Whincop Lecture October 2026 – Fleur Kingham

When seeing is no longer believing: How AI compromises evidence law and how we might respond

ACKNOWLEDGEMENTS

I acknowledge the Turrbul and Jagera peoples, the traditional custodians of the lands and waters we are meeting on and near, and I pay my sincere respects to their elders. I recognise that our laws, including those governing evidence, struggle to embrace traditional methods of preserving and transmitting knowledge, memory, and truth.

I thank Griffith University for the honour of presenting this lecture. In our brief acquaintanceship, Michael left a lasting impression of a young man driven by intellectual curiosity. I am sure he would have formed strong views about how to decide what to trust in this age of AI.

INTRODUCTION

When asked at a dinner party how he felt about dispensing justice, Sir Owen Dixon replied:

"I do not have anything to do with justice, madam. I sit on a court of appeal, where none of the facts is known. One third of the facts are excluded by normal frailty and memory; one third by the negligence of the profession; and the remaining third by the archaic laws of evidence."

My topic concerns those archaic laws of evidence and how well they serve justice in the age of artificial intelligence. This is a live topic for the Qld Law Reform Commission, I acknowledge with gratitude the research assistance provided by Secretariat staff and student interns. However, what I present to you does not reflect any view formed by the Commission, which has not yet formulated its approach to consultation on the topic.

AI poses many challenges to the administration of justice. I want to use a narrow lens and focus on one: deepfake imagery.

You be the judge. A video is tendered before you. It shows one party saying precisely the words on which the case turns.

The party tendering it swears it is genuine. The other party swears: “I never said that.”

Neither side has a digital forensic expert.

What do you do?

Our current rules would permit admission once a sufficient foundation has been laid. The ultimate question, whether it is genuine, would be left to the factfinder, judge or jury, when assessing the evidence.

But AI changes the factual premise against which those rules developed. A convincing audiovisual record can now be manufactured cheaply and quickly. What evidence rules should apply?

Winston Churchill is credited with saying: “The longer you can look back, the farther you can look forward.” So that is where I intend to start.

Looking back before there was visual evidence shifts our focus from the technology to the purpose of evidence law. AI may challenge, but it does not change that purpose.

HOW DID WE DECIDE WHAT TO BELIEVE?

What faith revealed

Before we had rules of evidence in anything like their modern form, proof could be a matter of faith, guided by the supernatural or divine. Truth was what faith revealed.

The medieval trial by ordeal, practised in Anglo-Saxon England and elsewhere, drew on older traditions of supernatural judgment. The gods’ revelation of hidden truth became the judicium Dei – the judgment of God.

In effect, people outsourced judgment on serious allegations to an all-knowing God and divine decree. The guilty hand became infected after carrying hot iron; the guilty person floated in blessed water.

Pope Innocent III helped end that system in 1215 when the Fourth Lateran Council prohibited clerical participation in ordeals. Tempting an uncontrollable God was no longer acceptable.

But trial by battle, which rested on a similar premise, endured. After the Norman Conquest, it was a principal means of resolving land disputes. The contest was physical, but the theory was divine: God would favour the rightful party.

These procedures seem alien now, but, like our laws today, they rested on assumptions about what can reliably reveal truth.

What we knew

Henry II's reforms in the Grand Assize of 1179 offered another method, an alternative to trial by battle in property disputes, which helped lay institutional foundations for the modern jury.

This marked a transition from what faith revealed to what we, or at least the jurors, knew. These early jurors were not passive recipients of evidence. They were selected because of what they already knew. That gradually changed, and by the eighteenth century, jurors were increasingly reaching verdicts based on information presented in court.

That shift created a problem that still defines evidence law: if the factfinder must decide from material presented in court, what material should be allowed to shape belief?

Faith remained central. Unsworn testimony could not be considered. Oaths, sworn by the right people in the right order, determined disputes tried by wager of law. The power of the oath was faith; the belief that a person would speak truly because they feared their immortal soul would be imperilled if they swore falsely.

That faith has withered. As one historian put it: "In our era … the fires of hell have smoldered. Oaths have lost their terror and even their meaning."

What we heard

Central to the adversarial trial is another belief: that cross-examination can expose error, inconsistency, and deception. This explains the primacy of first-hand accounts in the rule against hearsay. This rule of evidence and its many exceptions are not evidence-based. They assign evidential significance to assumptions about human behaviour, reflecting beliefs about when an account is likely reliable, despite the lack of confrontation.

Evidence law was already doing what it still does: constructing rules around imperfect human methods of knowing.

What we saw

And then came what we saw. People created documents when they had good reason to preserve information, most often out of institutional and landowning concerns: the Crown, Church, monastery, manor, or landlord. Signatures and seals were proved by attesting witnesses. The original document was produced to verify the copy, guarding against mistranscription in handwritten copies. These evidentiary rules emerged in a world where writing was exceptional, and documentary technology was crude. For most ordinary events there was no contemporaneous documentary record.

Then something extraordinary happened. We developed a machine that appeared capable of recording reality itself.

In 1826, the first permanent photographic image was fixed in France. Within a few decades, people used photographs in court. In 1852, French lawyers were reported to be using photographs to convince judges and juries with images “more eloquent than their words”.

Photography rapidly became cheaper and more accessible. Kodak’s Brownie camera democratised still photography, which proliferated. Moving pictures followed. CCTV became commercially available. Film gave way, remarkably quickly, to digital imaging. The camera then moved into the device most of us now carry everywhere: the smartphone.

Two hundred years after the first permanent image, photographic evidence is ubiquitous, and courts pay scant attention to either admissibility or authenticity.

The image became a medium of communication and a substitute for words.

And with it came another kind of faith: faith in what we see.

WHY DO WE BELIEVE WHAT WE SEE?

In the 1933 classic film Duck Soup, Chico Marx asks: “Who you going to believe, me or your own eyes?”

It is funny because the answer seems obvious.

Our language reflects this assumption: seeing is believing; what you see is what you get; a picture is worth a thousand words.

But why do we believe what we see?

Photographic technology emerged in the age of positivism, a new belief in scientific observation as the source of authoritative knowledge. The antithesis of faith in the supernatural. The camera was presented as an objective mechanical instrument that could register reality more accurately than the fallible human eye and hand.

But photographs have never been neutral. Someone chooses what is inside the frame and what is outside it, when to press the shutter and from what angle.

War photography gives striking examples. Alexander Gardner staged the corpse of a Confederate sharpshooter in a famous image during the American Civil War. Closer to home, Australia's official war photographer for both World Wars, Frank Hurley, used composite images to convey modern battle. He constructed an image no single camera exposure had recorded to convey what he considered the truth.

So, an image does not eliminate human judgment. It just relocates it from the person behind the eye to the person behind the machine.

What you see is not simply what you get

There is another complication. Seeing is an act of interpretation. What you get is not simply what you see.

Our visual system does not seamlessly convey a complete picture of the external world into consciousness. The brain selects, integrates and interprets incomplete sensory information. It draws on attention, expectation, memory and information from other senses.

What we "see" is our brain's interpretation of information, not a direct transmission of reality.

The invisible gorilla experiment is a famous demonstration. Participants focused on counting basketball passes miss a person in a gorilla suit walking through the scene. The information reaches the eyes. Attention determines whether it becomes part of the perceived event.

Our pre-existing attitudes can also affect visual interpretation. In one study, participants with positive and negative attitudes towards police watched the same video of an altercation between an officer and a civilian. Eye-tracking showed that the groups looked at the scene differently, paid more attention to the officer or the citizen and reached different judgments about the officer's conduct, depending on those pre-existing attitudes.

So before AI enters the picture, we already have two reasons to be cautious about "seeing is believing".

The image itself reflects choices about how reality is represented. And the viewer brings a human perceptual system that selects and interprets what the image contains.

AI adds a third problem. The event represented may never have happened at all. Can we still believe what we see?

CAN WE STILL BELIEVE WHAT WE SEE?

A deepfake primer

A deepfake is a digital photo, video or audio file that presents a realistic but false depiction of a real person doing or saying something they did not do or say.

The technical methods vary and keep evolving. Deep learning systems learn features from large collections of images or recordings and generate new content with those characteristics. Generative adversarial networks use competing algorithms to improve the fake: one generates content, the other evaluates its realism, creating a feedback loop, which improves the fake.

I find that design method ironic given our adversarial trial process. But the legal point does not depend on the deepfake architecture. The important facts are that convincing synthetic media can be generated cheaply, at scale, and with rapidly improving quality.

Deepfakes promote both belief and scepticism.

Deepfakes present two dangers.

The first danger is false belief: believing something that never happened.

A convincing fabricated video may influence a legal outcome before anyone discovers it is false. A couple of US cases illustrate the risk. In the first, a woman is reported to have spent two days in jail for violating a protective order made on evidence that included AI-generated text messages allegedly fabricated by her former partner; the charges were later dropped.

In the other, a judge who had agreed to receive a pre-recorded audio-visual statement of evidence ruled it inadmissible as a deepfake after she noticed glitches in the

recording. Digital forensic experts reportedly regarded her observation as fortunate: the technology is already moving beyond many of those obvious telltale flaws.

The second danger is false disbelief: we cease believing something that did, in fact, happen.

Awareness of deepfakes can give a dishonest litigant an offensive advantage. Genuine evidence might be dismissed as synthetic simply because deepfake evidence is now plausible. This is known as the "liar's dividend".

For example, in the prosecution of a participant in the January 6 attack on the US Capitol, his lawyer argued prosecutors should be required to verify all the online video evidence because of the proliferation of deepfakes. The evidence was admitted. Anecdotally, I have heard self-represented litigants raise the same argument in Qld Magistrates Court proceedings, such as traffic infringements, with the same result.

Those cases are reassuring about admissibility. They do not eliminate the deeper problem.

If authenticity remains a question of weight, the factfinder may be told, in effect: here is a video. It may show exactly what happened. Or it may have been manufactured by a machine. You decide.

Authentic evidence may lose persuasive value simply because technology has made deception plausible.

So AI destabilizes visual evidence in both directions. It can make the false believable and the true deniable. Are our rules good enough to withstand this destabilisation?

IS EVIDENCE LAW FIT FOR PURPOSE IN THE AGE OF AI

Admission and evaluation

There are two theories about the basis for admitting visual evidence.

Under the pictorial testimony approach, the image operates as a non-verbal expression of a witness's testimony. A witness who can speak to the scene confirms the image is a fair representation. The fact finder can use the image to understand the evidence from that witness but cannot draw their own inferences from the image.

Under the silent witness approach, the image is itself real evidence without the need for a witness to speak to the events. The court must be satisfied that there is an

adequate foundation for admitting the image, although in many cases this is not contested and it is admitted as a matter of course.

The foundation for admission concerns the provenance of the original capture and, for a copy, the accuracy of the copying process, drawing on the rules governing documentary evidence. The admission threshold is relatively low. Once met, disputes about whether the evidence is genuine remain for trial.

That division works tolerably well if convincing fabrication is difficult and exceptional.

Deepfakes challenge complacency about image manipulation.

A witness can identify the apparent scene. A custodian can explain where a file was found. Metadata can support a chain of custody. But synthetic media can be constructed to mimic the very features those traditional methods rely on.

The question is whether evidence law should demand more.

HOW CAN WE IMPROVE OUR RULES?

I suggest we organise our response around the stages through which digital evidence moves in the legal process.

We should ask questions at each stage of a matter progressing through the system. What should we do before the evidence reaches court, when it enters the courtroom, and when someone is ultimately asked to consider it? That framing centres the law and its objectives, not today's technology.

1. Before trial - preserve and disclose provenance

The first stage may be outside the scope of evidence law. It concerns professional responsibility, provenance and detection tools and pre-trial procedure.

A lawyer's paramount duty is to the court. If we can no longer safely assume the authenticity of digital material supplied by a client, what does that duty require before the lawyer discloses or tenders digital material?

Should lawyers require their clients to produce original digital files and associated metadata? I expect some already do. But should this be a professional obligation? Should a party be required to provide provenance information for digital evidence? Should the obligation arise only when authenticity is put in issue? And if a party strips

or fails to preserve provenance information that would ordinarily have been available, what consequences, if any, should follow?

Disclosure can involve large volumes of digital material, only a fraction of which will ever be used at trial. Requiring forensic authentication of everything disclosed would impose enormous cost for little benefit. A more proportionate regime might require preserving provenance across discoverable material, with more intensive authentication obligations applying only to evidence intended to be relied upon or where authenticity is genuinely disputed.

That distinction between disclosure and tender matters because the procedural response should not itself become a source of unnecessary cost and delay.

Existing civil procedure may be more easily adapted for this purpose than criminal procedure, where accused persons need a practical, accessible method to inspect the original file and associated records for material captured or interpreted by police systems.

An out-of-court disclosure and inspection process is a modest response: do not wait until the video is about to be played in the courtroom before asking where it came from. If authenticity disputes can be identified, specified, and narrowed before trial, courts need not be the first forum to explore technical questions.

Technology will provide some assistance in resolving authenticity disputes.

The emerging provenance infrastructure for digital evidence will play a role. Provenance products create cryptographic records of the origin and editing history, preserving an audit trail for digital material.

But an audit trail is not a deepfake detector. It may tell us something about the device or system that created the file, about subsequent edits, and about the continuity of the record. Conversely, the absence of a manifest or watermark proves little on its own. Not every camera or platform creates the same record. Metadata can be removed. Industry standards are not universal, although there are encouraging signs that the major technology companies are coalescing around Content Credentials, an open-source industry standard.

As provenance information is routinely created and preserved upon capture, its evidential significance will mature. That is familiar evidentiary reasoning: the significance of missing evidence depends on whether, in the circumstances, we would ordinarily expect it to exist.

Deepfake detection software is more problematic. A detector may purport to identify synthetic content, creating unwarranted confidence in the assessment. Detection methodologies may be opaque. The assessment may be unreliable when a file is reformatted or when generation technology changes. Stricter regulation of AI processes can help. Although Australia is slow to act, international approaches may drive developments in designing detectable elements in AI-created and AI-enhanced products. The European Union requires AI system providers to include design features that help detect and identify AI-generated or manipulated content. Some US states have adopted transparency measures with the same objective. China's Deep Synthesis provisions go further, mandating the marking of AI-generated material.

Nevertheless, we are in a detection arms race. A method that works for one generation of deepfake technology may fail against the next.

That suggests an important principle for reform: do not build an evidentiary rule around a particular provenance tool or detection methodology.

Procedural obligations can be designed around more durable questions: where did this material come from, what happened to it, what corroborates it, and can we explain and independently test the process by which its authenticity is assessed?

2. At admission – what is the authenticity threshold?

The second stage is admission.

Currently, the rules set a relatively low authenticity threshold for admission, with the trier of fact left to decide whether to accept the image as genuine. Deepfakes invite us to ask whether the admission process should be reformed.

Three broad approaches have been proposed in the United States.

An elevated authentication requirement

The first is an elevated authentication standard applied at a preliminary hearing. The tendering party would be required to provide additional corroboration of source or provenance.

I think the utility of this process depends on the trigger. If a litigant can secure a pre-trial hearing by simply asserting the possibility of a deepfake, the procedure would institutionalize the liar's dividend with associated cost and delay. That risk commends a particularized basis for the challenge: an inconsistency in provenance, an unexplained break in custody, evidence of manipulation, an anomaly in the file, or

another circumstance capable of raising a genuine issue. Only then should the proponent be required to go further.

What remains under this model is what to do with residual doubt about authenticity. Is the ruling conclusive or can authenticity remain a question for trial? It seems the latter.

Probative value and prejudicial effect

A second proposal uses the familiar balancing language of probative value and prejudicial effect.

Under this approach, once the challenging party provides a sufficient basis to allege fabrication or alteration, the tendering party would have to demonstrate the evidence should, nevertheless be received. The legal test proposed is that its probative value is greater than the danger it may mislead or unfairly influence the factfinder.

The attraction is that this works with a familiar concept of evidence law, but I wonder whether it is the right test. The difficulty a deepfake presents is more fundamental than whether the image might exert an irrational or disproportionate influence. If the depicted event did not occur, the apparent probative force of the evidence rests on a false premise.

Again, if admitted, disputes about authenticity remain live issues for trial.

Judge as authenticity gatekeeper.

The third proposal tackles this head-on.

The judge determines authenticity at the admission stage. If the evidence is admitted, the jury is not invited to speculate that it may nevertheless be synthetic. In fact, the jury is directed that, because it has been admitted, it can treat the evidence as genuine.

That directly addresses the liar's dividend. A party cannot lose an authentication challenge before the judge and then invite the jury to distrust the same evidence merely because deepfakes exist.

There is real force in that.

But it transfers a difficult factual judgment from the jury to the judge and invokes a range of issues. Are we comfortable asking judges to determine technologically contested questions with greater confidence than the available science may justify?

What standard of proof should apply? And what happens where the evidence establishes authenticity strongly, but not conclusively?

Judges might be reluctant to tell a jury, in effect, “You may treat this image as authentic”, when authentication technology itself is probabilistic and evolving.

I have some reservations about all three options. The first two require definition of the threshold for admission or discretionary exclusion, but may leave with the factfinder, judge or jury, the more difficult question about authenticity. The third places authenticity firmly with the judge as gatekeeper.

Whether the judge should determine authenticity on admission is, I think, the hardest question.

3. At evaluation - equip the fact-finder

Whoever has the ultimate say on disputes about authenticity, the third stage is critical: how do we equip judges and juries to evaluate the evidence rationally?

There are at least three parts to that: judicial education, expert evidence and directions to juries.

Judicial education - and the danger of confidence

Judges need enough technological literacy to understand what provenance evidence can establish, what a detector can and cannot establish, and when specialist assistance is required. This is a work in progress.

The US National Center for State Courts has developed bench guidance for judges dealing with suspected unacknowledged AI-generated evidence. The National Judicial College of Australia is developing focused AI programs.

Although recent international research led by an Australian academic suggests people can be trained to better detect AI-generated material, judges are unlikely to be effective deepfake detectors.

And the research comes with an important warning about overconfidence. The more confident we are in our ability to detect synthetic images, the more prone we are to error. This resonates with similar findings about confidence in detecting a liar. We should be very cautious about replacing “I can tell when a witness is lying” with “I can tell when a video is fake”.

Judicial education should focus on the questions judges need to ask and how they might be answered.

Digital forensic expertise

And that brings me to expert evidence.

A deepfake investigation cannot sensibly be reduced to a detector returning simply "real" or "fake".

A forensic analysis of suspect material must be multi-layered, involving answers to questions such as these:

- Where did the file come from?
- Who had custody of it?
- What happened to it before it was produced?
- What metadata or provenance information survives?
- Is there evidence of editing or recompression?
- Does surrounding evidence corroborate what the media depicts?
- What does forensic analysis show?
- Can the methodology and result be explained in court?
- Can they be independently verified?

That is orthodox forensic reasoning. No single feature carries the whole conclusion. Different sources of evidence are stacked to build - or undermine - confidence in authenticity.

It also reinforces why provenance matters before trial. The best expert in the world cannot analyse information that was never preserved.

This is an emerging discipline, and experts are thin on the ground. Courts must ask familiar questions in a new context. What qualifications establish expertise? What methodology is sufficiently reliable? Has it been validated against current generative tools? Can another expert reproduce the result? How much can the expert explain about a proprietary detector whose internal operation is opaque?

Jury directions and visual literacy

Finally, if a jury must assess disputes about the authenticity of visual evidence, how will they know how to do that? Just as we guide jurors about the fallibility of human

memory and the risks in eyewitness identification, we must build safeguards around the fallibility of visual belief.

I consider this important for all forms of visual evidence given what we know about the sensory and psychological factors involved in 'seeing'.

People do not passively receive visual information. Attention, expectation and prior attitudes affect what they perceive. The police-video study I mentioned earlier is encouraging. When participants were directed to give equal attention to both the officer and the citizen, their visual orientation changed, and their judgments became less polarised. We may not be able to eliminate prior beliefs, but we may be able to change how people attend to evidence.

There is a nascent body of academic research into juror decision-making and the impact of directions. I encourage researchers to test two propositions. First, that directions can affect how jurors attend to and evaluate visual evidence. Second, that jurors evaluate disputed digital evidence more effectively when directed to consider provenance, corroboration, continuity of custody, and the limits of visual intuition. We should explore whether particular formulations reduce overconfidence, and whether warnings about deepfakes inadvertently cause jurors to distrust authentic evidence.

The objective is the same for juries as for judges, to move them from an intuitive question - "Does this look real?" - to an evidentiary question - "What are the grounds for concluding that this is real?"

4. Regulatory design - durable law for fast-moving technology

There is one final issue that sits across all three stages: regulatory design.

To maintain confidence in our justice system, we must future-proof evidence law for technological advances we cannot conceive of now. The speed of technological change makes detailed prescription in an Act unattractive. We do not face a binary choice. Everything I have discussed contributes.

Evidence Acts should focus on durable evidentiary principles. Court rules can prescribe procedural obligations to test authenticity outside the courtroom. Practice directions can respond more quickly to changing technologies, forensic practices and technical standards, which are best placed to identify effective provenance and detection methods over time.

And then there is the question of national uniformity.

Digital evidence does not change character when it crosses a State border, but we already see divergences in how courts and tribunals respond to the use of AI in submissions and expert reports.

However, uniform evidence law has developed at a glacial pace that poorly matches the rate of technological change.

While national consistency is desirable, the mechanism used to achieve it must be adaptable and amenable to evolving consensus. A national dialogue about court rules and practice directions may be our best investment in national consistency.

CONCLUSION

Two hundred years ago, photography gave us something extraordinary: a mechanical means of capturing the world.

And we came to trust it.

We knew that the camera might be positioned selectively, that the photograph might be cropped, staged or manipulated. But beneath those qualifications remained a powerful assumption: something had been there, in front of the camera, to be photographed. Now we can readily produce compelling visual evidence of an event that never occurred, involving a person who was never there, saying words they never spoke.

Looking backwards, our dispute resolution processes have always reflected changing ideas about how to reveal the truth.

We trusted divine judgment. We trusted oaths. We trusted witnesses subject to cross-examination. We trusted documents. And eventually we placed a particular trust in the apparently objective eye of the camera. None of those forms of proof was infallible.

The history of evidence law can be understood as a centuries-long attempt to regulate how human belief forms in the courtroom. Looking forward, let us resist the temptation to look for a technological fix. Ultimately, we must rely on human judgment.

It is ironic that AI has exposed the fragility of “seeing is believing”. But in doing so, it may improve the way we think about all visual evidence.

It reminds us that seeing has never simply been believing. Seeing is perceiving, interpreting, contextualising and judging.

And that is precisely what the law of evidence has been trying to regulate for centuries.

Sir Owen Dixon may have been right to describe the laws of evidence as archaic.

But archaic does not necessarily mean obsolete.

The purpose of the laws of evidence remains as compelling an objective today as it did when the modern jury system emerged: to help imperfect human beings, using imperfect information and imperfect methods, decide what they have sufficient reason to believe.

AI does not challenge that purpose.

It simply makes the task more urgent.