Deepfakes and the Social Construction of Trust: A SCOT Analysis of Stakeholder Responses

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Baani Kaur

Spring 2025

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Travis Elliott, Department of Engineering and Society

Deepfakes and the Social Construction of Trust: A SCOT Analysis of Stakeholder Responses

Introduction to STS Topic

Artificial intelligence (AI) has led to the rapid rise of deepfakes which are synthetic videos, images, and audio that closely resemble real media. Once niche and experimental, deepfakes are now widespread and appear across social media platforms where they are often indistinguishable from authentic content. As a result, they are challenging public confidence in the information they encounter online. While some deepfakes are used creatively in entertainment or art, others are designed to deceive, spread misinformation, or impersonate individuals. These applications pose a threat to the credibility of digital media and to the platforms that host it. This paper looks at how deepfakes impact public trust in social media and explores how different groups understand and respond to this evolving threat. For analysis, the Social Construction of Technology (SCOT) framework is applied. SCOT emphasizes the role of social interpretation in shaping how technologies are perceived, used, and eventually regulated.

Introduction to STS Framework

The Social Construction of Technology (SCOT) framework proposes that technologies do not develop in isolation and are instead shaped by the values, needs, and interactions of different social groups (Bijker, Hughes, & Pinch, 1987). A central idea in SCOT is interpretive flexibility, which means that people understand and use the same technology in different ways depending on their perspectives and goals. These varying interpretations influence how technologies are designed, adopted, and regulated. Once relevant groups reach a shared understanding, a technology can achieve closure or stabilization, while it is in the process of becoming widely accepted and integrated into society. SCOT is particularly relevant to the study

of deepfakes, as this technology is seen differently by legal professionals, social media companies, governments, and the public.

Links between STS Framework and Topic

SCOT helps to explain why deepfakes remain controversial. Social media platforms may view deepfakes as a content management challenge, while lawmakers see them as a threat to democratic processes. Courts struggle with their impact on the reliability of evidence, and banks consider them serious security risks. These different perspectives reflect interpretive flexibility and help explain the lack of a unified approach. SCOT provides a useful way to analyze how society is currently debating the role of deepfakes and how these discussions influence laws, platform policies, and public trust. As concern about deepfakes continues to grow, SCOT can help track whether or not consensus emerges over how to regulate, detect, or accept them.

Background Context

Deepfake technology emerged publicly around 2017, when internet users began swapping celebrity faces into unauthorized videos (Encyclopædia Britannica, 2025). These early examples attracted attention for their shock value, but they marked only the beginning of a much broader and more sophisticated wave of media manipulation. The underlying engine of this innovation is the generative adversarial network (GAN). GANs consist of two AI models: a generator that produces fake content and a discriminator that evaluates its realism (Preeti, Kumar, & Sharma, 2023). These systems iterate back and forth until the output is virtually indistinguishable from authentic material. What once required significant technical expertise is now accessible to nearly anyone. Tools that allow users to replicate voices, faces, or entire performances can be purchased or downloaded freely, even through open source tools, making it

possible to create a convincing deepfake with only minutes of video or seconds of audio. For example, a popular deepfake audio tool, ElevenLabs, only takes fifteen seconds to produce a realistic human sounding audio.

Proliferation has been rapid. According to Surfshark, UK government sources estimated that approximately 500,000 deepfake videos were shared online in 2023, with projections suggesting the total could double every six months as the technology rapidly proliferates (Surfshark, 2024). As the technology's scale became evident, regulators and industry actors began to respond. Texas passed SB 751 in 2019, making it a criminal offense to create a deceptive deepfake video with intent to influence an election, while California enacted AB 602 the same year, giving victims of explicit synthetic media a private right of action (Texas Legislature Online, 2019; California Legislative Information, 2019). Meanwhile, the first wave of deepfake-detection startups launched, signaling that both policymakers and markets now viewed synthetic media as a pressing social and economic issue.

Analysis by STS Framework

The following section applies the SCOT framework to four groups in which deepfakes are already reshaping trust. These groups are social-media platforms, financial services, legal systems, and the general public. By treating each group as a case study, then comparing them in a synthesis, the analysis shows how interpretive flexibility produces contrasting meanings and different responses to the same technology. Where relevant groups begin to converge on a shared meaning, the path toward closure and stabilization becomes visible. However, where meanings continue to diverge, trust remains unsettled.

Early in the 2022 Russian invasion of Ukraine, a fabricated video of President Zelenskyy urging his forces to surrender raced across Facebook, Twitter, and Youtube, reaching millions of viewers before fact-checkers debunked it (Business Insider, 2022). Inside social media platforms, the incident was framed as a content-moderation crisis that threatened brand safety and advertiser confidence. Meta's response was a policy promise to label AI media by embedding metadata, but the company conceded it could not yet detect every deepfake posted through third-party tools (Sullivan, 2024). Policymakers, by contrast, regarded the same incident as a direct assault on democratic discourse. Texas, California, and several EU governments have since drafted or enacted statutes that criminalize election-related deepfakes in the weeks leading up to a vote (Texas Legislature Online, 2019; National Conference of State Legislatures [NCSL], 2024). Ordinary users interpret political deepfakes along partisan lines, and surveys show supporters of the targeted politician are more likely to recognize manipulation, while opponents treat the clip as evidence of scandal (Vaccari & Chadwick 2020). Moving toward closure will require platforms, lawmakers, and civil-society groups to agree on a common baseline for authenticity, perhaps through interoperable watermark standards, harmonized election laws, and joint rapid-response teams that can flag suspect content before it spreads. Until such coordination is in place, deepfakes will continue to exploit gaps between these groups.

In February 2024, a Hong Kong employee joined what appeared to be a routine video meeting with colleagues and the chief financial officer. Every face and voice on the call, except his own, was AI-generated, yet the employee still wired US$25 million to criminals (Magromo, 2024; U.S. Department of Homeland Security, Analytic Exchange Program, 2020). For years, banks had treated voiceprints as an easy and reliable way to verify a customer's identity. By 2023, almost half of tier-one global banks used voice biometrics in call-centers (Kharvi, 2024).

After a handful of costly scams, the industry quickly redefined the purpose of voice authentication from a way to secure identity to merely one factor among many. Major institutions now layer behavioral and contextual checks, such as typing cadence and device fingerprinting, on top of voiceprints, and several insurers have started discounting premiums when multi-factor controls are in place (Versasec, 2024; Arya.ai, 2024). Because financial losses are quantifiable and reputational risk is immediate, high-power actors in the financial sector have pushed the system closer to closure, showing SCOT's prediction that stabilization accelerates when dominant groups agree on the stakes.

Deepfake evidence is forcing judges to revisit long-standing rules of admissibility. In Maryland, a viral audio clip that appeared to capture a school principal using racist slurs generated public outrage and disciplinary action. Forensic analysts later proved the recording had been synthetically assembled by the school's athletic director, who faced termination proceedings (Nguyen, 2025). U.S. courts still rely on Rule 901, which allows authentication through witness testimony, yet judges increasingly acknowledge that testimony alone cannot verify AI media. Legal scholars therefore recommend special pre-trial hearings in which judges weigh expert forensic analysis before admitting disputed digital evidence, and California has already directed its Judicial Council to develop concrete guidelines for AI evidence by 2026 (National Conference of State Legislatures [NCSL], 2024). Competing proposals reveal ongoing interpretive flexibility inside the legal community because some judges view deepfakes as manageable with expert testimony, while others call for wholesale revision of evidentiary rules.

The general public may be the largest stakeholder group, but it often has the least direct influence over how deepfakes are governed. Recent surveys show a mix of curiosity and concern. Deloitte's 2024 Connected Consumer Study reports that 68 percent of respondents who

are familiar with AI worry that synthetic media could deceive them, and 59 percent acknowledge they usually cannot distinguish real from fake content (Deloitte, 2024). Large-scale user data echo these anxieties. Downloads of face-swap apps such as Zao surged to the top of the free download chart in China's iOS store within days of release (Huang & Murphy, 2019)." Academic work also reinforces this pattern. Vaccari and Chadwick (2020) found that viewers' judgments of political deepfakes track their existing partisan commitments, and Kharvi's 2024 study shows that many people believe others are more susceptible to deepfakes than they are themselves. SCOT helps explain why this outlook matters. Because interpretive flexibility remains high at the mass-user level, there is no shared baseline of media literacy that would allow the public to participate fully in negotiations over regulation or platform policy. Until users converge on common norms like routinely checking origin labels or using browser-based detection plug-ins, the technology cannot reach social closure. The public's uneven understanding prolongs societal uncertainty and slows the stabilisation of deepfakes within an information ecosystem.

Across different social groups, deepfakes erode three forms of trust that are intertwined: informational, transactional, and institutional. Social-media firms and regulators wield high structural power, but diverge sharply on solutions, leaving the platform sphere in low consensus. Banks combine high power with quick convergence on new security layers. Their domain is therefore approaching stabilization. Courts possess medium power and face rising caseloads, but lack unified procedures, placing them in a transitional state. The general public has low power and low consensus, which keeps overall societal closure out of reach.

Technical coalitions such as the Coalition for Content Provenance and Authenticity aim to attach tamper-evident metadata to new media files, with pilot adoption by Adobe, Microsoft,

and the BBC, and projected mainstream rollout by 2026 (Deloitte 2025). Legislative momentum is also building: the U.S. Deepfake Accountability Act and the EU AI Act both require clear labelling of synthetic media. If these technical and regulatory tracks mature in parallel, SCOT would predict a plausible path to closure in which authenticated content becomes the norm and untagged deepfakes are rapidly quarantined. A pessimistic scenario, however, envisions an arms race in which generative models continually outrun detectors, public literacy lags, and interpretive flexibility never narrows. Which path prevails will depend on how quickly high-power actors align their policies with public education initiatives.

The Coalition for Content Provenance and Authenticity is already testing metadata "watermarks" with partners like Adobe, Microsoft, and the BBC, and hopes to make them standard by 2026 (Deloitte, 2025). Legislators are headed the same way: the proposed US Deepfake Accountability Act and the EU AI Act both call for clear labels on AI-generated media. If these technological and legal fixes move forward together, the SCOT lens suggests deepfakes could settle into a stable norm in which verified files are trusted and unmarked ones get sidelined. If progress stalls, creators may keep outrunning detection tools, and deepfakes will stay an open problem. The outcome depends on how well platforms, governments, and educators coordinate their efforts.

Discussion

The analysis makes clear that deepfakes have fractured the social expectations that once sustained trust in online media. Social-media companies mostly frame the problem as a content-moderation and brand-safety headache, whereas banks see an existential threat to authentication and revenue. Courts occupy an uneasy middle ground, torn between long-standing evidentiary rules and growing pressure for stricter verification. Ordinary users, lacking the power

or the tools of these institutions, face a confusing stream of content that may or may not be genuine.

One lesson that emerges is that power drives the pace of change: institutions that suffer direct financial or legal losses move first. Banks have already begun to replace voice-only log-ins with context-aware checks, and several courts are piloting hearings that vet AI evidence before trial. Platforms, whose losses are mostly reputational, have responded more slowly, patching policies rather than overhauling them. A second lesson is that labeling requirements on their own will not close the trust gap. Watermarks and "AI-generated" tags will help, but only if citizens can interpret those signals and detection misses remain rare. Media-literacy programs must therefore roll out alongside technical standards, with special attention to communities that research shows are more vulnerable to visual misinformation. A third insight is that interpretive flexibility can become an asset rather than an obstacle. Artists and educators already use deepfake tools for parody, satire, and classroom demonstrations. By highlighting how synthetic video is made, such projects may teach critical viewing skills more effectively than fear-based warnings.

With these findings in mind, several policy and ethical steps follow. Regulators should implement a multi-factor proof of origin that pairs cryptographic watermarks with platform-level origin checks, and content that fails these tests should be looked into, lose algorithmic reach, or prompt a warning before sharing. Courts could adapt existing rules of evidence to mandate early reliability hearings whenever audio or video is contested, while governments could fund independent forensic labs so that smaller litigants can challenge suspect files. Annual public audits of each major platform's detection accuracy and response times would create market pressure to keep false-negative rates low and invest continuously in better safeguards. A portion

of any future AI-licensing fees could support libraries, schools, and community groups that provide hands-on verification training. Finally, researchers and vendors should hold regular "red-team" exchanges, sharing attack and defense techniques, so improvements arrive before adversaries exploit new gaps.

If these measures move forward together, verified media could become the norm within a few years. If cooperation stalls, generative tools will keep outrunning detectors, public skepticism will continue to grow, and attackers will pivot to smaller and more vulnerable targets such as local courts, small banks, and community newsrooms where defenses are weakest.

Main Points and Arguments

Deepfakes erode trust because platforms, courts, banks, and everyday users still read the threat in very different ways. Some analysts could say synthetic media is simply the next cybersecurity hurdle that society will tame in time, but routine threats rarely drain millions from company accounts or distort courtroom evidence. Industry advocates could argue that an "AI generated" label is enough, yet labels protect only when judges accept them, platforms enforce them, and citizens know how to read them. Critics could warn that stronger measures such as multi-factor log-ins or pre-trial reliability hearings could burden small institutions, but publicly funded forensic labs and shared detection services can reduce those costs. Technologists debate whether detectors or forgers will stay ahead, but the SCOT view shows that social coordination matters more. Common provenance standards, platform audits, and broad media literacy campaigns could make verified content the norm, while small unsystematic fixes risk letting deepfakes grow out of control.

Conclusion

This paper applied the SCOT framework to show how deepfakes erode trust through clashing interpretations held by platforms, banks, courts, and the public. Case studies found that groups hit by immediate financial or legal harm move fastest, while others lag and leave society exposed. The analysis points to multi-factor provenance checks, early reliability hearings, public audits, and wide-reach media literacy as the core measures for stabilizing trust. If social groups coordinate around these steps, authenticated content can become the norm and confidence can rebound. If they do not, deepfakes will keep spreading and public doubt will only grow.

References

Arya.ai. (2024, April 15). A comprehensive guide to AI-powered identity verification. https://arya.ai/blog/guide-to-ai-identity-verification

Bijker, W. E., Hughes, T. P., & Pinch, T. (1987). The Social construction of technological systems: new directions in the sociology and history of technology. MIT Press.

California Legislative Information. (2019). Assembly Bill No. 602, Chapter 493. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602

Deceptive audio or Visual Media (“deepfakes”) 2024 legislation. NCSL. (2024, November 22). https://www.ncsl.org/technology-and-communication/deceptive-audio-or-visual-media-deepfakes-2024-legislation

Deepfake statistics 2025: How frequently are celebrities targeted?. Surfshark. (2025, April 15). https://surfshark.com/research/study/deepfake-statistics

Deepfakes in the courtroom: Problems and Solutions. Illinois State Bar Association. (2025, March). https://www.isba.org/sections/ai/newsletter/2025/03/deepfakesinthecourtroomproblemsandsolutions

Deloitte. (2025, January 30). Deepfake Disruption: A cybersecurity-scale challenge and its far-reaching consequences. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/gen-ai-trust-standards.html

Deloitte. (2025, March 26). Earning trust as gen ai takes hold: 2024 connected consumer survey. Deloitte Insights. https://www2.deloitte.com/us/en/insights/industry/telecommunications/connectivity-mobile-trends-survey.html

Encyclopædia Britannica, inc. (2025, April 29). Deepfake. Encyclopædia Britannica. https://www.britannica.com/technology/deepfake

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. Cyberpsychology, Behavior, and Social Networking, 24(3), 149–152. https://doi.org/10.1089/cyber.2021.29208.jth

Huang, Z., & Murphy, C. (2019, September 2). China's red-hot face-swapping app provokes privacy concern. Los Angeles Times. https://www.latimes.com/business/story/2019-09-02/china-viral-face-swapping-app-zao-provokes-privacy-concern

Kharvi, P. L. (2024). Understanding the impact of AI-generated deepfakes on public opinion, political discourse, and personal security in social media. IEEE Security & Privacy, 22(4), 115–122. https://doi.org/10.1109/msec.2024.3405963

Magramo, K. (2024, May 17). British engineering giant Arup revealed as $25 million deepfake scam victim | CNN business. CNN. https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk/index.html

Nguyen, T. (2025, April 26). Baltimore High School athletic director used AI to create fake racist recording of principal, authorities say. USA Today. https://www.usatoday.com/story/news/nation/2024/04/25/baltimore-maryland-artificial-intelligence-teacher-deep-fake/73460123007/

Preeti, P., Kumar, M., & Sharma, H. K. (2023). A gan-based model of deepfake detection in social media. Procedia Computer Science, 218, 2153–2162. https://doi.org/10.1016/j.procs.2023.01.191

Press-Reynolds, K. (2022, March 17). Zelenskyy denies deepfake video of him surrendering after hackers broadcast it on Ukrainian TV Website. Business Insider. https://www.businessinsider.com/volodymyr-zelenskyy-deepfake-video-surrendering-ukraine-russia-forces-2022-3

Sullivan, M. (2024, February 6). Meta's plan to fight AI deepfakes is too little, too late. https://www.fastcompany.com/91024531/metas-plan-to-fight-ai-generated-misinformation-and-deepfakes-is-too-little-too-late

TechPark, A. (2024, September 23). The role of social media platforms in combating deepfakes. https://ai-techpark.com/role-of-social-media-platforms-in-combating-deepfakes/

Texas Legislature Online. (2019). Senate Bill 751, 86(R) – Enrolled version. https://capitol.texas.gov/tlodocs/86R/billtext/html/SB00751F.htm

U.S. Department of Homeland Security, Analytic Exchange Program. (2020). Increasing threats of deepfake identities. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

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(1). https://doi.org/10.1177/2056305120903408

Versasec. (2024, May 10). Lowering insurance premiums with MFA.

https://versasec.com/blog/lowering-insurance-premiums-with-mfa/