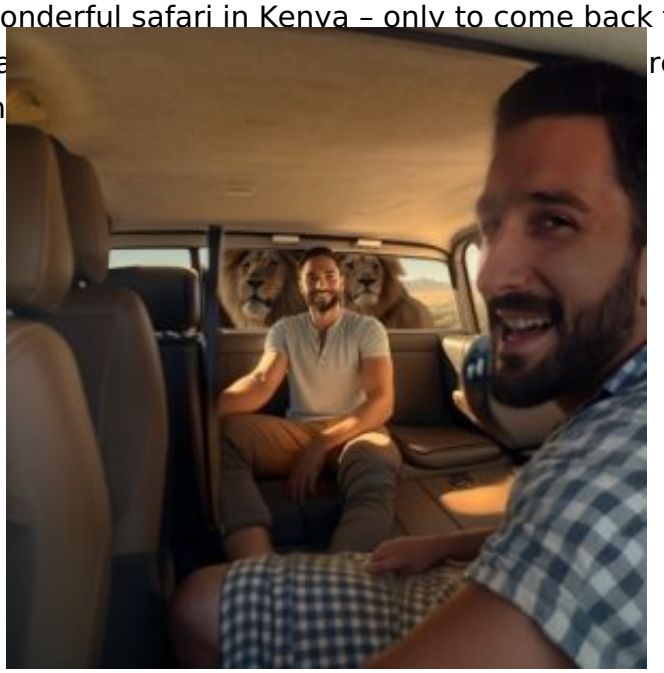
AI-generated evidence is on the horizon – but are we ready?

It has been a manic few weeks for me; busy work schedule aside, I have been lucky enough to recently squeeze in a hiking trip to Hawaii, a wonderful safari in Kenya – only to come back to win

Plea



res of d m

Figure 1 – My trek round Hawaii

Figure 2 – running into lions in Kenya

Figure 3 – me and my child client

A great few weeks all round..!

Well, I ought to come clean: I have never been to Hawaii, nor to Kenya. Similarly, the *'Best Children Lawyer in the World'* award is sadly nothing more than a figment of my imagination – and before anyone alerts the SRA, the smiling child in the image above is not my client. Indeed, that child doesn't even exist. The above images have all been generated using Artificial Intelligence (or AI), using publicly available platforms – and the potential of these platforms cannot be overstated.

What are AI-generated images?

An AI-generated image is simply an image that is created using artificial intelligence and machine learning. The technology draws up a computer-generated image based on a few lines of text that a user submits. Users can sign up to any one of a number of publicly available platforms for this – with the most prominent being Dall-E and Midjourney – and users can generate images that exist only digitally and do not, or cannot, exist in reality. With every use and every entry, the platform becomes cleverer, by *'learning'* from the user's input – meaning the results generated become more sophisticated, more suitable and more accurate with every click.

The potential is, quite frankly, limitless. Say, for example, you work for a high street supermarket,

and want to advertise Victoria sponge cake. In times of yore, you might commission an expensive advertising agency and design team, send a few cakes off to them with a vague brief to make them look alluring, and hope for the best. Now, you may consider shortcutting this all by entering the terms *'draw up a hyper-realistic photograph of a Victoria sponge, sliced, served at teatime'.* This would be the sort of response you would receive, 30-odd seconds later:

Figure 4 – not a real cake

The pictures above are not ones of a real cake; it is based on what the platform *'thinks'* a Victoria sponge should look like. Again – this cake is not real, as delicious as it looks. The platform generates the image based on similar images online, and the more specific you get, the more specific the image becomes. As shown above, the results can be spectacular, often impossible to discern from reality, and are only getting better.

And this domain is not just reserved for cake, or indeed for reality. Users can brief the generator to come up with weird and wonderful scenarios, grounded either in real life or in ludicrous fantasy.

The platform makes it possible for lawyers to be depicted as golden retrievers, for example:

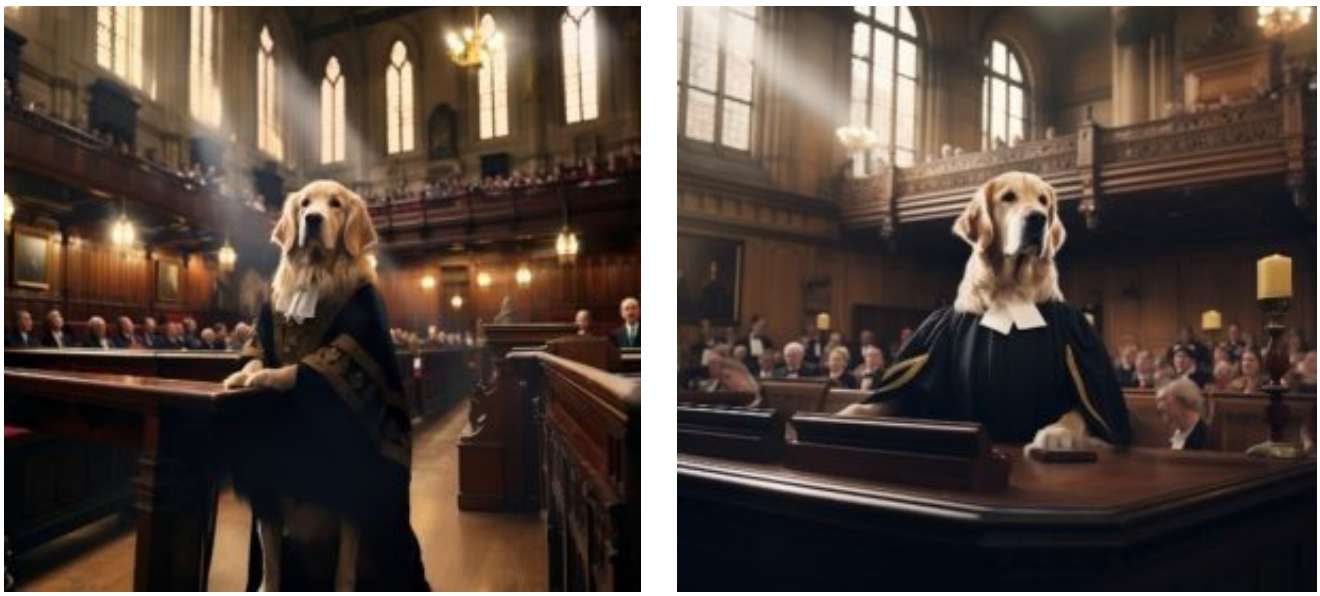


Figure 5 – ruff justice

Or, perhaps, for supreme court judges to be depicted as cats:

Figure 6 – our feline tribunal

The platforms are also increasingly user-friendly. Midjourney works on text prompts, as basic as *'generate an English courtroom setting but the lawyer is a golden retriever'*

. It took around 30-60 seconds to generate each of those images above, and thereafter they become immediately available for download and sharing. It allows you to zoom in, zoom out, crop things out, and add details (e.g., add neck bands to the dog, make the cats fluffier, add a Red Book, remove a gavel) – and with a cost of £9.50 per month, anyone with a computer and half an imagination can access this limitless portal of image creation. With a paid subscription, users *'own'* all the images they create (but notably, not their copyright), and thus they can even be used commercially. The technology is also only getting better; with Midjourney on its fifth incarnation within just a year, it now boasts over 18 million monthly users across the world.

But how accurate is it?

To fully test this, I spent a rainy Sunday afternoon submitting various photographs and throwing random text prompts into the mix. I wanted to see how a solicitor with zero technological skill (and even less artistic talent) could create fantastically believable digital imagery.

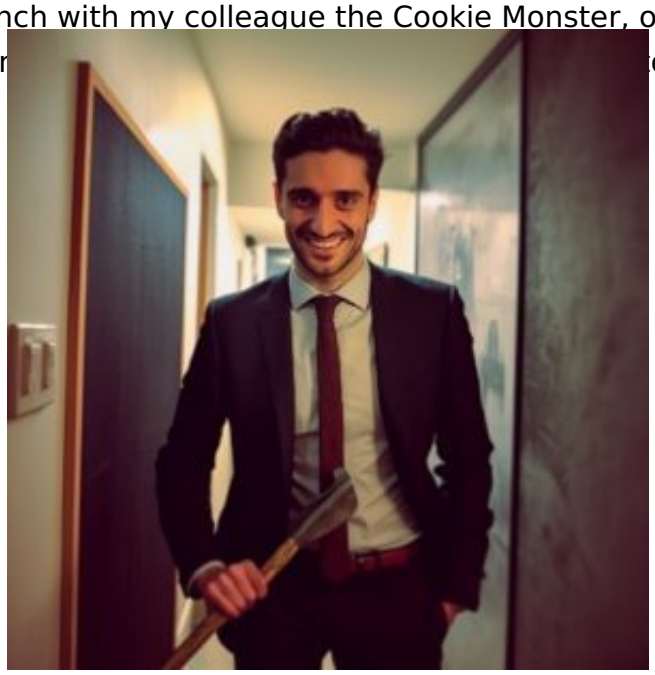
Within an hour or so, I had generated over 100 images. At first, the images were bordering on cartoonish. Others resembled movie posters or anime, and a few had missing limbs or were dreadfully blurred in parts. A number of results were incredibly corny, with *'fictional'* me having impossibly good hair, with fewer bags under the eyes and the like. However, with a few more prompts, and after giving it the opportunity to *'learn'* from my images, the results were striking. The images became less polished and fictitious, and became much more realistic.

Appearance wise, the first few images bore a vague similarity to me – in that they depicted what could be a fictitious second cousin, or a brother I never knew I had. As I continued, more images were generated and each had a slightly more realistic appearance. Later images had more of a passing semblance. With each click, the programme became more accurate, and by the end of the exercise I genuinely could not tell if particular images were the *'real me'* or the *'fictitious me'* at first blush. With the use of clever filters, some amendments, handy prompts and resizing the image, I would have been none the wiser. I went as far as sending the AI-generated images to family members without any prior warning; they failed to spot them being even slightly off.

See, for example, the images at the top: the subtle, AI-driven additions in each image made them look even more realistic – such as the candid companion on safari facing the camera, or the silhouette of a woman in shot with my bogus award. The images are detailed, complex, and entirely fake – and took mere minutes to manufacture. Remember, these images are all based on AI, with nothing much more than a steer from me and the programme scouring the internet to fill in

the blanks, I could be portrayed as sharing lunch with my colleague the Cookie Monster, or

ver



o Paris

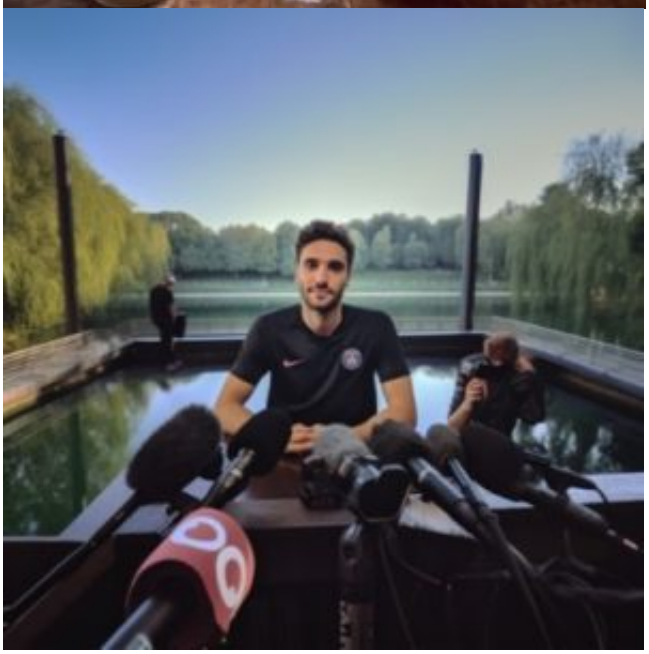


Figure 7 – not me with a muppet, not me with a spatula and definitely not me transferring to PSG from iFLG

So, I hear you ask – what does this little exercise in vanity show? It is incredibly easy to create

images in this way – perhaps too easy. But a more salient point arises: if my own family can't discern a real image of me from a bogus one of me, how could a judge possibly do so? How will this technology be used in an evidential arena, and how can a court control it?

AI-generated images in court proceedings

The use of AI in court proceedings is nothing new; you may have **spotted** the US attorneys named, shamed, and fined for using an AI chat service to help them draft their pleadings. They failed to note that their finished brief, submitted to court, contained completely non-existent cases. However, what can be done with AI-generated images, and what risks do they pose?

The most prominent issue is this: it is presently nigh-on impossible to detect if an image is computer generated. Whilst there are some tips and tricks to detect some signs of it being AI generated, a highly sophisticated digital image may pass as entirely *'normal'* and *'organic'*. There is currently no publicly available programme, app or website that can test the veracity of an image – leaving the average person (or indeed the average lawyer, or judge) completely clueless as to its provenance. In children proceedings in particular, the consequences of this could be huge.

Example – findings of fact hearings

Suppose, for example, a hypothetical couple are engaged in highly fraught Children Act proceedings and allegations of domestic violence are raised. Depending on the severity of the allegations, a judge is likely to be tasked with determining the truth of each allegation, and will be looking at a number of factors, to include a witness' oral and written evidence. This is a depressingly regular feature of children proceedings, and depending on the case, contemporaneous actions, messages and images could bolster an allegation, or torpedo it quickly.

For the sake of argument, let's say that the father alleges that the mother assaulted him in a fit of rage at 1pm on 1 April 2023, giving him a black eye. He says he then left the family home, spent the afternoon alone in Hyde Park, messaged a friend about the incident, and later received a WhatsApp message from the mother apologising for assaulting him. The mother denies this in its entirety, and a judge is to determine if this took place as described.

In advancing this allegation, the father produces a photograph he says he took following the alleged attack, *'time stamped'* to 1pm on 1 April 2023. This is the photograph:

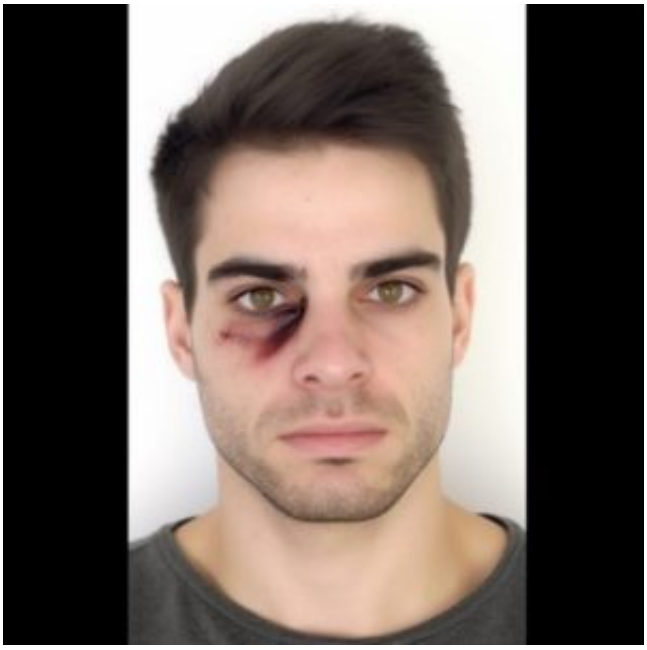
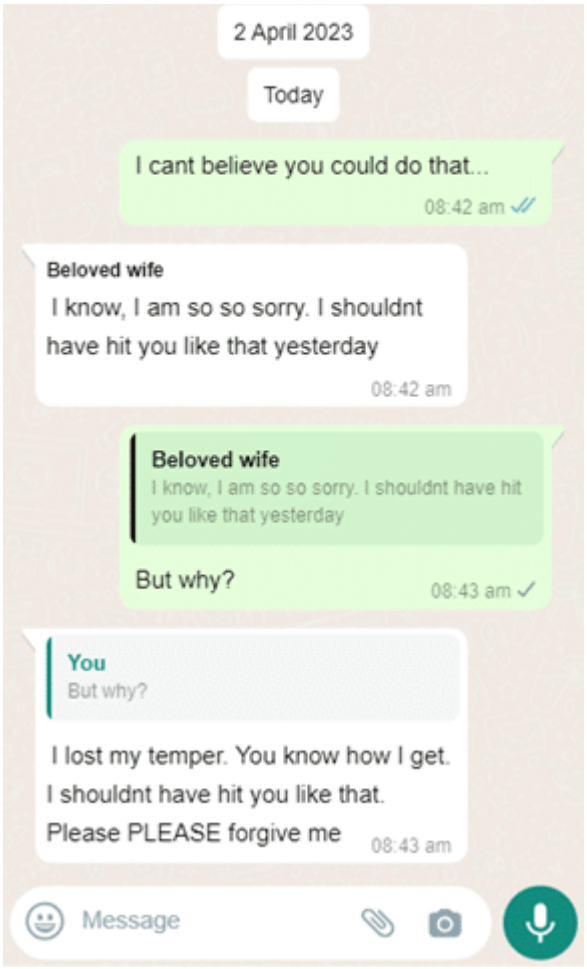


Figure 8 – a real shiner

The photograph's properties reflect it being taken on this date and time, and accompanying the photograph is the following chain of WhatsApp messages:



Alongside this, the father has corroborative messages (also on WhatsApp) with his friends about the incident. He also has a *'selfie'* and a Google Maps record of him being in Hyde Park on the day in question, alongside online searches (again timestamped) revealing searches for men's refuges, how to reduce swelling on a black eye, and enquiries about calling the police. The mother simply denies the incident took place.

A judge, who plainly cannot intimately know this family nor the father, considers the evidence in full. The judge is only likely to have seen the father once or perhaps twice in the actual courtroom; the images bear a real likeness to him, and the jigsaw pieces of his actions on the day all fit

together. The judge finds the ostensibly contemporaneous material particularly compelling, and goes on to determine that, *'on the balance of probabilities',* the mother did assault the father in the way he alleges. Arrangements for the welfare of the parents' children are thereafter determined with this forming part of the factual matrix, against a backdrop where findings of fact are notoriously difficult to appeal. The parties thereafter have certainty as to disputed allegations, and the family can move on accordingly.

However, the images above are entirely bogus. They are AI fakes. The bruised eye is a combination of my own passport photograph, mixed with other pictures, and the right prompts to bring out bruises. Indeed, unscrupulous litigants even have a choice of which delightful bogus image to deploy:

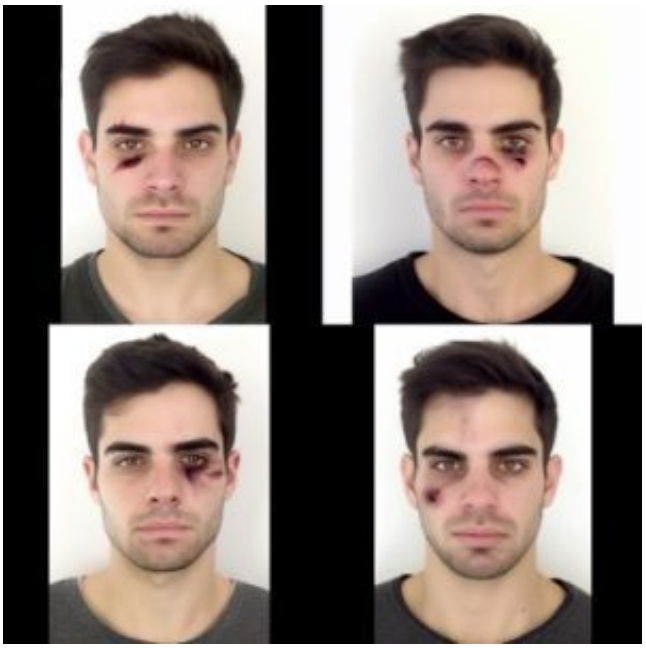


Figure 9 – spoilt for choice

The WhatsApp chat is also entirely AI-generated; and the same content can be changed to be a Snapchat chain, a Facebook message, a text chain, or even a LinkedIn post. The text can read to say anything one wants, and can be timestamped to any time and date. Search history can be similarly generated too. The father is able to bolster his case, hoodwinking the judge into believing his version of events. Given how accurate the exhibits look, and how difficult it is to disprove them, a judge can hardly be criticised for making this finding. Moreover, with any sort of technological examination of a mobile phone likely to be a disproportionate, costly and laborious exercise, what is a judge to do? With the foundational facts being conceived on the basis of complete falsehoods,

the outcomes are stark for this hypothetical family.

More to the point, this is just one allegation. The same tactic could be levied in support of multiple allegations. A pattern of *'behaviour'* – based once again on falsehoods – could emerge, and all it took was £9.50 and a bit of tinkering with images and text prompts.

Yet this can be extrapolated even further: AI can be used to generate any image, no matter how tasteless or alarming. There is next to no limit on this, and could easily include parents engaged in nefarious activities. Continuing the theme, in just a short space of time, I was able to swiftly generate a range of images, including threatening selfies, offensive weapons and further bruises or cuts. They can be customised in virtually any way – to include making the image look more sombre, or more threatening, or removing a smile. Within a matter of minutes, I could be depicted as wielding a hunting knife, being passed out drunk in my own living room, or even having joined ISIS. In the creation of a dishevelled, drunken scene, AI even thought it fit to remove my tie and undo my shirt's top button without me asking. Following my prompts, *'fake'* me is depicted on a sofa virtually identical similar to the one I own in real life.

On paper, safeguards against the production of overly false, vile, offensive images exist on the platform. However, any built-in safety features on the AI generator are easily bypassed. There are next to no limits to its absurdity.

Where do we go from here?

Even in passing, one can contemplate near endless, complex evidential questions around AI. AI could be used to generate false bank statements in financial proceedings, or a bogus air ticket in an abduction matter. It could be used to create immoral, obscene or illegal images, which are later surreptitiously found on another device. The creation of depraved images is certainly **entirely possible**. It can also be used to fabricate chats that – at least at present – are impossible to verify. Against this, the family court has arguably been significantly laxer about rules of evidence, certainly when compared to other jurisdictions; if a piece of evidence relates directly to an issue of a child's welfare, should it be easily excluded?

As the technology evolves, results are likely to be more and more believable, and harder and harder to disprove. With the parallel advent of *'deepfake'*, audio and video files can be digitally manipulated to portray likenesses of an individual – so we may well see videos of abusive behaviour or threatening calls, alongside AI-generated images, which all later turn out to be

entirely false. Conversely, is there a risk that, by raising the spectre of AI-generated images, will it plant a seed of doubt in relation to anything produced by a litigant? If AI manipulation cannot be disproved, then parties could realistically sound the AI alarm – against which there is presently no easily-available remedy.

Devious litigants, false documentation and bogus evidence are plainly not new to the family court. The court has faced similar challenges in the past when faced with either a forged signature, a curious testamentary disposition, a questionable bank statement, or a surprising *'dear John'* letter. Fraud using AI is no less deplorable; it is an exceptionally serious civil (and potentially criminal) matter. The difference now, however, is that AI makes this potential for fraud available to *anyone.* All you need is time, an internet connection, and an intention to mislead the court. It risks becoming far more widespread and far more challenging to discount. The court has always adopted a hard-line approach to the generation of fake documents – and when the first example of bogus AI-generated evidence does inevitably arise, one can imagine the justifiably dim and probably punitive view a judge is likely to adopt.

Previously, experts' reports and advancements in technology have stepped into the breach, hoping to assist judges through the murky process of evaluating evidence and credibility. This is likely to be an area of increasing importance in proceedings going forward; but until this develops, practitioners and judges must remain alive to the real risks of bogus documentation endemic in evidence.

And finally, in the words of Abraham Lincoln, you can't believe everything you read on the internet.

James Netto
james.netto@iflg.uk.com
The International Family Law Group LLP
www.iflg.uk.com