# Artificial Intelligence, Deepfakes and a Future of Ectypes

**Luciano Floridi**[1,2]

The art world is full of reproductions. Some are plain replicas, for example the Mona Lisa. Others are fakes or forgeries, like the "Vermeers" painted by Han van Meegeren that sold for $60 million (Kreuger and van Meegeren 2010). The distinction between a *replica* and a *fake* is based on the concept of *authenticity*. Is this artefact what it claims to be?[1] The answer seems simple but, in reality, things are complicated. Today, the paintings of the forger John Myatt are so famous that they are valued at up to $40,000 each, as "genuine fakes" (Furlong 1986). They are not what they say they are, but they are authentically painted by him and not by another forger. And they are beautiful. A bit as if one were to utter a beautiful lie, not any ordinary lie. And an artist like Magritte seems to have painted not only false Picassos and Renoirs during the Nazi occupation of Belgium (Mariën 1983), but also faked his own work, so to speak, in the famous case of the two copies of the painting "The Flavour of Tears" (1948), both by Magritte, but one of which he passed off as false—partly as a surrealist act and partly to make money. In this mess, and as if things were not confusing enough, digital technologies further reshuffle what is possible and our understanding of it.

Thanks to digital technologies, today it is much easier to establish the authenticity of a work. There are databases where you can check authors' signatures, and millions of images that can be viewed with a few clicks. Selling a fake is more difficult. Figure 1 shows a reproduction of the "Lodge on Lake Como" by Carl Frederik Peder Aagaard (1833–1895), a Danish landscape painter and decorative artist. It was on sale in 2016 on eBay. The painting is very popular on the web, and there are plenty of good replicas. Nothing wrong with them. However, if you check Fig. 1 carefully, you will notice that this is sold as an unsigned "original", which is misleading to say the least. Both the quality of the painting and the price are suspicious, and a Google image search quickly

[1] I have discussed the nature of questions and epistemic relevance in (Floridi 2008).

✉ Luciano Floridi
luciano.floridi@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

[2] The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK

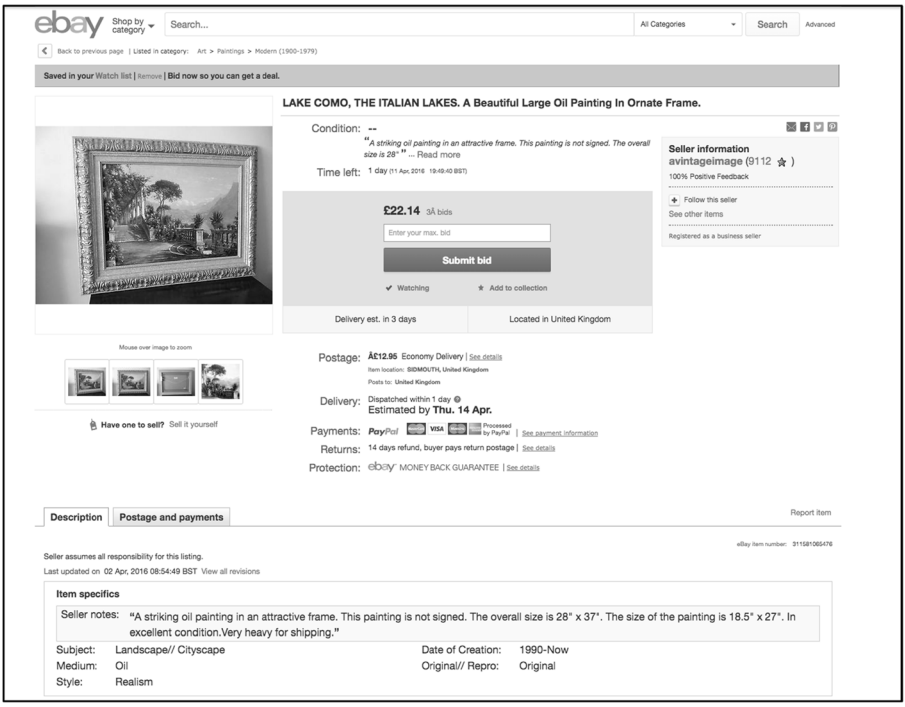


**Figure 1** A fake, the original is "Lodge on Lake Como" by Carl Frederik Peder Aagaard (1833–1895)

reveals that this is a mere replica. At the time of writing, the painting was no longer available and the seller did not seem to be active on eBay anymore.

Of course, fakes are not always reproductions; they can also be "new works" by a famous artist, like Pollock or Van Gogh. In this case, sophisticated scientific techniques to establish authenticity include tests run using AI. A research paper, published last November by Ahmed Elgammal, Yan Kang and Milko Den Leeuw (Elgammal et al. 2017) proposed "a computational approach for analysis of strokes in line drawings by artists", based on neural networks. The training collection consisted of a dataset of 300 digitised drawings with over 80,000 strokes, by Pablo Picasso, Henry Matisse and Egon Schiele, and a few works by other artists. By segmenting individual strokes, the system learned to quantify the characteristics of individual strokes in drawings, thus identifying the unique properties for each artist. The software managed to classify "individual strokes with accuracy 70%-90%, and aggregate over drawings with accuracy above 80%, while being robust to be deceived by fakes (with accuracy 100% for detecting fakes in most settings)". It turns out that the way in which individuals draw lines is as unique as their fingerprints or their gait, and AI can help one to discover it, as if it were a microscope.

But AI is not just for identifying fakes. Let us stay in the Netherlands, a very interesting project[2] by Microsoft, in collaboration with the Rembrandt House Museum, has led to the creation of a portrait of a gentleman, which both is and is not a Rembrandt (see Fig. 2).

[2] See https://news.microsoft.com/europe/features/next-rembrandt/

**Fig. 2** The Rembrandt that is not a Rembrandt. Microsoft Project with the Rembrandt House Museum

Analysing the known works of Rembrandt, an algorithm identified the most common subject (a portrait of a Caucasian man, 30–40 years old), the most common traits (facial hair, facing to the right, wearing a hat, a collar and dark clothing, etc.), the most suitable style to reproduce these characterising properties, the brushstrokes, in short, all the information needed to produce a new painting by Rembrandt. Having created it, it was reproduced using a 3D printer, to ensure that the depth and layering of the colour would be as close as possible to Rembrandt's style and way of painting. The result is a masterpiece. A Rembrandt that Rembrandt never painted, but which challenges our concepts of "authenticity" and "originality", given the painting's strong link with Rembrandt himself. I do not know the value of the painting. My bet is that it would be quite expensive if it were auctioned as reliably authenticated as *that unique* Microsoft's Rembrandt.

We do not have a word to define an artefact such as Microsoft's Rembrandt. So let me suggest *ectype*. The word comes from Greek and it has a subtle meaning that is quite useful here: an ectype is a copy, yet not any copy, but rather a copy that has a special relation with its source (the origin of its creation), the archetype. In particular, an ectype is the impression left by a seal. It is not the real thing, but it is clearly linked in a significant, authentic way with the real thing itself. Locke used "ectypes" to refer to ideas or impressions that correspond, although somewhat inadequately, to some external realities (the archetypes) to which they refer (Locke 2008). Digital technologies are able to separate the archetypal source—what was in the mind of the artist, for example—from the process (style, method, procedure) that leads from the source to the artefact (Floridi 2017). Once this link is severed, one can have ectypes that are "authentic" in style and content, but not "original", in terms of archetypal source, like Microsoft's Rembrandt. But one can also have ectypes that are "original" in terms of archetypal source (they do come from where they purport to come) yet not "authentic" in terms of production, performance, or method (they are not the ones used by the source to deliver the artefact). In other words, ectypes can be authentic but unoriginal artefacts, like Microsoft's Rembrandt, or inauthentic but original artefacts. A great example of an inauthentic original ectype was provided in March by an audio recording of John F. Kennedy's last speech. Despite being an ordinary speech from a decades-old campaign trail, it suddenly made headline news. Because it was the Dallas Trade Mart

speech of 22 November 1963, the text that JFK *would* have read, had he not been assassinated mere moments before, on his way to deliver it. The text is *original*: it comes from the source. But the voice that recites is *inauthentic*, because it was synthesised by software that analysed 831 recordings of Kennedy's speeches and interviews, in order to "learn" how to speak like him. The software finally gave voice to JFK's last speech 55 years late. So here is a Kennedy who is and is not a Kennedy, similar and yet different from the Rembrandt that is and is not a Rembrandt. They are both ectypes (see Table 1).

We saw that the production of ectypes does not stop at the work of art, but involves any artefact, from texts to photos, from audio recordings to videos. It is well known that the history of manuscripts, printing, photography, cinema and television is paved with fakes. Expect more ectypes too. In particular, artists love to break boundaries and it is easy to imagine that, like Magritte faking his own painting, they will start producing their own ectypes. Imagine a painter using the software developed by Microsoft to produce her own new works. It would still be an ectype, and this would explain why (with qualifications) the process would capture some authenticity. The reproduction of the work of art by mechanical means will have acquired a new meaning (Benjamin 2008).

With ectypes, we usually know where things stand. But someone could cheat. Last May, Google presented Google Duplex, a version of its AI assistant that simulates being human to help users with simple interactive tasks, like booking a restaurant table. The company was quick to state that it will not intentionally mislead anyone, and that it will make sure always to clarify when a user is interacting with an artificial agent. But someone else could use these technologies for criminal or evil purposes. This is what happens with Deepfake, a set of techniques used to synthesise new visual products, for example by replacing faces in the originals. The typical cases involve porn movies in which the faces of famous actresses like Gal Gadot or Scarlett Johansson (this is regularly about women's faces) are used to replace the original faces. In this case too, large databases are needed to instruct the software (which is available for free, and there is also an app), so if you are not a public figure the risks are lower. Deepfake also concerns politicians, like President Obama, for example.

What is the future ahead of us? Digital technologies seem to undermine our confidence in the original, genuine, authentic nature of what we see and hear. But what the digital breaks it can also repair, not unlike the endless struggle between software virus and antivirus. In our case, in addition to educating people, acquiring new sensitivities and having the right legal framework, there are at least a couple of interesting digital strategies. For artefacts that are already available, it is easy to imagine

**Table 1** Archetype, fake and ectypes

| | Original source | Authentic production |
|---|---|---|
| Leonardo's Mona Lisa | Yes | Yes |
| Han van Meegeren's forged Vermeers | No | No |
| Microsoft's Rembrandt | No | (Qualified) Yes |
| JFK's Trade Mart speech | Yes | No |

AI systems that give us a hand. It would be interesting to analyse Microsoft's Rembrandt and Kennedy's speech with an artificial system to see whether it discovered them to be ectypes. Research is already available on methods to expose Deepfake videos generated with neural networks (Li et al. 2018). In short, let us remember the software developed to analyse drawings: there are plenty of sophisticated tools for detection of image forgery. And more are likely to be developed as the demand for them increases. Next, as regards new artefacts, because originality and authenticity are also a matter of provable historical continuity from the source to the product through the process of production, the much-vaunted blockchain, or a similar solution, could make a big difference. Blockchain is like a register that stores transactions in an accruable, safe, transparent and traceable way. As a secure and distributed register of transactions, blockchain is being explored as a means of reliably certifying the origins and history of particular products: whether in terms of securing food supply chains, or in recording the many linked acts of creation and ownership that define the provenance of an artwork. In the future, we may adopt the same solution wherever there is a need to ensure (or establish) the originality and authenticity of some artefact, be it a written document, a photo, a video or a painting. And of course, a future artist may want to ensure, through a blockchain, that her work of art as an ectype is really what it says it is. At that point we shall have travelled full circle, for we shall have "genuine ectypes", like the Microsoft's Rembrandt, or Kennedy's speech.

## References


Benjamin, W. (2008). *The work of art in the age of mechanical reproduction*. London: Penguin.

Elgammal, A., Kang, Y., & Den Leeuw, M. (2017). Picasso, matisse, or a fake? Automated analysis of drawings at the stroke level for attribution and authentication. *arXiv preprint arXiv:1711.03536*.

Floridi, L. (2008). Understanding epistemic relevance. *Erkenntnis, 69*(1), 69–92.

Floridi, L. (2017). Digital's cleaving power and its consequences. *Philosophy & Technology, 30*(2), 123–129.

Furlong, M. (1986). *Genuine fake : a biography of Alan Watts*. Portsmouth: Heinemann.

Kreuger, F. H., & van Meegeren, H. (2010). *Han van Meegeren revisited : his art & list of works*. Delft: F.H. Kreuger.

Li, Y, Chang, M.-C., Farid, H., & Lyu, S. (2018). In Ictu Oculi: exposing AI generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*.

Locke, J. (2008). *An essay concerning human understanding*. New York Oxford: Oxford University Press.

Mariën, M. (1983). *Le radeau de la mémoire : souvenirs déterminés*. Paris: Pré-aux-Clercs.