Jim Craigmyle via Getty Images

Navigating Deepfakes in Litigation, Arbitration, and Mediation

Matthew F. Ferraro, Andrew Avsec, Ashley R. Riveira, and Megan Michaels

In the climactic scene of the 1992 legal comedy *My Cousin Vinny*, the voluble Brooklyn defense attorney Vinny Gambini (played by Joe Pesci) introduces into evidence a snapshot of the tire tracks left by the alleged murderers' car as it fled the crime scene. Gambini's tart-tongued fiancée Mona Lisa Vito (Marisa Tomei), testifying as an automotive expert, quickly deduces that the tire marks visible in the photo could not have been made by the defendants' "metallic mint green" Buick, but only by a 1963 Pontiac Tempest with an independent rear suspension and positraction. The local Alabama sheriff then reveals that two men fitting the defendants' description had been recently arrested driving a stolen Pontiac Tempest and carrying a gun matching the murder weapon. In the face of this overwhelming exculpatory evidence, the district attorney moves to drop the charges to whoops and cheers in the courtroom. Gambini wins two innocent men their freedom and, literally, drives off into the sunset with Vito.

No one in the movie thought to question whether the photograph on which the defendants' liberty hinged was a forgery generated by artificial intelligence (AI). But if Hollywood were to produce a remake, contemporary jurors might not believe their own eyes.

Today, affordable, easy-to-use, and highly advanced AI technologies grant nearly anyone the ability to produce or alter realistic images, audio, and videos, known as deepfakes.

Increasingly, courts are wrestling with the implications of these on-demand forgeries. AI-altered media could be admitted as genuine evidence, and fact finders may doubt the reliability of all media evidence, dismissing even legitimate images or audio as fabricated. Similar concerns also permeate alternative dispute resolution (ADR) proceedings, such as mediations and arbitrations, where evidentiary rules tend to be laxer than in litigation. Practitioners should prepare to navigate this new world.

What Is a Deepfake?

The term "deepfake" combines two words: "deep learning" and "fake." Deep learning is a subset of machine learning (ML), a branch of AI that aims to mimic how the human brain processes large volumes of data.

As the author Nina Schick defines them, deepfakes refer to images, audio, or video "either manipulated or *wholly generated* by AI" when the media is "used maliciously as disinformation." While definitions of deepfakes vary, here we exempt AI-generated media employed for salutary purposes—such as an AI-generated voice used by someone who has lost their ability to speak—or small-scale alterations in media common in traditional Adobe Photoshop editing and Instagram filters.

Several different tools and techniques can be used to generate deepfakes. Among the most common are "diffusion models," the technology behind such popular applications as OpenAI's Sora video generation tool and image generators, including Midjourney, Dall-E, and Stable Diffusion. A diffusion model generates media through a two-step process. First, in a "forward process," the model takes clean data (such as an image) and gradually adds noise to it until the original data becomes unrecognizable. Second, during a "reverse process," the model progressively removes the noise, reconstructing the original, clean data. After the model goes through this kind of training process repeatedly, it can generate new, high-quality data by starting from a random noise sample and iteratively removing the noise until clear media results.

These models, now accessible to essentially anyone with a smartphone and an Internet connection, have led to the mass propagation of deepfakes and a lower quality variant colorfully termed "AI slop." AI-generated content now fills social media feeds and superpowers fraudulent impersonations and intimate harms.

The growing prevalence of deepfakes poses social and legal harms by convincing viewers and hearers that fake media is real and by raising doubts that genuine media is fake. Professors Bobby Chesney and Danielle Citron labeled the latter phenomenon the "liar's dividend," whereby individuals can successfully deny the validity of media by claiming that the content has been altered by AI.

There exist effectively two technical countermeasures to deepfakes. First, technologies exist to detect deepfakes *after* they are created. For example, some websites offer users the ability to upload media to be analyzed for AI-generated forgeries, while other tools assess videos and still images and provide a confidence score on whether the media has been modified by AI.

The second method verifies photographs and other media at the "point of capture" in such a way that they cannot be altered after the fact without leaving evidence of the manipulation through what is known as provenance metadata or content credentials.

How Do Deepfakes Impact Litigation?

Deepfakes can impact evidentiary proceedings before courts, arbitrators, and mediators in primarily two ways.

First, a party can proffer AI-generated or manipulated content as genuine, untainted content. The fabricated content can then mislead factfinders about what actually occurred.

For example, in September 2025, the California Superior Court dismissed a pro se plaintiff's lawsuit for filing, in support of their motion for summary judgment, AI-generated videos purportedly of witnesses, photographs supposedly taken by a security camera, and text threads made to look like they were direct messages from social media. The court held that the plaintiffs violated California Civil Procedure Code § 128.7(b), which requires parties to certify that their submissions are backed by evidence and are not presented for improper purposes, by submitting "fabricated evidence." The court dismissed the case with prejudice. "[T]he use of deepfakes in a case significantly undermines the Court's ability to administer justice, significantly erodes the public's confidence in the judicial system, and significantly burdens under-resourced and overworked courts with the time-consuming task of assessing whether evidence presented to it during pretrial proceedings was a deepfake," the court wrote.

Second, a party could challenge genuine evidence, or jurors could themselves be skeptical of the veracity of media evidence, even after the evidence has been deemed admissible by a judge.

For example, in 2023, an electric car company argued in court that public statements made by its CEO regarding the safety and capabilities of one of its autonomous-driving features could be deepfakes, so the court should not compel the CEO to sit for a deposition about the statements. The judge rejected this argument, calling it "deeply troubling" because it would allow high-profile executives to disavow any public utterance.

Rules of evidence in federal and state courts establish the standards for the authentication of multimedia evidence and provide means through which litigants can challenge the admissibility of evidence that may have been fabricated by AI. Also, the American Bar Association (ABA) Model Rules of Professional Conduct bar attorneys from knowingly offering false evidence, such as a deepfake, and from challenging evidence without a good-faith basis. The Advisory Committee on Evidence for the Federal Rules of Evidence is considering modifying the federal rules to require a showing before a court can undertake an inquiry into whether evidence has been manipulated by AI.

Dealing with deepfakes in mediations and arbitration will require a different approach, where the rules around evidence are more varied and less formalized.

The Evidentiary Landscape of Mediation and Arbitration

Understanding the disruptive potential of deepfakes in ADR requires a clear grasp of the evidentiary standards governing its two principal forums: arbitration and mediation.

Arbitration: A Semiformal Evidentiary Approach

In arbitration, two or more parties enter into a legal contract—an arbitration agreement or clause—to resolve their dispute outside of court with the help of an arbitrator or a tribunal. Arbitration operates as a quasi-judicial proceeding where evidentiary standards are less rigid than in court litigation.

The arbitration agreement typically selects procedural rules, which in turn often only contain broad statements confirming the arbitrator's powers to assess evidence and weigh its relevance and materiality. On rare occasions, parties may in the arbitration agreement itself specify their adherence to an enumerated set of evidential rules or guidelines, including the International Bar Association Rules on the Taking of Evidence (IBA Evidentiary Rules).

Otherwise, an arbitrator, in consultation with the parties, will define the general procedural rules, including standards of evidence, in an initial procedural order, typically referred to as Procedural Order No. 1 (PO1). Often, PO1 delegates to the tribunal powers to determine which evidence to rely on rather than establishing particular evidentiary rules. For example, PO1 often establishes that a tribunal "shall" or "may" consider the IBA Evidence Rules, which state that an arbitrator "shall determine the admissibility, relevance, materiality and weight of evidence." The American Arbitration Association (AAA) Commercial Arbitration Rules, Rule 35, grants similarly broad powers to the arbitrator to admit relevant evidence or exclude evidence deemed to be "cumulative or irrelevant."

While arbitrators are not strictly bound by any particular standard, they typically apply principles of materiality and relevance when determining admissibility. The arbitral process is adjudicative, relying on documentary evidence, witness statements, and expert testimony presented to a neutral decision-maker whose rulings are generally binding. As noted, under commonly accepted standards, arbitrators have broad discretion to admit evidence, including hearsay and less conventional documentation, to facilitate efficient resolution.

Mediation: Informal Evidence and Psychological Impact

Mediation, in contrast, is a facilitated negotiation focused on reaching a voluntary settlement. Formal evidentiary rules are virtually absent. Evidence exchanged in mediation is not intended to prove legal facts, but rather to influence negotiation postures and shape perceptions of risk or potential outcomes. Lawyers may use a telling document or a "smoking gun" email in private conference to bolster their bargaining leverage, for the power of evidence derives primarily from its psychological impact on the parties' willingness to settle.

Deepfake Risks in Arbitration

The spread of deepfakes may undermine the efficiency and trustworthiness of arbitration. This risk may arise not only with respect to documentary or video evidence but also with regard to the authenticity of any remote interactions during a hearing (such as the cross-examination of witnesses and experts via videoconference or the legitimacy of all participants in virtual hearings).

In response to rising concerns about the use of AI in arbitration in general, several arbitral organizations have published guidelines for the use of AI in arbitration. The Silicon Valley Arbitration and Mediation Center (SVAMC) Guidelines on the Use of Artificial Intelligence in Arbitration state that "[p]arties, party representatives and experts shall not use any form of AI to falsify evidence, compromise the authenticity of evidence, or otherwise mislead the arbitral tribunal and/or opposing party(ies)." It suggests that all parties are responsible for any uncorrected errors or inaccuracies in the output of AI systems. Additionally, these Guidelines recommend that, if any arbitral tribunal determines that a party has violated them, it may consider "striking the evidence from the record," "deeming it inadmissible," "deriving adverse inferences, and taking the infringing party representatives' conduct into account in its allocation of the costs of the arbitration."

Likewise, the Chartered Institute of Arbitrators (Ciarb) Guideline on the Use of AI in Arbitration seeks to provide guidance on the use of AI to allow for its benefits while mitigating some of AI's "risk to the integrity of the process, any party's procedural rights, and the enforceability of any ensuing award or settlement agreement."

At the core of the Guideline is an emphasis on the value of a party disclosing the use of AI to the other party and the arbitral tribunal. If a party fails to disclose the use of such a tool "in contravention to the arbitrator's direction," the arbitrator can ultimately "take any measure to remedy that failure, make any further rulings on the use of AI, draw any appropriate conclusion (including drawing adverse inferences, if appropriate), or take such failure into account when awarding costs."

Specifically, it recommends countering the risk of AI-altered evidence through the use of deep fake detection tools to ascertain if evidence has been "fabricated."

The Ciarb Guideline suggests parties adopt an "agreement on the use of AI in arbitration," a model of which is included in the Guideline, under which the parties agree to refrain from using an AI tool to produce an output that may "mislead" the tribunal, including through the "fabrication or falsification of any evidence[.]"

Lastly, the American Arbitration Association-International Centre for Dispute Resolution (AAA-ICDR) has published AAAi Standards for AI in Alternative Dispute Resolution, which address the potential for AI-generated inaccuracies by requiring both arbitrators (or neutrals) and advocates to verify AI outputs against "trusted source materials," including evidence, party filings, and legal authorities. To ensure the integrity of the process, neutrals must prioritize "direct review of evidence" and reasoned deliberation, ensuring that automated data analysis or document summaries

"never overshadow firsthand examination" of the case materials. Additionally, the Standards task practitioners with examining AI outputs for "misstatements, bias, or dubious references." The Standards hold advocates responsible for using human scrutiny to correct errors promptly and encourage neutrals to seek clarification whenever AI-generated information is in doubt.

Deepfakes in Mediation: Manipulation and Erosion of Trust

In mediation, deepfakes may become instruments of psychological leverage rather than legal proof. Parties may introduce fabricated videos or documents to intimidate opposing counsel—especially where reputational harm could drive advantageous settlements. For instance, a plaintiff in an employment dispute could privately present the mediator with a realistic but fake video purporting to show the defendant's misconduct, exerting settlement pressure irrespective of authenticity.

In the employment context, the past could well be prologue: In April 2025, a Maryland school employee pleaded guilty to creating and distributing what turned out to be an AI-generated audio clip that impersonated the school's principal making racist and antisemitic statements. It was only after the principal was placed on leave that it was discovered the employee created the fake as retribution for the principal investigating the employee for fraud.

Beyond extortion risk, deepfakes have the potential to erode the trust and collaborative spirit necessary for successful mediation. Suspicion that a party is using AI-generated evidence—for example, by inflating damages with AI-generated images purporting to show significant destruction to insured property—may undermine the process entirely and push parties toward more costly litigation.

The IBA's Mediation Committee recently issued Guidelines on the Use of Generative AI in Mediation, which highlight that AI-generated content may include errors, biases, and inaccuracies and require users to assess outputs independently to maintain the integrity of the mediation. But the Guidelines are silent on concerns about the use of AI to create adulterated evidence that would be reviewed in mediation.

Safeguarding the Future of ADR

The rise of deepfakes necessitates proactive steps to ensure that ADR continues to provide value to parties. While each case will be different, practitioners may wish to consider the following:

First, in all forms of ADR, practitioners should educate themselves about deepfakes and other AI-related tools that may implicate the just resolution of their matters.

Second, practitioners should discuss AI use and deepfakes in preliminary conferences with arbitrators, mediators, and opposing parties. At a minimum, parties should agree on prohibitions on deepfake evidence; certify that no deepfake evidence is proffered; and disclose the extent to which AI has been used, including providing "before and after" versions of any media that AI has been used to process or touch up, even if the media itself may not qualify as a "deepfake." The parties may choose a rule that the knowing use of deepfakes to mislead will be punished by, for example, the drawing of an adverse inference for the spoliation of evidence.

Third, parties should consider explicitly adopting before the initiation of the ADR process one of the AI-ADR established guidelines discussed above (from the AAA-ICDR, Ciarb, SVAMC, IBA, or another such body) or incorporating deepfake-related rules from the Federal Rules of Evidence or a state court.

Fourth, if necessary, parties should draft and incorporate "deepfake clauses" in arbitration agreements or procedural orders that stipulate how the arbitrator or mediator will address challenges to digital evidence. This clause could require challenged digital evidence to be subject to third-party verification. Or, like the proposed amendments to the Federal Rules of Evidence, the proceeding could adopt a burden-shifting process whereby a party challenging authenticity must provide a threshold showing of potential fabrication before the proponent of the evidence must prove the evidence is more likely than not authentic.

Fifth, the parties could agree to use technical verification protocols, including AI detection tools or provenance or content credentials, if relevant, to establish that evidence was not manipulated.

Finally, if warranted, arbitrators can appoint a neutral forensic expert to evaluate evidence.

In sum, ADR practitioners must remain vigilant and adapt processes to preserve fairness and integrity against the evolving risks presented by AI.

Matthew F. Ferraro, a partner at Crowell & Moring, helps clients address complex business and regulatory matters at the intersection of advanced technology, national security, and crisis management. Previously the Senior Counselor for Cybersecurity and Emerging Technology to the Secretary of Homeland Security, he has, in private practice, counseled clients on the legal and business issues surrounding deepfakes since 2019.

Andrew Avsec, a partner at Crowell & Moring, advises clients ranging from start-ups to Fortune 100 companies on complex brand protection, right of publicity, copyright, advertising, trade secrets, AI, and social media issues. He has represented some of the world's most prominent brands in litigation, Trademark Trial and Appeal Board (TTAB) oppositions and cancellation actions, arbitration proceedings, and domain name disputes under the Uniform Domain-Name Dispute-Resolution Policy.

Ashley R. Riveira, a counsel at Crowell & Moring, is a member of the firm's International Dispute Resolution Group. She is a trusted advisor to global clients navigating complex cross-border business disputes. Renowned for her deep experience in international arbitration and litigation, Riveira's clients value her ability to craft tailored solutions for intricate legal challenges in international contexts.

Megan Michaels, a counsel at Crowell & Moring, is a member of the firm's Advertising and Brand Protection Group. She offers comprehensive legal services to clients seeking to protect their intellectual property assets. With a focus on trademark and copyright clearance, prosecution, and enforcement, she also has experience handling domain name disputes and Trademark Trial and Appeal Board (TTAB) proceedings. Her practice focuses on collaborating with clients to safeguard their brands and creative works.