

UNITED STATES DISTRICT COURT  
DISTRICT OF MINNESOTA

CHRISTOPHER KOHLS and  
MARY FRANSON,

Court File No. 24-cv-03754 (LMP/DLM)

Plaintiffs,

vs.

KEITH ELLISON, in his official capacity  
as Attorney General of Minnesota, and  
CHAD LARSON, in his official capacity  
as County Attorney of Douglas County.

**DECLARATION OF PROFESSOR  
JEFFREY HANCOCK IN SUPPORT  
OF MOTION FOR LEAVE TO FILE  
AN AMENDED DECLARATION**

Defendants.

I, Jeffrey Hancock, declare as follows:

1. I am writing to acknowledge three citation errors in my expert declaration, which was filed in this case on November 1, 2024 (ECF No. 23). I wrote and reviewed the substance of the declaration, and I stand firmly behind each of the claims made in it, all of which are supported by the most recent scholarly research in the field and reflect my opinion as an expert regarding the impact of AI technology on misinformation and its societal effects. Attached as **Exhibit 1** is a redline version of the corrected expert declaration, and attached as **Exhibit 2** is a redline version of the corrected list of academic and other references cited in the expert declaration.

2. The first citation error appears in paragraph 19 and cites to a nonexistent 2023 article by De keersmaecker & Roets. The correct citation for the proposition is to Hancock & Bailenson (2021), cited in paragraph 17(iv). The second citation error appears in paragraph 21, a citation to a nonexistent 2023 article by Hwang et al., and is identified

by the Plaintiffs in their motion to exclude my declaration. The correct citation for that proposition is to Vaccari & Chadwick (2020), which appears in paragraph 20. The third citation error appears in Exhibit C to the declaration (ECF No. 23-1, at 39): the citation to Goldstein et al. lists the first author correctly, but the remaining authors are incorrect. The correct authors are Goldstein, J., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. I discovered the errors in paragraph 19 and Exhibit C when Plaintiffs brought the error in paragraph 21 to the Court's attention, and I re-reviewed my declaration.

3. I apologize to the Court for these three citation errors, and I explain how they came to be below. I did not intend to mislead the Court or counsel. I express my sincere regret for any confusion this may have caused. That said, I stand firmly behind all of the substantive points in the declaration. Both of the correct citations were already cited in the original declaration and should have been included in paragraphs 19 and 21. The substantive points are all supported by the scientific evidence and correcting these errors does not in any way alter my original conclusions.

4. To provide an explanation for how these errors occurred and why they do not impact any of the scientific evidence or opinions presented in my declaration, I lay out more specifics about my expertise, some of the context regarding the production of the declaration, my workflow for researching and drafting this report, and the role that AI tools played in my process.

5. In mid-October I was contacted by counsel for Defendants to provide an expert declaration. I agreed to do so in my personal capacity as part of my private consulting work. I produced an expert declaration regarding the scientific literature on the

psychological and social implications of deepfakes, particularly in the context of credibility, their spread online, their influence on trust in the media, and the effectiveness of countermeasures. These issues are squarely within my area of expertise. I have written and published over 15 studies on AI and communication since 2017, and I co-wrote the foundational piece on AI-Mediated Communication, which has been cited over 400 times and is the most cited paper in this new field. I co-edited the first special issue on the social consequences of deepfakes in *Cyberpsychology and Behavior and Social Networking*, a high impact peer-reviewed journal that publishes research examining the social, behavioral and psychological impact of emerging technologies. I have published extensively on misinformation in particular, including the psychological dynamics of misinformation, its prevalence, and possible solutions and interventions, and I have published a study examining misinformation in virtual reality.

6. As a researcher and professor in the field of AI, my work at present already focuses on the challenges deepfakes pose. The literature has increased dramatically with the release of ChatGPT in November 2022, however, which has accelerated the development of tools that can be used to create deepfakes. For example, since the publication of the special issue on the social consequences of deepfakes that I co-edited in 2021, there have already been over 140 citations to our article leading that special issue (Hancock & Bailenson, 2021), which represents a very high citation rate in the social sciences for a recent article. For example, after only 3 years, this citation rate would almost rank in the top 25% of all social science publications (see Ionnadis et al., 2019). Since the release of ChatGPT in 2022, I have also published five peer-reviewed papers on the impact

of AI on trustworthiness and communication (Hohenstein et al, 2023; Jakesch, Naaman & Hancock 2023; Karinshak et al, 2023; Markowitz & Hancock, 2024; Markowitz et al, 2024). Because I cofounded the *Journal of Online Trust and Safety*, which is a leading journal on research related to, among other things, misinformation and deepfakes online, and because I am asked to present on my research and the impacts of AI on society, I regularly review the latest scientific literature in fields that intersect with my own. I also teach a graduate class on language and technology that focuses on AI and communication.

7. My workflow for this declaration involved three main phases: (a) surveying the literature, (b) analyzing the scientific evidence, and (c) drafting the declaration.

8. I began by surveying the literature on deepfakes, which I know well but wanted to refresh for the most recent scholarship. This a highly active research area, spanning several fields and disciplines, including computer science, engineering, psychology, communication, human-computer interaction and law, and there are thousands of citations related to the term “deepfakes.” To go through the large volume of scientific evidence related to deepfakes in the survey phase, I primarily used two tools (namely, Google Scholar and GPT-4o) to identify articles that were likely to be relevant to the declaration so that I could merge that which I knew already with new scholarship. Google Scholar is a tool that allows researchers to search across many scholarly disciplines and sources for journal articles, conference proceedings, books, online repositories and professional societies. Google Scholar provides indexed information about publications, how often they have been cited and by which other publications, and it provides links to locate copies of the publications. GPT-4o is a generative AI tool referred to as a large

language model (LLM) that powers chatbots and other agents, such as ChatGPT, as well as an increasing number of online services and technologies, including search, translation, drafting, and document summarization and analysis. Indeed, the use of generative AI tools is on the rise given that most current tools used for writing and research, such as Microsoft Word, Google Docs, Adobe Acrobat and even email services like Gmail and Outlook, offer AI-based functions that complete or suggest sentences or provide initial drafts of communication. In general, I use tools like GPT-4o to enhance the quality and efficiency of my workflow, including search, analysis, formatting and drafting. For this declaration, during the survey phase I used GPT-4o to assist with searching for and identifying articles that would likely be relevant to the declaration. I am already familiar with many of these articles and conclusions, and the search results included articles I authored.

9. In the analysis phase, one part of my process is to use GPT-4o to summarize some of the relevant articles identified in the survey phase so that I might identify themes and research questions that emerged across studies. Examples of such themes and research questions include how deepfakes may be detected by humans or whether deepfakes can undermine trust in news or media, which included themes and research questions that are already known to me and new ones that more recently emerged in the field (such as new ways of analyzing human perception of deepfakes). I used the GPT-4o tool to both verify my knowledge of literature that I had both read and written and to make sure that there was no new scholarship that I needed to be aware of as I was writing my declaration. I used GPT-4o and Google Scholar to produce an initial list of references that I would consider citing in the declaration.

10. The citation errors here occurred in the drafting phase, and as such, I explain my process in granular detail here. The drafting phase involved two parts – the substance and the citations. As to the substance, I began by outlining the main sections of the declaration in MS Word. I then outlined the key substantive points for each section, also in MS Word. I continued to engage Google Scholar and GPT-4o.

11. The two citation errors, popularly referred to as “hallucinations,” likely occurred in my use of GPT-4o, which is web-based and widely used by academics and students as a research and drafting tool. “Hallucinated citations” are references to articles that do not exist. In the drafting phase I sometimes cut and pasted the bullet points I had written into MS Word (based on my research for the declaration from the prior search and analysis phases) into GPT-4o. I thereby created prompts for GPT-4o to assist with my drafting process. Specifically for these two paragraphs, I cannot remember exactly what I wrote but as I want to try to recall to the best of my abilities, I would have written something like this as a prompt for GPT-4o: (a) for paragraph 19: “draft a short paragraph based on the following points: -deepfake videos are more likely to be believed, -they draw on multiple senses, - public figures depicted as doing/saying things they did not would exploit cognitive biases to believe video [cite]”; and (b) for paragraph 21: “draft a short paragraph based on the following points: -new technology can create realistic reproductions of human appearance and behavior, -recent study shows that people have difficulty determining real or fake even after deepfake is revealed, -deepfakes are especially problematic on social media [cite].”

12. When I inserted the bullet points pertaining to paragraphs 19 and 21 into GPT-4o I also included the word “[cite]” as a placeholder to remind to myself to go back and add the academic citation. As I explained earlier, both of the now corrected cites were articles that I was very familiar with – one of which I wrote myself. I did not mean for GPT-4o to insert a citation, but in the cut and paste from MS Word to GPT-4o, GPT-4o must have interpreted my note to myself as a command. The response from GPT-4o, then, was to generate a citation, which is where I believe the hallucinated citations came from. This only happened in these two instances and nowhere else in my declaration.

13. When GPT-4o provided me these answers, I cut and pasted them from the online tool into my MS Word declaration. I then edited my declaration extensively as to its substance, and where I had notes to myself in both instances to add the citation, GPT-4o had put them in for me incorrectly and deleted the “[cite]” placeholder I had included to remind myself to go back and include the right citation. Without the “[cite]” placeholders, I overlooked the two hallucinated citations and did not remember to include the correct ones. This was the error on my part, and as I stated earlier, I am sorry for my oversight in both instances here and for the additional work it has taken to explain and correct this.

14. Finally, the last part of the drafting phase involves the citation list. I asked GPT-4o to generate a reference list in APA format using the in-text citations already in the draft declaration. Given the relatively small number of citations, I did not run this through the reference software I typically use when I write my academic papers, which might have caught the hallucinated citations. Here too, I realize that I should have vetted this through my usual process. I then mistakenly included the two citations that were created by GPT-

4o. I meant to cite to the source that was immediately before the erroneous one in both implicated paragraphs.

15. Given that background, the citation error identified by plaintiffs occurred in paragraph 21, which stated: “Moreover, the difficulty in disbelieving deepfakes stems from the sophisticated technology used to create seamless and lifelike reproductions of a person’s appearance and voice. One study found that even when individuals are informed about the existence of deepfakes, they may still struggle to distinguish between real and manipulated content. This challenge is exacerbated on social media platforms, where deepfakes can spread rapidly before they are identified and removed. (Hwang et al., 2023).”

16. The correct citation for paragraph 21, which is cited immediately above in paragraph 20, is to Vaccari & Chadwick (2020), a paper that I was already familiar with as part of my research. This paper reports on a large study that found that even when people were informed that a video is a deepfake, approximately 44 percent of the participants were deceived by or uncertain of (i.e., struggled with) the manipulated video. This citation directly supports the proposition regarding the study described in paragraph 21.

17. This substantive point in the declaration has also been empirically supported by other research, including consulted citations in the references. For example, one study finds that even when participants were warned about the existence of deepfakes, their ability to detect deepfakes remained low (Köbis et al, 2021). Another recent study, published in *The International Journal of Press/Politics* (a Sage journal with a high impact factor, which was not cited in the original declaration) also supports this claim (Weikmann et al, 2024). After revealing to participants that they had been exposed to a deepfake,



participants' self-efficacy for detecting deepfakes declined. That is, participants' confidence in detecting deepfakes declined after they were informed about deepfakes.

18. The citation error that I identified after re-review appears in paragraph 19, which stated "Research indicates that deepfake videos are more likely to be believed than text-based misinformation because they engage multiple senses simultaneously, creating a stronger illusion of authenticity. The realistic portrayal of individuals, especially public figures, engaging in fabricated actions or statements exploits the cognitive biases that lead people to accept visual and auditory information as truth. (De keersmaecker & Roets, 2023)."

19. The correct citation is to Hancock & Bailenson (2021) for paragraph 19, which is cited above in paragraph 17. I co-authored this article, and it lays out why the visual medium is so dominant in human perception and why communication research indicates that misleading audiovisual information may be more likely to be trusted than verbal messages (Hancock & Bailenson, 2021, p. 150). Specifically, the article states: "the impact of deception by deepfake has the potential to be greater than that of verbal deception because of the primacy of visual communication for human cognition. Deepfakes not only change verbal content, but they also change the visual properties of how the message was conveyed, whether this includes the movement of a person's mouth saying something that he or she actually did not, or the behavior of a person doing something that he or she did not. The dominance of visual signals in human perception is well established."

20. The claim that the visual medium is dominant in human perception is not controversial. It has also been empirically supported by additional and more recent research

on deepfakes. In one study that sought to examine why doctored videos shared online lead to strong reactions among users that resulted in the wrongful death of innocent people, participants found video misinformation to be more credible than audio or text-based misinformation, and they were more likely to share it. This finding, published in the *Journal of Computer-Mediated Communication*, one of the top ranked journals in communication (ranked #6 out of 227 journals in the communication category), is entirely consistent with the substantive point in paragraph 17 (Sundar et al, 2021).

21. Finally, as noted above, upon the re-review I identified a minor error in the authorship in a citation (Goldstein et al, 2023). The correct author list for this source is: Goldstein, J., Sastry, G., Musser, M., DiResta, R., Gentzel, M., and Sedova, K. As noted in paragraph 14 above, I did not run the citation list through the reference software I typically use when I write my academic papers, resulting in the inclusion of the two hallucinated sources. I believe GPT-4o also generated an incorrect list of secondary authors for Goldstein et al.—a real paper with which I was already familiar.

22. In conclusion, despite the presence of these three mistakes, I stand firmly behind the substantive points presented in the report. As demonstrated in the scientific evidence, the correct citations already used in the report provide empirical support for the specific arguments made. In particular, both Hancock and Bailenson (2021) and Vaccari and Chadwick (2020) already appear in the declaration immediately before the now corrected paragraphs.

***[Signature on the following page]***

PURSUANT TO 28 U.S.C. § 1746, I DECLARE UNDER PENALTY OF PERJURY THAT EVERYTHING I HAVE STATED IN THIS DOCUMENT IS TRUE AND CORRECT.

Dated: November 27, 2024

  
 \_\_\_\_\_  
 JEFFREY HANCOCK

## References

Ioannidis, J. P., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLoS biology*, *17*(8), e3000384.

Markowitz, D. M., Boyd, R. L., & Blackburn, K. (2024). From silicon to solutions: AI's impending impact on research and discovery. *Frontiers in Social Psychology*, *2*, 1392128.

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps?. *Journal of Computer-Mediated Communication*, *26*(6), 301-319.

Weikmann, T., Greber, H., & Nikolaou, A. (2024). After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media. *The International Journal of Press/Politics*, 19401612241233539.

## Five Hancock peer-reviewed AI & communication publications since ChatGPT release in 2022

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J.T., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*(1), 5487.

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), e2208839120.

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1-29.

Markowitz, D. M., & Hancock, J. T. (2024). Generative AI are more truth-biased than humans: A replication and extension of core truth-default theory principles. *Journal of Language and Social Psychology*, *43*(2), 261-267.

Markowitz, D. M., Hancock, J. T., & Bailenson, J. N. (2024). Linguistic markers of inherently false AI communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1), 63-82.