The Elusive Definition of ‘Deepfake’

Originally published at https://www.unite.ai/the-elusive-definition-of-deepfake/

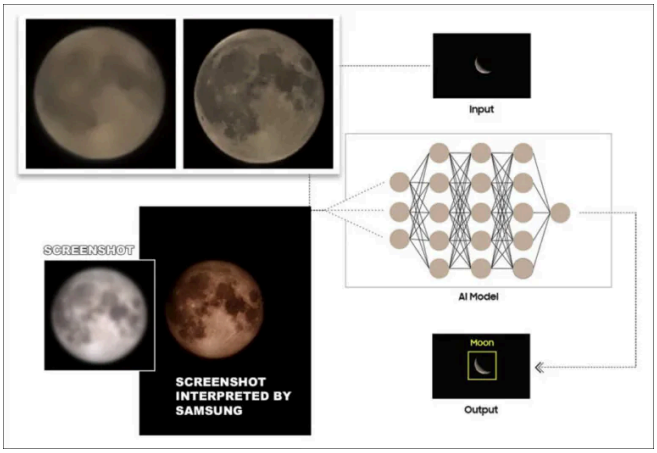
A compelling new study from Germany critiques the EU AI Act’s definition of the term ‘deepfake’ as overly vague, particularly in the context of digital image manipulation. The authors argue that the Act’s emphasis on content resembling real people or events – yet potentially *appearing* fake – lacks clarity.

They also highlight that the Act’s exceptions for ‘standard editing’ (i.e., supposedly minor AI-aided modifications to images) fail to consider both the pervasive influence of AI in consumer applications and the subjective nature of artistic conventions that predate the advent of AI.

Imprecise legislation on these issues gives rise to two key risks: a ‘chilling effect,’ where the law’s broad interpretive scope stifles innovation and the adoption of new systems; and a ‘scofflaw effect,’ where the law is disregarded as overreaching or irrelevant.

In either case, vague laws effectively shift the responsibility of establishing practical legal definitions onto future court rulings – a cautious and risk-averse approach to legislation.

AI-based image-manipulation technologies remain notably ahead of legislation’s capacity to address them, it seems. For instance, one noteworthy example of the growing elasticity of the concept of AI-driven ‘automatic’ post-processing, the paper observes, is the ‘Scene Optimizer’ function in recent Samsung cameras, which can replace user-taken images of the moon (a challenging subject), with an AI-driven, ‘refined’ image:



Top left, an example from the new paper of a real user-taken image of the moon, to the left of a Samsung-enhanced version automatically created with Scene Optimizer; Right, Samsung's official illustration of the process behind this; lower left, examples from the Reddit user u/ibreakphotos, showing (left) a deliberately blurred image of the moon and (right), Samsung's re-imagining of this image – even though the source photo was a picture of a monitor, and not the real moon. Sources (clockwise from top-left): https://arxiv.org/pdf/2412.09961; https://www.samsung.com/uk/support/mobile-devices/how-galaxy-cameras-combine-super-resolution-technologies-with-ai-to-produce-high-quality-images-of-the-moon/; https:/reddit.com/r/Android/comments/11nzrb0/samsung_space_zoom_moon_shots_are_fake_and_here/

In the lower-left of the image above, we see two images of the moon. The one on the left is a photo taken by a Reddit user. Here, the image has been deliberately blurred and downscaled by the user.

To its right we see a photo of the same degraded image taken with a Samsung camera with AI-driven post-processing enabled. The camera has automatically 'augmented' the recognized 'moon' object, even though it was not the real moon.

The paper levels deeper criticism at the Best Take feature incorporated into Google's recent smartphones – a controversial AI feature that edits together the 'best' parts of a group photo, scanning multiple seconds of a photography sequence so that smiles are shuffled forward or backward in time as necessary – and no-one is shown in the middle of blinking.

The paper contends this kind of composite process has the potential to misrepresent events:

> *'[In] a typical group photo setting, an average viewer would probably still consider the resulting photo as authentic. The smile which is inserted existed within a couple of seconds from the remaining photo being taken.*
>
> *'On the other hand, the ten second time frame of the best take feature is sufficient for a mood change. A person might have stopped smiling while the rest of the group laughs about a joke at their expense.*
>
> *'As a consequence, we assume that such a group photo may well constitute a deep fake.'*

The new paper is titled *What constitutes a Deep Fake? The blurry line between legitimate processing and manipulation under the EU AI Act*, and comes from two researchers at the Computational Law Lab at the University of Tübingen, and Saarland University.

Old Tricks

Manipulating time in photography is far older than consumer-level AI. The new paper's authors note the existence of much older techniques that can be argued as 'inauthentic', such as the concatenation of multiple sequential images into a High Dynamic Range (HDR) photo, or a 'stitched' panoramic photo.

Indeed, some of the oldest and most amusing photographic fakes were traditionally created by school-children running from one end of a school group to another, ahead of the trajectory of the special panoramic cameras that were once used for sports and school group photography – enabling the pupil to appear twice in the same image:

The temptation to trick panoramic cameras during group photos was too much to resist for many students, who were willing to risk a bad session at the head's office in order to 'clone' themselves in school photos. Source: https://petapixel.com/2012/12/13/double-exposure-a-clever-photo-prank-from-half-a-century-ago/

Unless you take a photo in RAW mode, which basically dumps the camera lens sensor to a very large file without any kind of interpretation, it's likely that your digital photos are not completely authentic. Camera systems routinely apply 'improvement' algorithms such as image sharpening and white balance, by default – and have done so since the origins of consumer-level digital photography.

The authors of the new paper argue that even these older types of digital photo augmentation do not represent 'reality', since such methods are designed to make photos more pleasing, not more 'real'.

The study suggests that the EU AI Act, even with later amendments such as recitals 123–27, places all photographic output within an *evidentiary* framework unsuited to the context in which photos are produced these days, as opposed to the (nominally objective) nature of security camera footage or forensic photography. Most images addressed by the AI Act are more likely to originate in contexts where manufacturers and online platforms *actively promote* creative photo interpretation, including the use of AI.

The researchers suggest that photos 'have never been an objective depiction of reality'. Considerations such as the camera's location, the depth of field chosen, and lighting choices, all contribute to make a photograph deeply subjective.

The paper observes that routine ‘clean-up’ tasks – such as removing sensor dust or unwanted power lines from an otherwise well-composed scene – were only *semi*-automated before the rise of AI: users had to manually select a region or initiate a process to achieve their desired outcome.

Today, these operations are often triggered by a user’s text prompts, most notably in tools like Photoshop. At the consumer level, such features are increasingly automated *without* user input – an outcome that is apparently regarded by manufacturers and platforms as ‘obviously desirable’.

The Diluted Meaning of ‘Deepfake’

A central challenge for legislation around AI-altered and AI-generated imagery is the ambiguity of the term ‘deepfake’, which has had its meaning notably extended over the last two years.

Originally the terms applied only to video output from autoencoder-based systems such as DeepFaceLab and FaceSwap, both derived from anonymous code posted to Reddit in late 2017.

From 2022, the coming of Latent Diffusion Models (LDMs) such as Stable Diffusion and Flux, as well as text-to-video systems such as Sora, would also allow identity-swapping and customization, at improved resolution, versatility and fidelity. Now it was possible to create diffusion-based models that could depict celebrities and politicians. Since the term’ deepfake’ was already a headline-garnering treasure for media producers, it was extended to cover these systems.

Later, in both the media and the research literature, the term came also to include *text-based impersonation*. By this point, the original meaning of ‘deepfake’ was all but lost, while its extended meaning was constantly evolving, and increasingly diluted.

But since the word was so incendiary and galvanizing, and was by now a powerful political and media touchstone, it proved impossible to give up. It attracted readers to websites, funding to researchers, and attention to politicians. This lexical ambiguity is the main focus of the new research.

As the authors observe, article 3(60) of the EU AI Act outlines four conditions that define a ‘deepfake’.

1: True Moon

Firstly, the content must be *generated or manipulated*, i.e., either created from scratch using AI (generation) or altered from existing data (manipulation). The paper highlights the difficulty in distinguishing between ‘acceptable’ image-editing outcomes and manipulative deepfakes, given that digital photos are, in any case, never true representations of reality.

The paper contends that a Samsung-generated moon is arguably authentic, since the moon is unlikely to change appearance, and since the AI-generated content, trained on real lunar images, is therefore likely to be accurate.

However, the authors also state that since the Samsung system has been shown to generate an ‘enhanced’ image of the moon in a case where the source image was not the moon itself, this would be considered a ‘deepfake’.

It would be impractical to draw up a comprehensive list of differing use-cases around this kind of *ad hoc* functionality. Therefore the burden of definition seems to pass, once again, to the courts.

2: TextFakes

Secondly, the content must be *in the form of image, audio, or video*. Text content, while subject to other transparency obligations, is not considered a deepfake under the AI Act. This is not covered in any detail in the new study, though it can have a notable bearing on the effectiveness of *visual* deepfakes (see below).

3: Real World Problems

Thirdly, the content must *resemble existing persons, objects, places, entities, or events*. This condition establishes a connection to the real world, meaning that purely fabricated imagery, even if photorealistic, would not qualify as a deepfake. Recital 134 of the EU AI Act emphasizes the 'resemblance' aspect by adding the word 'appreciably' (an apparent deferral to subsequent legal judgements).

The authors, citing earlier work, consider whether an AI-generated face need belong to a real person, or whether it need only be adequately *similar* to a real person, in order to satisfy this definition.

For instance, how can one determine whether a sequence of photorealistic images depicting the politician Donald Trump has the intent of impersonation, if the images (or appended texts) do not specifically mention him? Facial recognition? User surveys? A judge's definition of 'common sense'?

Returning to the 'TextFakes' issue (see above), words often constitute a significant portion of the act of a *visual* deepfake. For instance, it is possible to take an (unaltered) image or video of '*person a'*, and say, in a caption or a social media post, that the image is of '*person b'* (assuming the two people bear a resemblance).

In such as case, *no AI is needed*, and the result may be strikingly effective – but does such a low-tech approach also constitute a 'deepfake'?

4: Retouch, Remodel

Finally, the content must *appear authentic or truthful to a person*. This condition emphasizes the *perception* of human viewers. Content that is only recognized as representing a real person or object by an algorithm would *not* be considered a deepfake.

Of all the conditions in 3(60), this one most obviously defers to the later judgment of a court, since it does not allow for any interpretation via technical or mechanized means.

There are clearly some inherent difficulties in reaching consensus on such a subjective stipulation. The authors observe, for instance, that different people, and different types of people (such as children and adults), may be variously disposed to believe in a particular deepfake.

The authors further note that the advanced AI capabilities of tools like Photoshop challenge traditional definitions of 'deepfake.' While these systems may include basic safeguards against controversial or prohibited content, they dramatically expand the concept of 'retouching.' Users can now add or remove objects in a highly convincing, photorealistic manner, achieving a professional level of authenticity that redefines the boundaries of image manipulation.

The authors state:

‘We argue that the current definition of deep fakes in the AI act and the corresponding obligations are not sufficiently specified to tackle the challenges posed by deep fakes. By analyzing the life cycle of a digital photo from the camera sensor to the digital editing features, we find that:

‘(1.) Deep fakes are ill-defined in the EU AI Act. The definition leaves too much scope for what a deep fake is.

‘(2.) It is unclear how editing functions like Google’s “best take” feature can be considered as an exception to transparency obligations.

‘(3.) The exception for substantially edited images raises questions about what constitutes substantial editing of content and whether or not this editing must be perceptible by a natural person.’

Taking Exception

The EU AI Act contains exceptions that, the authors argue, can be very permissive. Article 50(2), they state, offers an exception in cases where the majority of an original source image is not altered. The authors note:

‘What can be considered content in the sense of Article 50(2) in cases of digital audio, images, and videos? For example, in the case of images, do we need to consider the pixel-space or the visible space perceptible by humans? Substantive manipulations in the pixel space might not change human perception, and on the other hand, small perturbations in the pixel space can change the perception dramatically.’

The researchers provide the example of adding a hand-gun to the photo a person who is pointing at someone. By adding the gun, one is changing as little as 5% of the image; however, the semantic significance of the changed portion is notable. Therefore it seems that this exception does not take account of any ‘common-sense’ understanding of the effect a small detail can have on the overall significance of an image.

Section 50(2) also allows exceptions for an ‘assistive function for standard editing’. Since the Act does not define what ‘standard editing’ means, even post-processing features as extreme as Google’s Best Take would seem to be protected by this exception, the authors observe.

Conclusion

The stated intention of the new work is to encourage interdisciplinary study around the regulation of deepfakes, and to act as a starting point for new dialogues between computer scientists and legal scholars.

However, the paper itself succumbs to tautology at several points: it frequently uses the term ‘deepfake’ as if its meaning were self-evident, whilst taking aim at the EU AI Act for failing to define what actually constitutes a deepfake.

First published Monday, December 16, 2024

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More