a practice “has been shown to produce superior performance” (Druery, J & McCormack, N & Murphy, S, 2013, p.111); these articles show specific guidelines towards the use or identification of deepfakes. The last theme, critical analysis, included sources that examine ideas, perspectives, or critical opinions.

After classifying the sources in these categories, statistics were carried out to show the aspects and relationships of the data sample. For each electronic database, three-bar diagrams were made for each category (academic field, theory approach, and theme) to show the relation between them. Finally, a total bar chart was made with results from all four electronic databases to have an overview of the approach of published sources concerning deepfakes (Figure 1).

Analysis

The systematic review of the sources through the three categories allowed us to make some general and preliminary conclusions. First, checking the data, most of the articles belong to the STEM area. This may confirm that the academic community is interested in further developing the technology behind deepfakes and may imply that the main focus of researchers is related to the development of automatic programs allowing for advanced and faster responses.

From the 1049 articles, 300 of them belong to the area of Humanities and Social Sciences, and most of the sources in this field were focusing on state-of-the-art, literacy, and critical analysis. This may show that academics in those areas are establishing ground that informs users about deepfakes, answering questions such as “What are deepfakes? What are the possible effects? or What are the threats to democracy?” These documents generate debate about what practices scholars should study to recognize deepfakes and mitigate their negative effects. In response to this, in the database, we found that some academics understand the significance of writing literature that monitors the production and detection of deepfakes. From the database, 53 sources explore topics related to public policies, and 35 sources studied the best practices. STEM produced 381 articles that in different ways promote innovation concerning deepfakes. 530 sources have a descriptive approach, implying that deepfakes are a new topic requiring discussion through the observation of models and the prediction of results. The lopsidedness in prescriptive and descriptive approaches (descriptive was almost double that of prescriptive) may show that there could be a need increase the production of articles that focus on the technical automation of detecting deepfakes.

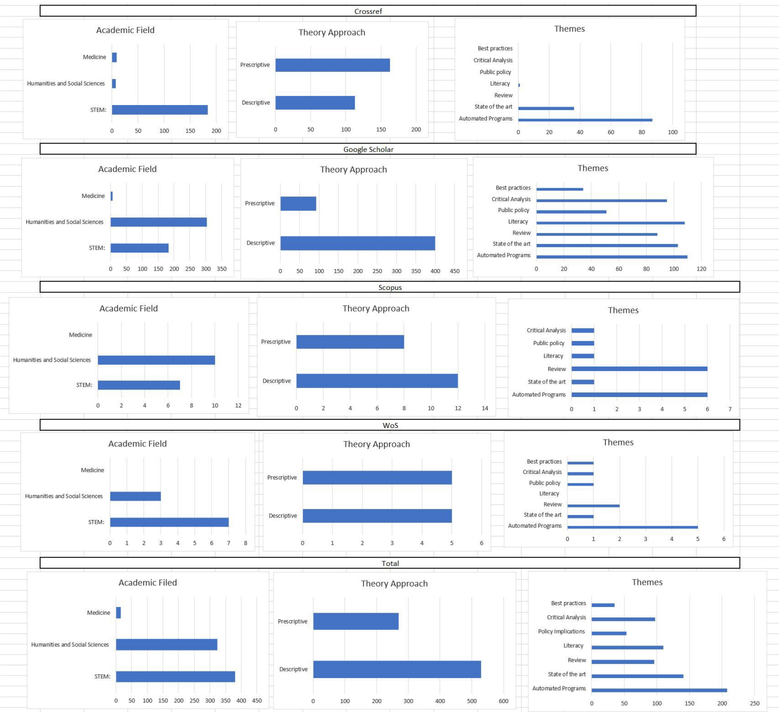

Figure 1 Results from coding deepfake scholarly literature

References

Paris, B. and Donovan, J. (2018). Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. *Data & Society*. Retrieved from Data & Society: https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf

Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. *Technology and Engineering Teacher*, 70(1), 30-35.

Druery, J & McCormack, N & Murphy, S. (2013). Are Best Practices Really Best? A Review of the Best Practices Literature in Library and Information Studies. *Evidence Based Library and Information Practice*. 8. 110-128. 10.18438/B8RC9S.

Huang, M. H., and Chang, Y. W. (2008). Characteristics of Research Output in Social Sciences and Humanities: From a Research Evaluation Perspective. *Journal of the American Society for Information Science and Technology*, 59(11), 1819-1828.

Ullrich, C. (2008). Descriptive and Prescriptive Learning Theories. In *Pedagogically Founded Courseware Generation for Web-Based Learning* (pp. 37-42). Berlin: Springer.