The Orphanhood of the Archive in Contemporary Times: Between the Public and the Private

Carol Sabbadini

Abstract

Currently, new perspectives have emerged, allowing a reevaluation and expansion of the notion of orphanhood in relation to archives.
This goes beyond issues of authorship and ownership, embracing a critical and activist vision that contemplates alternative definitions impacting both the public and private spheres. On the other hand, technological developments that are increasingly prevalent in our daily lives, specifically machine learning and artificial intelligence (AI), are leading us to question the concept of archiving, the dynamics of orphaning or recreating, and the role of memory in the digital environment.

Keywords

Orphan, Orphaning, Archives, Authorship, Artificial Intelligence, Datasets, Memory, Public and Private.

Orphanhood

Through an exercise of questioning pre-established categories within traditional archival practices and exploring the archaeology of media through artistic languages, critical studies, and the management of audiovisual archives, there is a proposal to create spaces for discussion that reconsider the classification and valorization of documents designated as orphans due to their characteristics and typologies.

With a focus on the management, access, dissemination, and reuse of such documents and their respective contents, the category of "orphans" was established in the late 20th century to identify documents whose ownership and rights are unknown or cannot be traced. These documents often have doubtful or unidentifiable origins, unclear rights and ownership, varied topics, fragmented nature, lack of credits or attached metadata for identifying places, people, dates, or situations, and problematic or almost non-existent usage and circulation.

Multiple factors contribute to this, such as authors omitting credits or production years in their works, dissolution or disappearance of the owning companies or organizations (producers /distributors) without clarifying the rights to the materials, or uncertain and repeated distribution of works, making it nearly impossible to identify or trace the authors or owners of the archives.

Furthermore, the orphaning of archives can be attributed to the decisions of public or private institutions and the dilemma they face regarding preservation versus commercialization. These types of documents lack commercial value, cannot be distributed, disseminated, and their public access cannot be guaranteed due to the legal limbo they inhabit. This complicates their preservation due to the high costs involved and copyright laws that, in many cases, do not allow their release into the public domain, rendering them orphans of the world.

Expanding on canonical notions of archival orphanhood, the definition of the term can encompass documents that have been abandoned, marginalized, silenced, or disappeared from archives due to neglect during their custody and management. This is particularly true for works by amateur creators of small scale, undistributed films, discarded shots, censored materials, obsolete medical films, government surveillance images, documents created by marginalized communities (minorities or underrepresented groups), ephemeral films, advertisements, unfinished works, amateur films, family records, newsreels, ethnographic and scientific films, whose contents and the memories they preserve over time have been forgotten or erased.

Additionally, Foucault (2005) introduces the concept of the archive not only as mere repositories of historical information but as a representation of a system of power and control over knowledge and memory. [1]

According to his perspective, archives become spaces where objects and documents are divorced and separated from their original context, stripped of their meaning, and reorganized according to prevailing power structures, transforming objects into mere immutable relics devoid of their ability to generate new interpretations and narratives.

Building upon this, the orphanhood of some archives refers to their extrapolation from the original context in which they were created or recorded, resulting in isolated fragments. This could occur because their timeline, sequence, or original narrative was altered (bastardized) due to the intrinsic workflow, particularly in television, and the use of analog media (reuse of media for new recordings, deletion or rewriting during the editing or broadcasting phase). This led to multiple fragments with similar or diverse thematic content, colloquially known as "chunks" in some archives.

These "chunks" are associated with media and are primarily found on Betacam, U-matic, VHS, and Betamax videotapes, which, due to their intrinsic characteristics, allowed for quick recording and viewing, as well as manual and frequent rerecording, reediting, and erasure. This facilitated the creation of multiple archives on such analog supports for several decades but also led to the disappearance of many significant events recorded on these types of media.

An example of this is the news events recorded during the 1970s and 1980s, forming a prolific collection of newsreel "chunks" and original camera recordings (rushes), preserved in various institutions in different countries as Colombia.

In the first case, a sequentiality is evident, associated with newsreel sections and a narrative presented by anchors, a result of editing for broadcast. In the second case, there are a series of fragments, often unpublished, with similar or dissimilar thematic content, emphasizing the perspective and active role of the cameraperson not only as an operator but as a creator and narrator of a timeline, with meaning and sequentiality derived from the original moment of recording events.

On the other hand, orphanhood can also be considered in the context of orphan technologies, a term coined by Kittler (2018) to refer to devices and formats that become obsolete or discontinued, leaving behind a technological legacy without adequate support. These orphan technologies pose a challenge for information preservation since the archives and memories stored in those formats become inaccessible or difficult to retrieve. This undoubtedly enriches the debate on digital preservation, collective memory, and the relationship between technology and society. [2]

Computer Science: Artificial Intelligence and Machine Learning

How do we orphan the archive from digital and technological dynamics?

Currently, in the realm of computer science, orphaned files are defined as fragments or remnants of files or data that remain marginalized in databases, as ghosts of existing programs due to failures or errors in computer operating systems, or in the installation and uninstallation processes of programs and applications. This critical definition allows approaching archives from the perspective of errors and technical operability, raising questions about the various ways in which digital, technological dynamics, and computer science (Artificial Intelligence - AI and Machine Learning) approach the notion of orphan and the action of orphaning.

From computer science and the emergence of new technologies such as big data, which refers to the technologies and tools used to store, manage, process, and analyze extremely large and complex datasets that cannot be adequately managed and processed through traditional data processing methods, new terms associated with these dynamics emerge, such as databases and datasets.

Databases refer to these large organized systems for storing, managing, and retrieving data in a structured way, allowing users to perform various data operations and manipulations, such as viewing, browsing, and searching.

On the other hand, datasets are structured or unstructured collections that can contain various types of data, such as text, images, numbers, audio, videos, among others. They can vary in size and complexity, from small datasets used for testing and development to large datasets in terabytes or petabytes. These datasets can be public data available for download and use or private data that requires authorization and compliance with privacy and security policies. They are typically used for various purposes, such as training machine learning models, conducting statistical analyses, scientific research, decision-making, and other applications.

In line with the aforementioned, current image generation applications based on artificial intelligence (AI) systems like Dall-e, Midjourney, Stability AI, DeviantArt, or Stable Diffusion, among others, compile, store, combine, and generate images using trained algorithms and neural networks. These AI systems can learn patterns and visual characteristics from a dataset of images or files previously collected.

These images or files are processed from a wide variety of sources, such as public databases, online platforms, and private collections, which are used as training data to teach AI models to recognize and understand different visual elements, such as objects, shapes, textures, and colors. Once the dataset is collected, it is stored on appropriate servers and storage systems for later access and processing. This allows AI algorithms to use techniques such as overlay, fusion, and manipulation of existing images to generate new visual compositions resembling those found in the original databases.

For this reason, it is crucial to focus on the sources (databases) that feed these systems.
Quoting David Holz (founder of Midjourney), it can be stated:

> *There is really no way to get a hundred million images and know where they come from. It would be great if images had embedded metadata about the copyright owner or something. But that's not a thing; there's no registry. There's no way to find an image on the internet and then automatically trace it back to an owner and then have some way of doing something to authenticate it.*

This has become evident through cases like that of Getty Images, which came to light in 2018. During this year, it was discovered that some AI applications were generating images using Getty Images' vast image repository without paying for usage rights. In the processing for generating new images, the distinctive watermarks of Getty Images were not properly removed, leading to the appearance of these marks in the resulting images (Figure 1). The main concern was that these AI-generated images would be widely shared online, potentially leading to massive infringement of the copyright of the original sources.

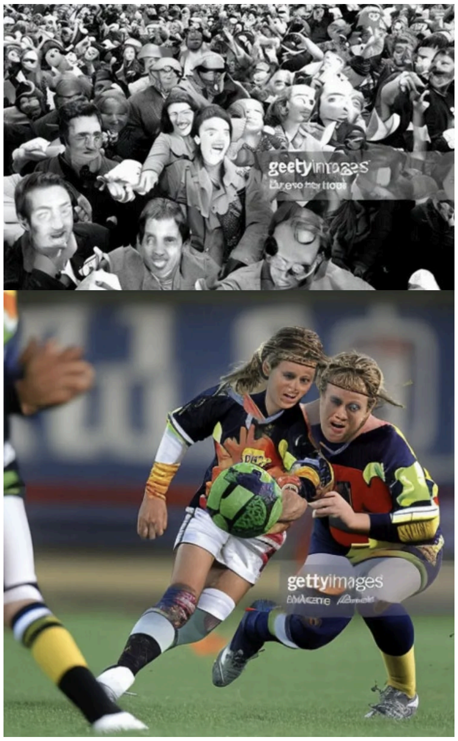

Figure 1. Examples of images created by AI with the Getty Images logo.

This is how the issue of ownership and copyright becomes increasingly relevant in contemporary times, leading many organizations responsible for the management, preservation, and conservation of archives to be cautious and resistant to the use and subscription to massive data processing services. They aim to prevent their databases from being used for the creation of repositories and AI training datasets without prior authorization, as copyright regulations remain unclear.

For instance, there are platforms like Kaggle that provide datasets for training AI systems. These datasets contain collections of hard-to-find materials, such as film records from various sources, in both black and white and color, spanning various years, with rights seemingly released or authorized for different types of uses. This highlights that many of these systems lack specific databases to train their algorithms, usually belonging to archives of public or private institutions that they don't have access to, becoming desired objects, as is the case of film records. [3]

The situation has become more complex as AI processing systems have become more sophisticated and addressed their errors, making it challenging to identify the original sources of files used for generating new products.

The ease with which these systems can create new content immediately and efficiently, solely through a prompt (a sequence of words, instructions, or commands), has led to the proliferation of a large number of new images or digital content. This raises questions about the authenticity of these productions, making it more difficult to distinguish the real from the fictional. This has been evident in images simulating non-existent humans (Figure 2), fake historical or news events (Figures 3 and 4), and even the recreation of expanded works of art derived from the original, adding new characters, objects, and extending the originally created scenes (Figure 5).

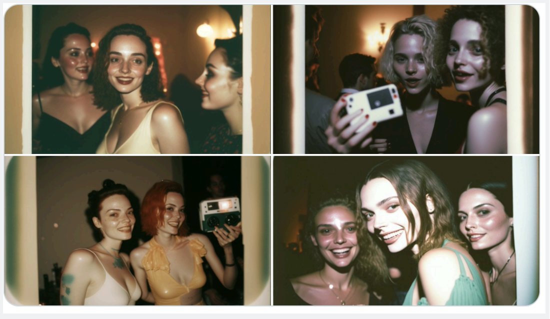

Figure 2. Photographs of non-existent people created with AI
Source: Miles [@mileszim] (2023). [4]

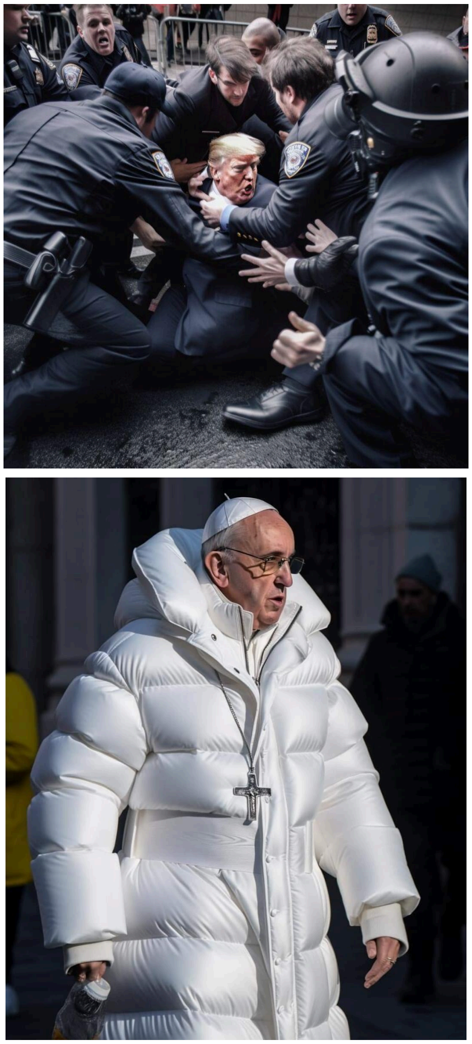

Figure 3. Fake images of the arrest of Donald Trump (Top) and Pope Francis (Jorge Mario Bergoglio) with a rapper look (Bottom), created with AI in Midjourney

V5, Eliot Higgins.

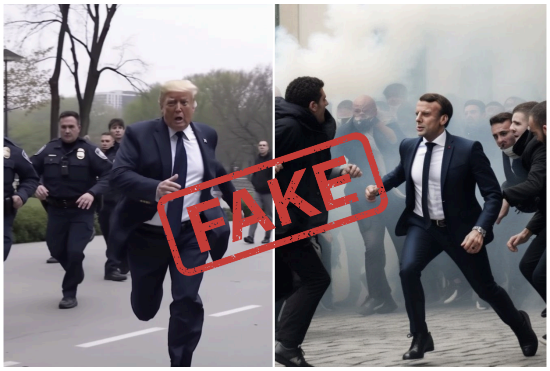

Figure 4. Fake images of Donald Trump (left) and Emmanuel Macron (right) created with AI
Source: Christine Vanden Byllaardt.

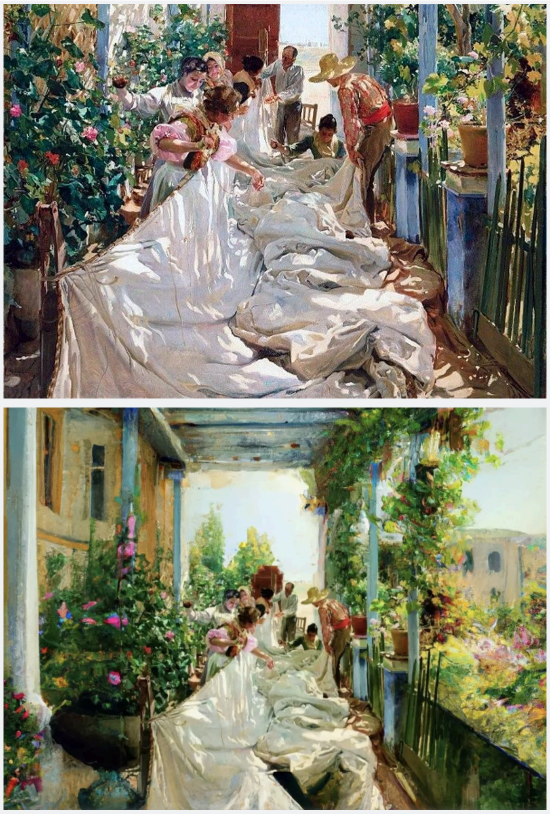

Figure 5. Joaquín Sorolla, Painting "Cosiendo la vela" (Sewing the Sail), 1896 (Top). Image created by AI from the original painting (Bottom).

On the other hand, nowadays, there is no regulation that covers everything originating from AI, and it has been legally declared that all content generated by AI is not subject to copyright.

This has made it clear that these new technological development systems are generating new dynamics that allow for orphaning of archives and establishing new categories to denote what is orphaned.

Firstly, in terms of processing itself, images are pulled/extracted from their original collections (public or private), and in this process, it can be established that they become orphans from their sources and original context since it is almost impossible to trace them and establish their provenance.

Secondly, all these new databases generated through AI, not covered by copyright, have become new collections of orphaned documents that belong to no one. This raises questions about how machine-created data is stored, the future of these new creations, and what legal handling should be given to them regarding their management, preservation, and conservation.

For example, in many of these AI applications, only the 50 most recent generations are recoverable unless users manually save the images individually, which are then stored in a large pool of images that cannot be searched (Figure 6).

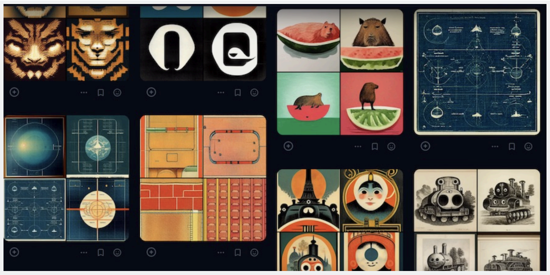

Figure 6. Example of visualization of a collection of images created in AI applications.

In conclusion, the discussion expands regarding the notion of orphaned archive, the mechanisms of orphaning, and the significance of memory in today's context. This highlights the challenges and ethical implications associated with the use of AI in the generation and manipulation of images, calling for guidelines and regulations to ensure copyright, responsible use of images, and the concept of memory. It is important to address that such systems are subject to inherent biases based on the data they are trained on, which can impact how historical information and collective memory are presented. Similarly, the information manipulation capability raises questions about the reliability of memories and the authenticity of history (Figure 7).

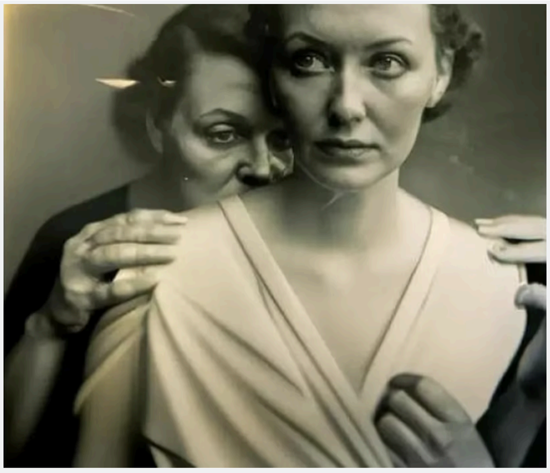

Figure 7. Pseudomnesia. Photograph created with AI that won the creativity award at the Sony World Photography Awards 2023. Boris Eldagsen.

References

[1] Foucault, M. (2005). The Archaeology of Knowledge. Siglo XXI.

[2] Kittler, F. A. (2018). The Truth of the Technical World: Essays on a Genealogy of the Present. Fondo de Cultura Económica.

[3] Kaggle. (n.d.). Morocco 1930's footage. https://www.kaggle.com/datasets/anashamoutni/morocco-footage-from-1930

[4] Miles [@mileszim]. (2023, January 13). Midjourney is getting crazy powerful—none of these are real photos, and none of the people in them exist [Tweet]. Twitter. https://twitter.com/mileszim/status/1613965684937224192?