UNITED STATES DISTRICT COURT
DISTRICT OF MINNESOTA

CHRISTOPHER KOHLS and MARY FRANSON,

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.

Defendants.

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

**EXPERT DECLARATION OF PROFESSOR JEFF HANCOCK**

I, Jeff Hancock, declare as follows:

1. I am the founding director of the Stanford Social Media Lab and the Harry and Norman Chandler Professor of Communication at Stanford University. I am also the Faculty Director of the Stanford Internet Observatory and co-Director of the Stanford Cyber Policy Center, and a senior fellow at the Freeman Spogli Institute. I am also Founding Editor of the Journal of Trust & Safety, the leading journal focusing on research related to spam and fraud online, misinformation and deepfakes, child exploitation and non-consensual intimate imagery, suicide and self harm, hate speech and harassment.

2. I am a leading expert in social media behavior and the psychology of online interaction. I conduct studies on the impact of social media and artificial intelligence (AI) technology on misinformation and trust, on psychological well-being, digital literacy and how we use and understand language. Recently I have begun work on understanding the mental models people have about algorithms and AI, as well as working on the ethical

issues associated with computational social science. I have received numerous awards for my research, and I have published over 100 journal articles and conference proceedings. My research has been supported by funding from the U.S. National Science Foundation and the U.S. Department of Defense.

3. A full overview of my professional experience and publications is provided in my full academic curriculum vitae, a copy of which is attached as **Exhibit A**, and my shorter expert witness curriculum vitae, a copy of which is attached as **Exhibit B**.

4. I have further identified the academic, scientific, and other materials referenced in this declaration in the references attached as **Exhibit C**.

5. I have been retained by the Office of the Minnesota Attorney General to provide expert opinion and testimony regarding how AI is influencing misinformation on social media and the psychological impact of deceptive deepfakes, particularly deepfakes shared on social media. I am being compensated at my government rate of $600 per hour.

6. I have reviewed Minnesota's electoral deepfake law, Minnesota Statutes section 609.771, and the Complaint in this case. In this declaration, I provide my expert views, based on my training and experience, with reference to recent research, on:

- how AI is changing the way that misinformation is shared and received on social media;
- the psychological impact of deepfakes, with a particular focus on how deepfakes undermine trust, including in the electoral context; and
- the limits of corrections, fact-checking, and other counterspeech and how they are particularly ineffective in the context of deepfakes.

## I. BACKGROUND ON AI, SOCIAL MEDIA, AND THE SPREAD OF DEEPFAKES.

### *Overview of AI and its Use on Social Media Platforms Today*

7. Artificial Intelligence (AI) refers to computer systems designed to perform tasks that typically require human intelligence, such as recognizing speech, making decisions, solving problems, and understanding natural language (Russell & Norvig, 2021). AI systems learn from data and can improve their performance over time without being explicitly programmed for each specific task.

8. Generative AI is a subset of AI focused on creating new content—such as text, images, or videos—by learning patterns from large datasets. A prominent example of generative AI is Large Language Models (LLMs), which are trained on extensive collections of text from books, articles, websites, and other sources. These models generate human-like text by predicting word sequences based on the input they receive (Radford et al., 2019); LLMs can now be used frequently in human communication contexts like social media (Hancock, Naaman & Levy, 2020).

9. On social media platforms, generative AI and LLMs are increasingly used in various ways:

i) *Content Creation*: AI assists users in drafting social media posts, blogs, and comments. Users can input a brief prompt, and the AI generates a full post or article in a conversational or professional tone. This can dramatically accelerate massive content production and allows for personalized communication.

ii) *Chatbots and Virtual Assistants*: Many platforms use LLMs to power chatbots that interact with users by answering questions or engaging in conversations. These AI systems simulate human dialogue to provide customer service, product recommendations, or social companionship.

iii) *Content Moderation*: AI is employed by social media platforms to detect and flag inappropriate content. LLMs help recognize harmful language, disinformation, or policy violations. However, AI-based moderation can be imperfect and subject to biases inherent in the training data.

iv) *Personalized Recommendations*: AI recommends content to users based on their past interactions, preferences, and engagement patterns. This includes suggesting posts, videos, or advertisements that align with the user's interests, enhancing user engagement.

v) *Image and Video Generation*: Generative AI creates or enhances images and videos. Tools enabling users to produce stylized profile pictures or AI-generated artwork have gained substantial popularity on social platforms by allowing users to more easily create and modify images and videos.

***Overview of Deepfakes and the Impact of AI on Deepfakes***

10. Deepfakes are highly realistic, AI-generated manipulations of digital content—typically videos or images—where a person's likeness, voice, or actions are convincingly altered or fabricated. These forgeries utilize advanced artificial intelligence techniques to simulate human behavior and appearance, making it difficult for the average person to discern that the content is fake. The term "deepfake" combines "deep learning" (a subset of AI) and "fake," highlighting the use of complex AI methods to create these deceptive materials.

11. Deepfakes are primarily produced using a type of AI architecture known as Generative Adversarial Networks (GANs). GANs consist of two neural networks: the generator and the discriminator. The generator creates fake content, while the discriminator attempts to distinguish between real and fake content. Through this adversarial process, the generator progressively improves its ability to produce realistic media that can mimic human faces, voices, and movements with remarkable accuracy (Farid, 2022).

12. To create a deepfake, substantial amounts of data—such as photos, videos, or audio clips of the target individual—are input into the AI model. The AI analyzes this data to learn the person's facial expressions, voice patterns, and mannerisms. It then applies these learned characteristics to another piece of footage or generates new, fabricated scenes. The result is often a seamless video or image that can deceive viewers into believing the person is performing actions or speaking words they never actually did.

13. Deepfakes distinguish themselves from traditional digital manipulations due to their high level of realism and ability to replicate nuanced human behaviors. Deepfakes capture subtle details like facial muscle movements and speech intonations, making them significantly more challenging to detect without specialized tools (Farid, 2022).

14. Advancements in generative AI, particularly through GANs, have significantly transformed the production of deepfakes. The period between 2017 and 2019 marks a significant shift in the ease of creating deepfakes. While GANs were introduced in 2014, it was the subsequent improvements in model architectures and the release of accessible tools that lowered technical barriers. These developments allowed individuals without specialized expertise to generate convincing deepfake content, leading to increased prevalence and concern over the misuse of such technology. These technological developments have lowered the barrier to creating deepfakes, requiring minimal technical expertise. What was once the domain of skilled programmers is now accessible to the general public via user-friendly generative AI tools and open-source platforms. This increased accessibility has led to a surge in the creation and dissemination of deepfakes.

15. This democratization of deepfake technology raises serious concerns due to its potential misuse in spreading disinformation and manipulating public perception. Deepfakes have been employed to impersonate individuals in fraudulent schemes, manipulate media content, and craft false narratives on social media platforms. As these AI-generated forgeries become more sophisticated, they pose a growing threat to trust in authentic media and public figures, especially when rapidly disseminated through social platforms. Deepfakes can erode public trust and have significant psychological and societal effects. Deepfakes contribute to the spread of misinformation, create confusion about the authenticity of legitimate content, and potentially harm individuals' reputations and well-being (Hancock & Bailenson, 2021).

16. The combination of generative AI's capabilities and the viral nature of social media presents a significant threat to online trust and safety. In response, researchers and platforms are working to develop advanced detection technologies and establish policy frameworks to counter the rise of deepfakes. However, the rapid evolution of AI tools continues to present challenges in effectively mitigating these risks.

***AI is Transforming the Misinformation Landscape***

17. The advent of generative AI models has significantly transformed the landscape of misinformation by changing how it is created, shared, and consumed. These systems have dramatically lowered the barriers to producing convincing misinformation at scale. This transformation presents new challenges compared to traditional forms of media:

i. *Proliferation of Misinformation Due to Reduced Costs ("Flooding the Zone")*: Generative AI has made it cheap and efficient to produce large volumes of misleading or false content, the low cost and high speed of AI-

generated content enable malicious actors to "flood the zone" with misinformation, overwhelming information ecosystems and making it difficult for consumers to discern truth from falsehood. This saturation diminishes the visibility of accurate information and can manipulate public opinion by sheer volume (Goldstein et al., 2023).

ii. *Tailoring and Personalization*: AI technologies allow for the tailoring of misinformation to specific audiences or even individuals. By leveraging data on user preferences and behaviors, AI can generate personalized content that resonates with the target's beliefs and biases. This personalization increases the effectiveness of influence operations by making messages more persuasive and harder for individuals to dismiss (Goldstein et al., 2023).

iii. *Difficulty in Detection*: The sophistication of AI-generated content makes it increasingly difficult to detect misinformation. Advanced models can produce text, images, audio, and video that closely mimic authentic content. Traditional detection methods, which rely on identifying inconsistencies or known patterns of falsehoods, are less effective against AI-generated misinformation. As AI models improve, the artifacts that once signaled fake content diminish, complicating efforts by platforms and researchers to identify and remove such content (Farid, 2022).

iv. *Enhanced Persuasive Power Through Audio and Video*: The ability of AI to generate realistic audio and video deepfakes amplifies the persuasive power of misinformation. Audio and visual content have a stronger impact on human perception and memory compared to text. By creating convincing deepfake videos or audio recordings of individuals appearing to say or do things they never did, AI can significantly influence public opinion and damage reputations. Deepfakes in audio and video formats can exploit the trust people place in sensory information, making the misinformation more effective and much more difficult to refute (Hancock & Bailenson, 2021).

## II. THE PSYCHOLOGICAL IMPACTS OF DEEPFAKES.

18. The media characteristics of deepfakes significantly enhance the persuasiveness of misinformation due to their ability to mimic real-life sensory experiences. Humans naturally place a high degree of trust in what they see and hear, making deepfakes particularly effective in influencing beliefs and opinions.

19. 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).

20. Deepfakes can significantly influence political beliefs by presenting convincing false narratives that are difficult to refute. The visual and auditory realism of deepfakes can undermine trust in legitimate media sources and political institutions, leading to confusion and polarization among the public (Vaccari & Chadwick, 2020).

21. 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).

22. People are more likely to doubt the authenticity of a political video if the content is inconsistent with their perceptions of the politician’s typical behavior or known viewpoints. Familiarity with the politician and the implausibility of the statements made in the deepfake increased skepticism among viewers. However, if the deepfake aligns with the viewer’s expectations or biases, they are less likely to question its authenticity (Hameleers et al, 2024).

23. The audio and video aspects of deepfakes not only enhances their persuasive impact but also poses significant risks to democratic processes and societal trust. The ability to fabricate credible evidence can be used to discredit public figures, incite unrest, or manipulate electoral outcomes. As such, deepfakes represent a potent tool for malicious actors seeking to influence political beliefs, disrupt social cohesion and undermine trust in institutions (Hancock & Bailenson, 2021).

## III. THE LIMITS OF FACT-CHECKING AND OTHER COUNTERSPEECH.

24. Traditional fact-checking methods are less effective in combating deepfakes due to the sophisticated and deceptive nature of these manipulated audio and video files. Deepfakes pose unique challenges because they can produce highly realistic fabrications of individuals appearing to say or do things they never did, which are difficult to detect without specialized technical tools and expertise. Fact-checkers relying on conventional techniques may struggle to verify the authenticity of such content, as deepfakes often lack the textual inconsistencies or factual errors that traditional methods uncover.

25. Another significant challenge in combating deepfakes is the rapid speed at which they can spread on social media platforms, often outpacing traditional forms of content due to their sensational and visually compelling nature. Deepfakes are designed to capture attention and provoke strong emotional responses, which increases their likelihood of going viral quickly and reaching large audiences before detection and removal are possible (Goldstein, 2023). For example, in 2021, a series of deepfake videos featuring actor Tom Cruise emerged on TikTok under the account "@deeptomcruise." The sophistication of the deepfake technology used made it difficult for viewers to distinguish

these videos from authentic footage. Within days, the account amassed over 11 million views and attracted hundreds of thousands of followers, demonstrating how quickly such content can captivate and deceive large audiences on social media platforms (Metz, 2021).

26. One traditional approach to combatting deepfakes is labeling. Research demonstrates that labeling content as a "deepfake" can significantly influence how audiences perceive both manipulated and authentic information. Research suggests that when a video is labeled as a deepfake or as manipulated by AI, people are more likely to be skeptical of it, suggesting labels can be effective. But other research suggests labeling can also be used to undermine authentic information. In one study (Hameleers & Marquart, 2020), when an authentic political speech was labeled as a deepfake, participants perceived it as less credible and less authentic. This indicates that while labeling can alert viewers to potential manipulations, it can also undermine the credibility of genuine content when misapplied. Erroneously labeling authentic content as a deepfake can delegitimize truthful communication and diminish public confidence in accurate information.

27. While labeling strategies can be effective in helping audiences identify manipulated media, labeling can also produce a "liar's dividend," where individuals exploit the existence of deepfakes to deny the authenticity of real events by claiming they are fabricated. Misapplication of deepfake labels can provide malicious actors with plausible deniability, allowing them to dismiss genuine evidence as false.

28. Another important concern about deepfakes is their potential to create false memories, in which a person's recollection matches the deepfake version of an event rather than the actual event. In one recent study by false memory scholar Elizabeth Loftus and

her colleagues (2024), participants were first presented with original images to establish a baseline, and after a filler task they viewed AI-modified versions of that image, including a condition with AI-generated videos. The AI modifications included changes like an increased military presence. When participants' memories were assessed, the percentage of false memories were two times higher in the deepfake video condition than the control, suggesting that deepfakes can significantly increase false memories.

29. Repeated exposures to deepfakes are also an important problem. A recent large scale study across 8 countries (Ahmed et al., 2024) investigated how deepfakes contribute to the Illusory Truth Effect (ITE)—a psychological phenomenon where repeated exposure to misinformation increases its perceived accuracy. The study found that individuals who had previously been exposed to deepfakes were more likely to perceive them as accurate compared to those encountering them for the first time, for both political and non-political deepfakes. The study found that social media news consumption amplifies the ITE for all individuals, irrespective of their cognitive ability levels. This suggests that even individuals with higher cognitive abilities are not immune when they heavily engage with news on social media.

30. These psychological studies confirm that exposure to deepfakes can undermine trust in institutions by making it harder for individuals to discern true information from falsehoods and to recall what is true versus a false memory. The exposure to deepfakes contributes to increased uncertainty, not only about the veracity of information, but about the veracity of our memories. This skepticism can extend to legitimate news and official communications, thereby undermining trust in traditional

information gatekeepers and institutions. By making false political information more believable, deepfakes can potentially influence public opinion and electoral outcomes, posing a threat to democratic institutions and processes (Hancock & Bailenson, 2021).

## IV. CONCLUSION

31. The advancement of generative AI and the proliferation of deepfakes present significant challenges to the integrity of information and the trust placed in institutions. Deepfakes leverage sophisticated AI technologies to create highly realistic and persuasive misinformation that can spread rapidly on social media platforms. The psychological impacts are profound: deepfakes can manipulate perceptions, create false memories, and exploit cognitive biases, making it difficult for individuals to discern truth from falsehood.

32. Traditional methods of fact-checking and counterspeech are insufficient to address the sophisticated and rapidly spreading nature of deepfakes. Regulatory measures, combined with technological solutions and public awareness efforts, are necessary to combat the risks associated with deepfakes. In my expert view, enacting and enforcing laws that specifically target the production and dissemination of deceptive deepfake content during elections are critical in preserving the integrity of democratic institutions and protecting the foundational trust upon which they rely.

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: Nov 1, 2024

JEFF HANCOCK