Post-Truth

Archives, GPT-2 and Fake News

Marion Carré

Since 2017, I have been examining the relationships between art and artificial intelligence (AI) from different perspectives: as an entrepreneur, teacher, speaker, author, and artist. I am based in Paris, but work on international projects. I am convinced that art is really helpful in better understanding AI and its impact on society. It can make visible, comprehensible, and tangible various issues that would not otherwise be perceived. My artistic practice consists of evoking new comparisons, of creating contrasts and anachronisms in order to help us take a step back. My work is rooted equally in the past and in the future, between which I build bridges to seek to understand the present. In this exploration, AI, new technologies, and archives occupy an important place, alternately as a medium or subject of study.

The research underlying this paper was carried out as part of one-year-long artist residency in 2021 called 'New Forms of Togetherness'. It was organized by the Alliance Française de Glasgow, the Goethe-Institut in Glasgow, and the French Institute of Scotland, and also included the following partners: the National Library of Scotland, the Neon Digital Arts Festival, and the Soba lab (a neuroscience laboratory). The residency was divided into several phases: a research phase, design phase, and production phase, punctuated by several workshops with artists and researchers. During the residency, I was able to present the progress of my work on a number of occasions such as the symposium 'Alt+Shift+Archive—unpacking the past, present and future of digital archives', organized by researchers and archivists as well as the Neon Digital Arts Festival. Finally, my work was the subject of a solo exhibition titled *Is it true? The Post-Truth Archive Factory* at Briggait in Glasgow.

Archive and Truth

Constructing the truth may seem counter-intuitive, as if it were a raw material that can be manipulated. This is, however, precisely what interested me: how to construct and manipulate the aura of truth to communicate false information (Carter 2007; Kastenhofer 2015), thus trying to capture the essence of this complex notion and ex-

plore its attributes and malleability. Archive and truth are two notions often used in the singular. The singular, however, obscures entire segments of the reality of these two notions (Ogilvie 2017). It anchors an idea of absoluteness and immutability, erasing the conditions of their production (Ketelaar 2006). This use of the singular is much more present today than in the past and says something about the relationship our society has with it. The archive, through the direct access it seems to give us to past events, embodies the truth to the point of replacing it. By contrast, I wanted to make archives and truths visible in their plural and relational forms, complementing and contradicting each other. Their incomplete, fragmentary, and piecemeal nature will always defeat our quest for the absolute. So, how can we deduce anything from what remains, without knowing everything that has been?

Archives and truths share two common ingredients: chance and choice, which act as filters in the process of production and transmission. Chance is perhaps the one that fascinates us the most because it escapes us: why freeze this event, why did this one come to us, why this one and not another? The randomness it represents shapes our memory through a multitude of factors, unrelated to the content of the documents, but which nevertheless leave their mark (Breakell 2008). By freezing trivial or extraordinary events, it selects and distinguishes a portion of facts and individuals of which we would otherwise have retained no trace. Far from acting based solely on transmission, the archive is produced by chance. But choice is never far away and is just as decisive as chance.[1] What should be told or not? What should be preserved or destroyed? The destruction of archives shapes our relationship with the past at least as much as their preservation (Pène 2016). One might think that this filter would be easier to expose than that of chance, but it is sometimes quite the opposite when the rules are not clearly stated. When it comes to archives and truths, we have as much to learn from what we choose to tell and amplify as from what we leave out.[2]

The phenomenon of the rapid accumulation of traces nowadays concerns both archives and truths, in connection with digital media. The information that they encapsulate for some and the form in which they circulate for others are copied and multiplied to the point of uncontrollable proliferation (Klein/Lemay 2013; Derrida 1995). And yet the latter contributes to their dissolution by precipitating brutal selections. The more the ocean of archives and information available expands, the more it

1 See the podcast 'Les petits papiers de l'archiviste', available at: https://www.radiofrance.fr/franceculture/podcasts/matieres-a-penser/les-petits-papiers-de-l-archiviste-3544601 (all URLs here accessed in August 2023).

2 See the podcast 'La mise en archives: histoires', available at: https://www.radiofrance.fr/franceculture/podcasts/matieres-a-penser/la-mise-en-archives-histoires-2427974, and "L'art des rapprochements", available at: https://www.radiofrance.fr/franceculture/podcasts/matieres-a-penser/l-art-des-rapprochements-1927870.

limits what we have access to. We consult only a tiny fraction of what is available. The algorithms of online platforms open up a path through the vastness of available data, while reducing what we have access to. We thus access the scum of reality, which we nevertheless take to be the sum of all that exists. We live at a time where we put truth before everything. Faced with this narrowing of our horizons and in a context where the lines between truth and falsity blur, we set out in search of the truth. This solitary and radical quest for truth isolates us and fragments reality. We mobilize facts as evidence to prove the validity of our own versions of the truth. Taking advantage of this extensive conception of truth, we are all the more vulnerable to artefacts assuming the appearance of reality placed in our path. Through our misperceptions and our sharing, we endorse them as truths. We have thus turned reality into a sum of private truths that clash with each other (Robert 2021).

Becoming a Forger

In order to carry out these examinations, I decided to become a forger, by building up my own collection of false archives. I was particularly interested in historical narratives, not those of a grand history, but those about ordinary people. Inspired by the approach of the French historian Arlette Farge, who was one of the first people to work on popular history based on judicial archives, I wanted to work with this type of archive (Farge 2013). Their brittleness, which brings us fragments of facts and lives that we would never have heard of without them, seemed to me to be an asset in creating in my turn false archives that might melt into the hubbub of history. Although I decided not to tackle well-known historical facts, creating an entire collection of fake archives was still a monumental task. Even if I stuck to the content of archives without trying to reproduce their material form, it required both a lot of imagination to think of plausible events and also deep historical knowledge of a society at a given time so as to limit anachronisms. Not to mention the fact that English is not my first language and that Old English is even less familiar to me.

To support me in this forgery mission, for which I had no particular skills, I decided to mobilize an artificial intelligence. To generate the texts of my fake archives, I chose the algorithm called GPT-2 developed by OpenAI. Pre-trained on millions of textual data, it is possible to re-train this algorithm on the dataset of one's choice so that it generates texts in the same style. It was important for me to choose a widely accessible algorithm that did not require a lot of technical knowledge to implement, in order to be merely one forger among many potential ones. To achieve plausible results in my forgery generation endeavour, it was important to take great care in building the dataset to train the artificial intelligence. At a time when most of the artificial intelligence algorithms that can be used to create are accessible to the greatest

number of people, producing the dataset seems to me to be one of the decisive steps in the artistic gesture.

Figure 1: An original broadside from the National Library of Scotland collection.

An Account of the last Dying Words and Execution of MRS. SHUTTLEWORTH, for the Murder of her Husband, on Friday the 7th of December inst. 1821, at Montrose, near Edinburgh.

I had a wide range of possibilities in the datasets made available by the National Library of Scotland through its Data Foundry. After several explorations and exchanges with the teams of the National Library of Scotland, I set my sights on a rather special archive: The Broadsides.[3] Among them, I was particularly interested in the crime broadsides, texts relating to sensationalist trials or events, the raw material that comes closest to a court record. This archive was even better insofar as these texts have a complex relationship with the truth, which they are known to have distorted or embellished for commercial purposes. They operate a form of *mise en abyme* vis-à-vis the truth by narrating events in these case trials with the aim of establishing the truth. Moreover, these texts carry a certain emotional charge that further disturbs our appreciation of the facts they relate. I therefore set out to read and sort through 1728 digitized texts available as open content from the National Library of Scotland's Data Foundry.[4]

3 https://www.nls.uk/collections/rare-books/collections/broadsides/.

4 https://data.nls.uk/data/digitised-collections/broadsides-printed-in-scotland/.

In the course of this long and tedious selection work, I sometimes had surprising encounters, such as that with Margaret Dickson. Condemned and hanged in 1823 at the age of 22 on suspicion of having killed the child she gave birth to alone after concealing her pregnancy, she miraculously woke up in her coffin (?!) and lived another 25 years in freedom. This type of archive, however authentic, went far beyond fiction and blurred the lines even more. After several months of work, I selected 500 texts with which to train the artificial intelligence. This was followed by several months of work to choose which texts to keep among all those generated by the artificial intelligence. This is also a very important step in the process of creating with artificial intelligence. Among the texts generated as a whole, I wanted to retain a spectrum ranging from partially incoherent content to plausible content without incoherence. The 100 texts selected have a common strangeness and at the same time a historical atmosphere that creates a kind of 'uncanny valley' of the archive. One becomes increasingly confused, since the results come so close to authentic archives that it becomes impossible to distinguish them.

Artistic Processing

Once I had created my collection of 100 fake archive texts, I wanted to stage it by combining it with authentic archives in three works offering different points of view on the process of creating fake news and adhering to a form of chronology in their approach. This triptych stages three temporal sequences in the life of archives and fake news: production (Living Organism), selection (Selective Memories), and transmission (The Post-Truth Archive).

Living Organism

Installed on the high bookshelves of the École Nationale des Chartes, surrounded by old books and archives, printers copy the texts of original broadsides and AI generated ones on long rolls of paper. Here, the artificial intelligence endows the archive with a reproductive function. Through the process of generation, the mother (original) archive creates daughters, (generated) fictitious archives. This process underscores the living character of the archive as well as of the truths that we perceive as fixed, whereas they can evolve with time, along with our interpretations of them.

Through generation, artificial intelligence defies the finitude of the archive. Nourished by the original Broadsides archive, it randomly composes new texts, which are then added to the old ones, thus leading to an exponential and limitless accumulation process. Does the current flood of traces and the exponential growth of archives have anything to do with the presentiment of our own imminent disappearance? At a time when we are delegating more and more tasks to machines,

even to the point of attributing human characteristics such as intelligence to them, might we ask them to continue writing history without us? To what extent can we delegate tasks to machines?

Figure 2: The installation Living Organism *in the École Nationale des Chartes.*

Selective Memories

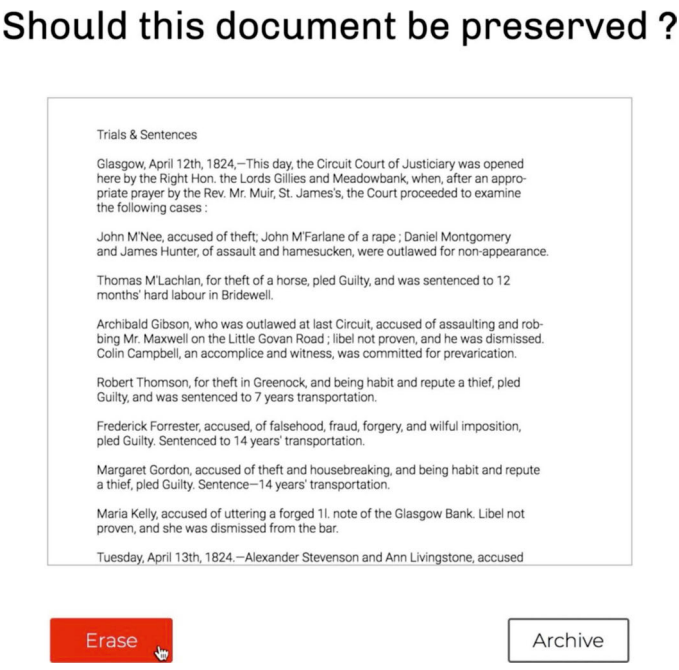
An online platform offers its users the possibility to archive or destroy texts presented to them, both original and generated. On this platform, the texts in the fake archive are presented on the same level as the real ones, in an identical format, thus reproducing the phenomenon of loss of reference that social networks create with respect to the information they relay. In order to respond to the instruction given, users will still have to find indicators to decide which archives to retain or destroy. Will verisimilitude be the only criterion? Will the emotion that certain texts may provoke get in the way? Can 'fake' texts succeed in producing such emotions in their readers? What place does emotion have in our understanding of truth and falsity?

While we conduct such processes of discriminating between information to separate the true from the false on a daily basis, this work, which is often unconscious, is not without consequences for our memory, into which false information can slip, to the point of our developing false memories. If our individual memories are corruptible, how immune is our collective memory to false information? Does the sharing of memories protect us from the manipulations to which we may be exposed as individuals? Is cooperation and cohesion a barrier to the penetration of false informa-

tion? Or is the collective, by contrast, a redoubtable sounding board for the spread of false information?

Alongside the artistic version of the project, real and fake archives from the dataset were also used to analyse how generated human-like texts are evaluated by general audiences. With the researchers Kohinoor Darda and Emily S. Cross, our goal was to investigate how people react to texts generated algorithmically, whether they are indistinguishable from original, human-generated texts, and the value people assign these texts. The findings of this study have been published in the Royal Society Open Science Journal. The key elements are that participants in this study were unable to distinguish between AI-generated archives and original archives. We also noticed that biases against artificial intelligence are found when participants are aware of the source of the archive or mistake it for an AI-generated archive.

Figure 3: The Selective Memories *platform.*



The Post-Truth Archive

The Post-Truth Archive is the output of collaborative work done by participants on the Selective Memories online platform. The red register brings together all

the texts that participants decided to archive on the platform. It is surrounded by texts—crumpled and scattered on the ground—that people decided to destroy. It blurs the line between truth and fake because original and generated broadsides can be found not only in the register, but also lying on the ground. It thus explores the perception mechanism that enables us to give credit to and propagate the fake. Is crowdsourcing reliable in this situation?

The approval of these texts by other peers can potentially introduce a confirmation bias in viewers. How things are presented can also have an impact and be a source of bias, and this is why the register makes use of codes for how official archives documents can be presented. Furthermore, the very status to be given to this body of text is questionable. It could be considered corrupted, since it would mix truth and falsehood. Yet the treatment of these texts may have even more to teach us. Is this corpus only a reflection of the true/false cleavage? Does the fact that some false archives were retained by certain participants give them a special status? Can the value given to this corpus go beyond the question of its veracity?

Figure 4: The exhibition Is it true? The Post-truth Archive Factory *at The Briggait in Glasgow.*

Detecting Fake News

Artificial intelligence has the paradoxical aspect of being able to be both the evil and a solution, and more and more applications for detecting forgeries are emerging in

art and culture. What we can blame AI for is giving anyone who wants to produce a forgery sophisticated tools. Fakes did not have to wait for artificial intelligence to exist before they could be produced on a large scale. And there is currently a race to improve existing techniques to make them even more efficient. The impact of this is that we can no longer rely on content that we see or hear online. We already see the huge challenges this creates for democracies, but it is now also spreading to delinquency: tomorrow you could receive a call from someone with your father's voice asking you to send him money. But like any tool, AI is not consequently good or bad, and can also be part of the solution.

Artificial intelligence requires training on large datasets, this is why datasets have also been created to enable them to be used for fake news detection projects (Liar, BuzzFace, Facebook Hoax …). These datasets are interesting because they evince some common patterns in fake news, such as many occurrences of first and second person pronouns for textual content. The AI models used to detect fake news use a different approach (Salazar 2020). Some of them examine content propagation and interaction between people. Researchers have found that false news stories on Twitter are 70 per cent more likely to be retweeted than true stories. Others focus on the content itself, which is more difficult when it is not text, but instead images or video (Vosoughi/Roy/Aral 2018). Some models focus on trying to detect whether one particular content propagates already known fake news, while others focus on the structure of the text or image itself. Researchers, big tech firms, and start-ups are developing such models. To name a few examples, Meta has developed a model called SimSearchNet++ to detect variations of an image with a very high degree of precision.[5] It uses OCR to detect texts in images. The start-up Sensity has developed a solution to detect deepfakes or images of faces generated with GANs.[6]

Research tends to show that—as is often the case with AI—methods employing a mix of human annotators and AI tools are more capable of detecting fake news. But, as always, this comes with risks of misuse: those technologies might also be used by governments aiming to censor information.

Beyond fake news, AI can also become a powerful tool to help us better detect art forgery. Lately, researchers such as the team of Ahmed Elgammal at Rutgers University or the team at Case Western Reserve University (CWRU) have been worked on using AI to be able to authenticate an artwork based on tiny details of brushstrokes that cannot be controlled by the artist (Elgammal/Kang/Leeuw 2018). In 2021, the CWRU team published a paper showing that they can identify which of the participants in the study (four students) actually made the painting with greater than 95 per cent accuracy (Ji/McMaster/Schwab et al. 2018). In conclusion, I believe that education is a powerful safeguard against fake news. Deepfakes necessitate the devel-

5 https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

6 https://sensity.ai/deepfakes-detection/.

opment of critical thinking skills in an era when seeing is no longer believing. This is where museum professionals have an important role to play in helping to raise awareness about fake news.

References

Breakell, Sue (2008). Perspectives: Negotiating the Archive. Tate Papers 9. Available online at https://www.tate.org.uk/research/tate-papers/09/perspectives-negotiating-the-archive (all URLs here accessed in August 2023).

Carter, Rodney G. S. (2007). Tainted Archives: Art, Archives, and Authenticity. Archivaria. Available online at https://archivaria.ca/index.php/archivaria/article/view/13128.

Derrida, Jacques (1995). Mal d'archive: une impression freudienne. Paris, Editions Galilée.

Elgammal, Ahmed/ Kang, Yan/ Leeuw, Milko Den (2018). Picasso, Matisse, or a Fake? Automated Analysis of Drawings at the Stroke Level for Attribution and Authentication. Proceedings of the AAAI Conference on Artificial Intelligence 32 (1). https://doi.org/10.1609/aaai.v32i1.11313.

Farge, Arlette (2013). Le Goût de l'archive. Paris, Éditions Seuil.

Ji, F./McMaster, M.S./Schwab, S. et al. (2021). Discerning the Painter's Hand: Machine Learning on Surface Topography. Heritage Science 9 (1). https://doi.org/10.1186/s40494-021-00618-w.

Kastenhofer, Julia (2015). The Logic of Archival Authenticity: ISO 15489 and the Varieties of Forgeries in Archives. Archives and Manuscripts 43 (3), 166–80. https://doi.org/10.1080/01576895.2015.1074085.

Ketelaar, Eric (2006). (Dé) Construire l'archive. Matériaux pour l'histoire de notre temps 82 (2), 65–70. https://doi.org/10.3917/mate.082.0065.

Klein, Anne/Lemay, Yvon (2013). Les archives à l'ère de leur reproductibilité numérique. In: Joumana Boustany (Ed.). La médiation numérique: renouvellement et diversification des pratiques. Actes du colloque Document numérique et société, Zagreb 2013. Louvain-la-Neuve, De Boeck Supérieur, 37–50. https://doi.org/10.3917/dbu.chron.2013.01.0037.

Ogilvie, Denise (2017). Paradoxes de 'l'archive'. Sociétés & Représentations 43 (1), 121–34. https://doi.org/10.3917/sr.043.0121.

Pène, Matthieu (2016). Le temps des archives et la périodisation historique: une histoire découpée en boîtes? Questes 33, 31–43. https://doi.org/10.4000/questes.4315.

Robert, Anne-Cécile (2021). Dernières nouvelles du mensonge. Montréal, Lux.

Salazar, Armida (2020). AI Tools on Fake News Detection: An Overview and Comparative Study. Available online at https://www.researchgate.net/publication/3

43859545_AI_Tools_on_Fake_News_Detection_An_Overview_and_Comparative_Study.
Vosoughi, Soroush/Roy, Deb/Aral, Sinan (2018). The Spread of True and False News Online. Science 359 (6380), 1146–51. https://doi.org/10.1126/science.aap9559.