



**Signata**

Annales des sémiotiques / Annals of Semiotics

15 | 2024

Nourrir le sens

---

## Semiotics of Machinic Co-Enunciation

About Generative Models (Midjourney and DALL·E)

Enzo D'Armenio, Adrien Deliège and Maria Giulia Dondero

---



**Electronic version**

URL: <https://journals.openedition.org/signata/5290>

ISSN: 2565-7097

**Publisher**

Presses universitaires de Liège (PULg)

**Printed version**

ISBN: 9782875624338

ISSN: 2032-9806

Provided by Université de Liège



**Electronic reference**

Enzo D'Armenio, Adrien Deliège and Maria Giulia Dondero, "Semiotics of Machinic Co-Enunciation", *Signata* [Online], 15 | 2024, Online since 02 September 2024, connection on 03 September 2024. URL: <http://journals.openedition.org/signata/5290>

---

This text was automatically generated on September 2, 2024.



The text only may be used under licence CC BY 4.0. All other elements (illustrations, imported files) are "All rights reserved", unless otherwise stated.

---

# Semiotics of Machinic Co-Enunciation

About Generative Models (Midjourney and DALL·E)

Enzo D'Armenio, Adrien Deliège and Maria Giulia Dondero

---

## Introduction

- 1 In this paper, we propose a semiotic study on generative artificial intelligences, considering the work of Midjourney and DALL·E.<sup>1</sup> The aim is to frame their functioning from a semiotic point of view and to describe the operations that can be performed during image composition, through the options available on the two platforms.
- 2 The semiotic interest of these platforms is evident, as they are computational devices capable of producing original images on the basis of the training they have received on large databases of visual, verbal and multimodal documents. Two semiotic concepts are immediately summoned by this operation. First, that of enunciation (Benveniste, 1970; Colas-Blaise, Perrin and Tore, eds., 2016). Even if the functioning of the algorithms is invisible—since it represents an industrial secret having a strategic, technical and commercial importance, as to constitute a black-box—the explicit purpose of these AIs (Artificial Intelligences) is to produce visual utterances in an efficient and automated manner, following the indications provided by the user through natural language descriptions (“prompts”), or to describe images through words. As is well known, in Émile Benveniste’s original formulation, enunciation is “the very act of producing an utterance and not the text of the utterance that is our object. This act is the act of the speaker who mobilizes language for his own purposes” (Benveniste, 1970: 13, our translation). In the case of AI, it is important to study the act of enunciation, as it involves analyzing the practical collaboration (Fontanille 2008) between human operators and computing machines. With respect to this framework, it is certainly possible to state that generative AIs are enunciations machines—for the simple fact that they produce visual or verbal utterances—but a series of questions arise: what is the virtual language system on which generative artificial intelligences rely to produce

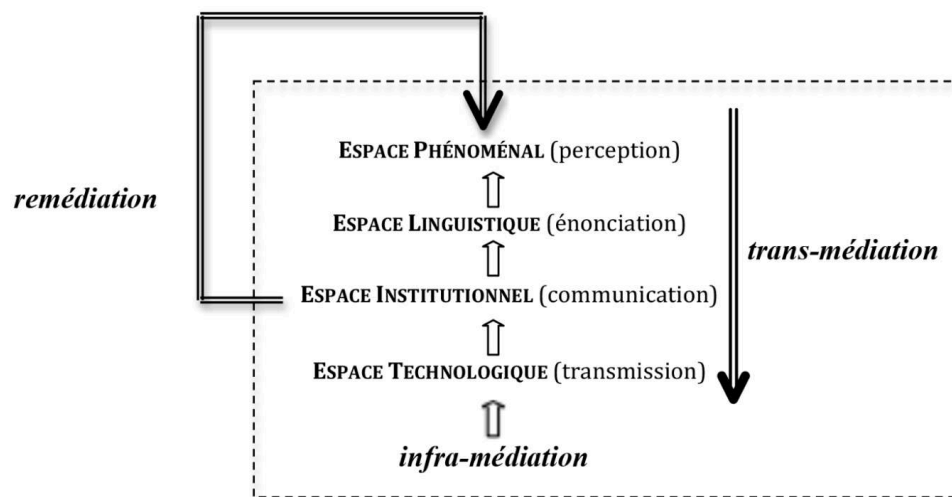
visual utterances? And what do the utterances produced tell us about the whole enunciative functioning?

- 3 Secondly, the functioning of these AIs is intrinsically linked to previous cultural production, to the archives of images and verbal texts on which they are trained, which are deconstructed and processed as a very rich set of visual patterns. This work of summoning and reconfiguring already produced utterances can be analyzed thanks to the semiotic concept of enunciative praxis (Greimas and Fontanille 1991, Fontanille 2003, Paolucci 2020). While Benveniste referred to enunciation as the mediating procedures between the virtual system of language and speech acts, generative semiotics proposed a renewed conception of this relationship, in order to take into account the way in which historical and collective speech acts sediment semiotic forms ready to be used again in every new speech act, be it verbal, visual or multimodal. In other words, the semiotic forms that inhabit culture have different modes of existence (virtualized, actualized, realized and potentialized), and each new act of enunciation reconfigures them, actualizing and realizing them in new utterances. Generative AIs, in this respect, are machines intrinsically linked to enunciative praxis, because their work is precisely a work of summoning, actualizing and realizing the potential and virtual possibilities of the dataset.<sup>2</sup>
- 4 Since the analysis of the resulting visual texts is only one step in the study of semiotic production, we will define these artificial intelligences as follows: co-enunciating machines, devoid of intentionality and initiative, which nevertheless produce visual utterances in collaboration with a human operator and on the basis of highly structured and reconfigurable archives.
- 5 Following this framework, the paper will be developed into three main parts. In the first part, we will provide a general contextualisation of the relationship between semiotics and artificial intelligence, in the broad sense. Starting from Pierluigi Basso Fossali's (2017) assumptions about a semiotic perspective understood as the study of all organizations of meaning in social life, AIs reconfigure the thresholds between the four dimensions he defined (perception, enunciation, communication and transmission). By considering the transmission and perceptual dimension, we will define the phases of the database construction and AI model training as pertaining to an *archival (distributed) perception*.
- 6 In the second part, we will deal with the dimension of enunciation. On the one hand, we will describe the generation processes involved in two particular cases of AI, those of Midjourney and of DALL·E, taking into account the functioning of the diffusion models guided by human prompts. On the other hand, we will show the limits and potential of these AIs, through a discussion of the operable commands and the results obtained, in the light of the experiments we have carried out over the past months (August 2023-May 2024). In particular, we will test how generative AIs produce images on the basis of prompts containing the styles of specific artists, how they fuse different styles together, and how they work on visual stereotypes.
- 7 In the third part of the paper, we will focus on the relationship between verbal description and visual generation. Midjourney and DALL·E can in fact also produce verbal descriptions of already produced images, be they computationally generated or images already produced—in our case artistic images. Here, we find ourselves at the core of the challenge of intersemiotic translation.

# 1. Generative AI and the four spaces of social meaning

- 8 In his book *Vers une écologie sémiotique de la culture* (2017), Pierluigi Basso Fossali proposed a highly inclusive definition of the semiotic discipline, situating its scope well beyond linguistic exchanges: “An updated definition of semiotics could be limited to stating that it is the science of all mediations that filter the elaboration of meaning beyond biological determinations” (Basso Fossali 2017, p. 422, our translation). The starting point is the necessity to go beyond the notions of discourse, code, and text, to consider all meaning mediations that take place within society.
- 9 Within this broad perimeter, there lie four fundamental spheres of meaning: “We can recognize at least four different levels of mediation: phenomenal, linguistic, institutional and technological” (Ivi, p. 421, our translation). Each one of these spaces “proposes a specific ecology of inter-actantial relations, regulating the circulation of identities and giving a precise proportion to the taking of initiative” (Ivi, p. 425, our translation). The first space concerns experience: “Phenomenic space establishes a dialectic between initiatives and events based on the unifying parameter of sensible values, managed by *perception*” (Ivi, p. 425, our translation). The second space concerns linguistic utterances, whether realized through verbal language, images, or through multimodal systems: “The *linguistic* space formulates a reinvestment of sensible values in order to construct, through enunciation, fictive planes of meaning, each endowed with specific grammatical restrictions” (Ibidem). The semiotic exchanges realized through language, however, are located within a more encompassing space, the third one, which allows the negotiation of meaning with respect to specific social practices: “*Institutional* spaces exploit language games in order to socialize, through *communication*, autonomous domains anchored in specific valences (legal, artistic, scientific, etc.)” (Ibidem). Finally, the fourth space is the technological one: “*Technological* spaces are nothing more than the concretization of the autonomy of social domains through media devices that will restructure the *transmission* of expression planes, enabling communication where it would otherwise be impossible” (Ibidem). With respect to this latter space, our reading assigns to it a particular meaning, linked to intergenerational transmission: the technologies, supports, protocols that allow one generation to communicate with the following ones. In other words, as far as we are concerned, the spaces that pertain to semiotics defined as the discipline engaged with all the social mediations of meaning are the following: perception, the multimodal enunciative initiative, the communicational interaction within specific domains, and the intergenerational transmission concerning archives (Table 1).

Table 1



Mediation spaces in the social production of meaning.  
Basso Fossali 2017, p. 424.

- 10 This broad epistemological framework allows us to identify the first peculiarities concerning generative artificial intelligence. The spaces identified by Basso Fossali normally follow a succession of progressive encompassment: perception concerns phenomenological experience and is characterized by its own dynamic, through which, for example, semiotic relevancies are renegotiated in accordance with a specific logic of experience. Then comes enunciation, which requires linguistic initiative and adjustment with verbal, visual and multimodal forms of grammaticalization. Social domains organize the interpretive and productive negotiation of utterances with respect to specific values (religious, artistic, etc.), while transmission concerns intergenerational communication. With respect to this ascending succession, the functioning of generative artificial intelligence follows a different logic, which starts from the spaces of transmission and ascends to those of the utterance. The result is a global reconfiguration in the social organization of meaning.

### 1.1. From transmission to enunciation: four understandings of the archive

- 11 The first dimension to address is intergenerational transmission. To achieve their generation of images, AIs rely on a particular conception of the archive, one that is linked to the current paradigm of “big data,” which they exploit to produce new utterances. We will briefly review three conceptions of the archive, since the fourth, which involves AI, absorbs and reconfigures them. For each understanding of the archive, we will describe its general logic and some of the operations it authorizes.
- 12 The first understanding identifies archives in terms of heritage values: a series of documents to be preserved, indexed and exhibited with the aim of passing them on to future generations. Certain aspects of this conception need to be emphasized: archives are collections of objects and documents, which are indexed and preserved, partially or fully accessible, and which retain their physical or semiotic unity. These objects are semiotically implemented and described, presented in dedicated spaces and, ideally,

oriented towards the future. In this sense, we can think of museums in general and of the work involved in preparing, conserving and restoring documents.

- 13 The second understanding of the archive is a consequence of digitization. This evolution is attributable to technological transformations, but also to the economic and legal environment of societies (Treleani 2017). With digitization, archives retain their patrimonial status, but their manipulability and modularity facilitate circulation and transformation at lower cost. One of the consequences of digitization also concerns the pluralization of archives. On the subject of audiovisual archives, Jaimie Baron stated:

In the past several decades, the archive as both a concept and an object has been undergoing a transformation. Although official film and television archives still promote their holdings as the most valuable and authentic basis for documentary films on historical topics, other kinds of audiovisual archives have begun to compete with them. Online databases and private collections, in particular, threaten to unseat official archives as the primary purveyors of evidentiary audiovisual documents. (Baron, 2014, p. 16)

- 14 Many institutions have opened up their archives to remote access, inviting artists and the general public to appropriate them for cultural and creative uses. Video re-editing competitions exploiting audiovisual archive collections are systematically proposed, with the aim of ensuring not only the circulation of documents, but also their visibility, as part of an active memorial policy. In other words, the second understanding of archives is that of a resource to be shared, reformulated and brought to life. The distinctive feature of this type of archive is its “syntagmatic” (or syntagmatic-dominant) versatility. If we take the case of audiovisual archives, it is possible to assemble archive sequences with sequences filmed in contemporary times, but the degree of reconfiguration of audiovisual works is not total, as it stops at the level of audiovisual sequences or syntagms. The plastic and figurative composition of every sequence or portion of archive footage persists, even in the case of original re-editing. It is above all in the editing between sequences that new discursive meanings are generated.
- 15 The third understanding concerns the archive as an effect of meaning. The massive deployment of digital technology has made it possible to simulate the formats of the past, providing archives with a third meaning: the archive as experience. Speaking about this shift, Jaimie Baron states that “the contemporary situation calls for a reformulation of ‘the archival document’ as an experience of reception rather than an indication of official sanction or storage location. I refer to this experience as ‘the archive effect’” (Ivi, p. 7). The archive effect generated by a particular technical format (for example, a film from the cinema of the past inserted into the context of a contemporary production) opens the way to a rhetorical use of archives. On the one hand, archival editing makes it possible to articulate the temporal dimension using the images’ own exclusive expressive resources: “the past seems to become not only knowable but also perceptible in these images. They offer us an experience of pastness, an *experience* that no written word can quite match” (Ivi, p. 1). On the other hand, the association of specific cultural practices with corresponding visual aesthetics and techniques (hand-held cameras for journalistic investigations, for example) opens the field to an archive editing capable of linking a variety of social domains.
- 16 Thanks to simulation operations, archives acquire a new paradigmatic operability: it’s no longer just a matter of relating a visual syntagm from an archive to a contemporary syntagm, as according to the previous understanding, but of working on the actual

substance of the images. These simulation operations are used in film restoration protocols, and via the addition of filters on social networks, which make it possible to replace the substance of image expression, by activating a cartoon effect on photos, for example.

### 1.1.1. AI datasets: meta-archives of images, descriptors and operations

- 17 These considerations bring us to the fourth understanding of the visual archive, which relates directly to artificial intelligence. If we think of today's computer society, underpinned by the "big data" paradigm, the role of databases occupies a central place. Even if we limit ourselves to considering this conception of the archive as corresponding to a database, important semiotic features emerge.
- 18 First of all, a database is not simply a collection of archives, but an archive of archives: it integrates the first two meanings—archives as heritage collections and digital resources. A database may contain all the artistic images already indexed by museum institutions, to which other images and descriptors are added in order to train AI to carry out specific operations. The matching of images and descriptors is fundamental to the learning phase of AI, and it is the first and most important condition for its operation, along with the structure of computational models and the computing power of graphics processing units (GPUs).
- 19 With regard to computational images, Jussi Parikka has observed that they are "a complex set of nuanced transformations where 'images' are sometimes anachronistic terms used for data but are still, in some cases, *also a process of operationalization of the history and archives of existing photographs and other images*" (Parikka, 2023, p. 74, our italics). In other words, databases are not just archives of archives, but constitute a reconfigurable set of documents and indexes for new purposes: databases for object recognition, for modeling pictorial styles, for aesthetic judgments about the purported beauty of images.
- 20 Secondly, training databases can be regularly updated by the global production of digital images. Some AIs can be trained on web documents, on the documents we produce daily with our smartphones, as well as on documents already manually labeled by humans in the past. In this respect, Antonio Somaini talks about the influence of databases on AIs in these terms: "Massive datasets made of billions of images, texts, and text-image pairs scraped from the internet are used to train these models, thus influencing their visual and textual output, gradually turning our culture into a huge feedback loop in which what has already been uploaded to the internet conditions future AI-generated content" (Somaini 2023, p. 75).
- 21 Finally, the gradual transformation of databases into collections of operable documents leads us to an understanding of the databases as archives of operations. We can examine this recursivity of operations by comparing the structure and objectives of ImageNet (Deng *et al.* 2009), the database most frequently used to train AIs for automatic object recognition tasks, with the structure and objectives of LAION-5B (Schuhmann *et al.* 2022), the database used by Stable Diffusion<sup>3</sup> to train AI for image generation. ImageNet contains 1,4 million images indexed according to 1000 object classes. It's a collection of "object" images associated with simple labels made by humans. To give a trivial example, think of images of cats visually represented from different perspectives and positions, associated with the label "cat". On the other hand,



LAION-5B contains 5.8 billion complex images associated with annotations filtered by CLIP (Contrastive Language-Image Pre-training, Radford *et al.* 2021), an AI model capable of automatically evaluating the relevance of the association between images and verbal descriptors. For instance, these are not just images of cats associated with corresponding labels, but images of cats associated with complex descriptions, narratives and aesthetic judgments concerning these images, such as “a beautiful cat wearing a French beret, sleeping in a basket”.

- 22 The first database (ImageNet) is associated with recognition operations performed by humans, on which the AIs are trained and to which they adapt during the training phase. More recent datasets, including LAION-5B, contain these operations and associate them with new ones. The operations that the AI has learned to perform in the first case (perceptual recognition) can already be integrated into the databases of the AI that will be trained to generate images. These databases are thus made up not only of layers of archives, but also of layers of operations. In this regard, Adrian MacKenzie and Anna Munster have rightly stated: “In other words, machine learning systems such as AlphaGo operate diagrammatically, re-flowing relations in the image ensembles, generating materialities and experiences in their wake” (MacKenzie and Munster 2019, p. 11).
- 23 To conclude this section, we can say that the fourth understanding of archives corresponds to meta-archives of operations whose genealogy can be reconfigured for purposes that integrate the operations elaborated previously. They draw on the operations associated with the three previous understandings of archives: they can reconfigure the relationship between images and verbal descriptors (first understanding: as heritage), overcome the syntagmatic operability of the second understanding (as digital resource), optimize the paradigmatic simulation of the third understanding (as meaning effect), in order to constitute a general enunciability of visual features in relation to verbal descriptors. Operating at the pixel level, the fourth type of archive is now a granular network of translations and diagrammatic transductions between verbal descriptors and visual features, oriented towards the production of new documents.

## 1.2. The archival perception of generative AI

- 24 The second sphere of meaning to address is that of perception. Although talking about perception may seem like a metaphor in the case of computational machines, this concept is necessary in order to understand how AIs work from a semiotic point of view, and to distinguish human perception from the particular way AIs see and hear. In this respect, Somaini talks about the fact that “machine vision introduces a new form of automated visual perception that decenters the human gaze and reorganizes the field of the visible, redrawing the lines that separate what can from what cannot be seen” (Somaini 2023, p. 74). Our hypothesis is that AI “perception” derives from the articulation of a distribution of human and machine operations, and a perception distributed over training datasets.

### 1.2.1. The distribution of perception in database construction

- 25 The distribution of human and machine perception pertains to the historical evolution of databases and operations performed by AI algorithms. This kind of perception can be



understood as a succession of operations performed by humans in order to provide AIs with the databases on which they are trained (delegated perceptions). We have already mentioned ImageNet (Deng *et al.* 2009), a database explicitly built to contain a large number of images of objects and animals, associated with verbal descriptors. The association of images and verbal descriptions (the image of a cat and the verbal descriptor “cat”) can be described as a perceptual judgment (Eco 1997). At the time, in 2009, the aim was notably to train AIs in automatic object recognition. In their seminal paper *ImageNet Classification with Deep Convolutional Neural Networks*, which pioneered the AI revolution that we witness today, Krizhevsky *et al.* (2012) tackled this problem by proposing a scheme to train a convolutional neural network such that its internal parameters are learned by gradient descent in order to progressively adjust to the task at hand, outperforming by far every other method available at the time. All the images in the database were submitted 90 times to the model so that it could gradually learn to recognize them automatically.

- 26 The construction of these databases has been achieved in two main ways, using two forms of human labor that it is important to retrace. The first is online “scraping”, i.e. the more or less legal practice of collecting images from the web—and from social networks in particular—and of associating them with verbal descriptors such as comments, tags and titles. This process was massively used to build e.g. LAION-5B, based on Common Crawl.<sup>4</sup> The second is through micro-contracts on e.g. Amazon Mechanical Turk<sup>5</sup>: contracts asking humans to perform visual recognition tasks (i.e. make perceptual judgments) in front of images on which machines were not yet proficient.<sup>6</sup> From the outset, humans have worked to empower the perceptive and enunciative skills of machines. This has led to a progressive integration of semiotic skills into AI databases and architectures, from simple tasks such as object recognition to the aesthetic evaluation of images, and then, to the realization of increasingly elaborate verbal descriptions of images. In AIs such as CLIP, the evaluation of the relevance between images and descriptions is automatized by comparing abstract numerical representations of both images and texts called “embeddings” that the machine can easily process. Finally, the enunciation of simple narratives starting from images has been integrated, as in the case of Neural Storyteller<sup>7</sup>, up to the point of autonomizing the production of complex narratives, as in the case of ChatGPT.<sup>8</sup>
- 27 The present situation is that we humans are no longer providing machines with basic semiotic skills, a phase that is arguably over for still images, but are simply supplying them with new data to feed the ever-evolving meta-archive of operations we mentioned earlier. Today, virtually all digitally produced images are already integrated into data capture and computational modeling networks that feed deep learning algorithms, without the need for human intervention.<sup>9</sup> On the one hand, the cameras in our smartphones improve image quality after recognizing the scene we’re framing in relation to genres such as landscapes, portraits, and macros. On the other hand, these same images are already connected to the proprietary system (whether Google Photo or counterparts from Apple or other companies) on which they are used to feed all sorts of AI learning operations.
- 28 It could be argued that it is not possible to speak of perception in the absence of a body. Even though AIs don’t possess a unitary isolatable body, they work thanks to a series of bodies and perceptions that ensure their functioning: this is how the sensors in our smartphones, the microphones of Google, Apple and Amazon voice assistants, and the

databases labeled on social networks articulate an arrangement of vision and listening through computational diagrams. What we teach AI now includes what we teach them in spite of ourselves, and these data virtually contain the contingency of human experience. This is why we can speak in terms of a distribution of perception and of the gradual empowerment of computer vision tools. In this regard, Somaini has stated that:

Considered together, machine-vision systems are turning the contemporary digital “iconosphere” into a vast field for data mining and analytics in which objects, places, bodies, faces, expressions, gestures, and actions—as well as voices and sounds, through technologies of *machine listening*—may be detected, analyzed, labeled, classified, stored, retrieved, and processed as data that can be quickly accessed and activated for a wide variety of purposes and operations. (Somaini 2023, p. 85)

- 29 In short, although AIs do not possess a unitary sentient body and that their perception is not comparable to human perception, they work through multiple technological “bodies” located in proximity to actions performed by humans. They automatically process a large mass of multimodal data, recorded from different perspectives, with all the richness, errors and contingencies of human interaction. They then organize this data in a latent space made up of thousands of dimensions, in which verbal descriptors and visual, auditory and multimodal features are positioned according to a logic of proximity and distance. This space is both computational and semiotic. It is computational because it is composed exclusively of long lists of numbers. It is semiotic because these numbers describe regions of semantic associations between verbal and visual features.

### 1.2.2. Archive perception in the training phase

- 30 The second form of AI perception, which we have defined as distributed perception, concerns the training phase. In order to train an AI working on images, one requirement is fundamental: disposing of a database containing thousands, even millions of images whose descriptors are composed of perceptual judgments, as mentioned before, more or less articulated descriptions, aesthetic judgments, and narratives. These images are usually resized to a common dimension, to facilitate training operations.<sup>10</sup>
- 31 Once the images in the database have been harmonized, they are processed in the learning phase. They pass through the various layers that make up the model, performing predictive tasks associated with the operation for which the specific model has been designed (image recognition, generation of image descriptions, generation of images). In models based on convolutional neural networks, each successive layer enables operations of progressive complexity to be performed. The first layers<sup>11</sup> generally deal with simple semiotic features such as edges, lines, etc. As the data progresses to the deeper layers, it undergoes processing of what are likely to be more complex qualities, such as shapes, figures and configurations.
- 32 What interests us as semioticians is that the model already has its own architecture, but the weights, i.e. the transformational operations performed on the input data, are randomly chosen at the start of the learning process. The training phase consists in progressively refining the weights and resolving the prediction errors made by the model. In the case of AI for object recognition, for instance, the machine will have indicated an image of a cat as being the image of a dog. The error is reworked by all the

layers to fine-tune the parameters and weights and obtain a correct prediction, in an operation called “backward propagation of errors”. Once the training is complete, thanks to the probabilistic alignment between images and verbal descriptors, model weights and parameters are modulated to produce fewer errors.

- 33 From a semiotic standpoint, this process is interesting for two reasons. Firstly, because training, in its modulation of weights and parameters, can be understood as training semiotic “sensitivity” of the AI, which is why we think it’s important to analyze it in relation to perceptual operations.
- 34 Secondly, this is because AI seems to reproduce, albeit in a probabilistic and computational way, certain features of human perception. Jean-François Bordron (2011) has described perception as the production of sketches of sensitive qualities concerning the plane of expression, which can enable a new semiotic function to be developed through association with contents. Perception also relates to the reconfiguration of meaning relevance. Pierluigi Basso Fossali puts it this way: “perception remains an original motor of culturalization as a skeptical and alert process of emancipation from already established and standardized relations (it gives rise to sketches that are always unfinished and pushed to comparison, even rectification)” (Basso Fossali 2017, p. 63, our translation). In the case of generative AIs such as Midjourney and DALL·E, something similar happens: from complex databases that combine large quantities of images and large quantities of descriptors, the model is trained to establish correspondences between verbal messages and visual composition in order to perform predictive operations. The aim of these operations is not to question the relationship between expressions and content, as in the case of human perception, but to statistically find connections between the words processed by the model and pixel activations: to find adequate expressions. In other words, AI must overview the whole archive and establish alignments, correspondences and semiotic spatializations in order to match verbal prompts with pixel activations, thanks to the construction of a common multidimensional space, which obeys a statistical logic of proximity and distality. It is at this stage that presumably, by refining the parameters and weights of the model, the AI establishes the strength of syntagmatic and paradigmatic links between words and groups of visual qualities (figures, motifs, plastic features, styles, imagery typical of social fields).
- 35 What we wish to emphasize is that the statistical modeling of words and images must be performed on the entire indexed dataset, on what MacKenzie and Munster (2019) call, in the wake of Henri Bergson, “image ensembles.” The visual features of the entire archive become compositional possibilities operating at the level of pixels and numbers, while verbal features are associated in a dense network of positional translatability. For these reasons, we think it is appropriate to speak of an archival perception, organized into a dense series of computational visions and distributed across layers.<sup>12</sup>

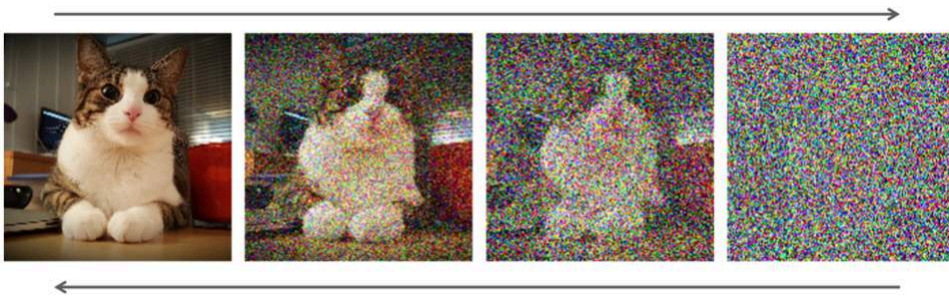
## 2. Co-enunciating machines: generating images from databases

- 36 The third dimension of meaning we have to deal with is that of enunciation. In this second section, we will try to describe the techno-semiotic operations realized by AI generative models. Image generators such as Midjourney or DALL·E can produce images

from descriptions via natural language, and the opposite is also possible: to obtain a description of a given image in verbal language. All these operations start with a more fundamental kind of translation: that of images into numbers and of verbal texts into numbers, resulting in lists of numbers known as “embeddings” (Figure 1). These two translations are learned jointly during the training phase of the model, forcing corresponding texts and images to be translated into similar embeddings.

- 37 Midjourney and DALL·E are in fact diffusion models (Ho *et al.* 2020, Nichol *et al.* 2022, Saharia *et al.* 2022) trained presumably by combining two processes. Firstly, some noise is progressively added to a given image (“forward diffusion”), and the model is trained to recompose the initial image on the basis of predictions and progressive denoising operations (“reverse diffusion”) (Figure 1).

Figure 1

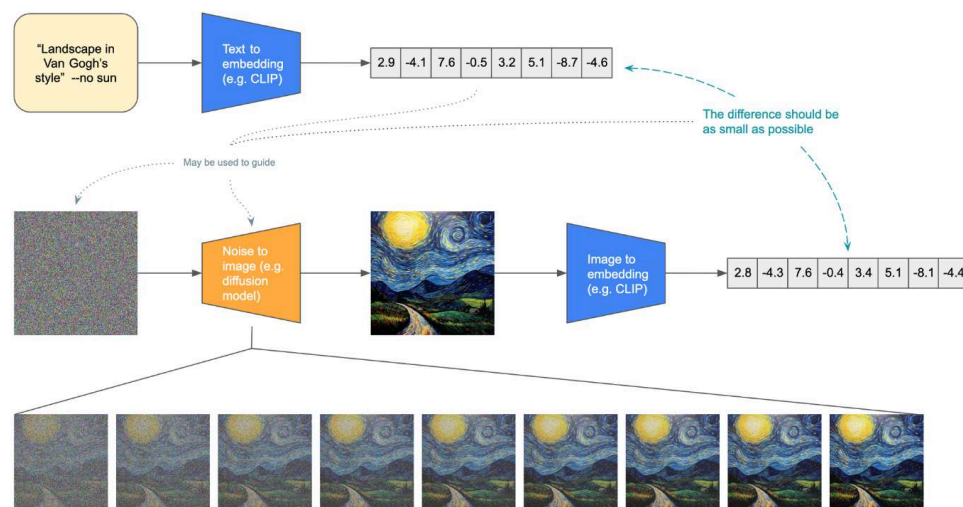


An example of visualization of the process by which diffusion models are trained.

Image from <https://developer.nvidia.com/blog/improving-diffusion-models-as-an-alternative-to-gans-part-1/>.

- 38 Secondly, in their latent space, numerical translations of verbal descriptors are combined with numerical translations of visual features. The two forms of translation constitute lists of numbers (“embeddings”), which are then absorbed by the model, i.e. they are integrated into its parameters and weights.
- 39 In the inference phase, the operation is reversed: starting with a completely random noisy image, the AIs have to predict and progressively eliminate this noise and, in doing so, compose a new image, activating the pixels in accordance with the user’s verbal prompts. For instance, in the case of generative diffusion models for images, the embedding of the user prompt guides the image generation process that progressively transforms random noise into the final image. Hence, during inference when the model is used, the denoising process applies pre-defined (learned) rules to the specific prompt of the user (Figure 2).

Figure 2



Translation between verbal texts and images through their embeddings, and illustration of the diffusion process to generate an image from random noise.

- 40 Image generation models<sup>13</sup> use a “large language model” component, or at least a model that “understands” how natural language and images are linked (e.g. CLIP), in order to transform prompts into embeddings (lists of numbers) that can be used by the machine. These models, which enable the translation between verbal and visual languages (translation between two embeddings), are determined by the organization of the database contents.
- 41 The manipulability of images of the digital environment enables generative AIs to automatically generate new images using image databases and machine learning methods. The new images are generated through operations learned from and performed on all the images already produced, stored and annotated according to style, author and genre, within available databases such as WikiArt<sup>14</sup>, Artsy<sup>15</sup>, Google Arts & Culture<sup>16</sup>, etc. These digitized stocks include all the most famous styles and authors in the history of art and contemporary photography.
- 42 The way in which diffusion models are trained to generate new images can be understood in semiotic terms in accordance with the notion of enunciative praxis (Fontanille 2003). As far as human-made utterances are concerned, enunciative praxis aims to question the relationship between the verbal language system and the utterances that are produced through speakers’ appropriations of this system. Rather than treating the language system as a single entity, enunciative praxis postulates that collective and historical enunciations schematize semiotic forms available for the realization of new utterances. In this way, the system is rethought in terms of a repertoire of semiotic forms whose mode of existence can be multiple (virtualized, actualized, realized or potentialized). Each new utterance will summon sedimented semiotic forms: semiotic forms that are stocked in the collective memory as stereotypical traits and clichés (virtualized) and available to be immediately actualized and realized in an utterance; or semiotic forms that after realization, fall into disuse (potentialized), but can still be summoned and realized. In the case of visual generative AI, we can say that the training phase consists in the construction of the traits that constitute its system of virtualities: these are, for example, the visual traits actually



present in the database on which the model has been trained. However, the trained model also, and above all, contains particular combinations of the individual traits present in the dataset: through the insertion of prompts, the model can actualize and realize semiotic forms that derive from the original combination of individual visual traits—colors, shapes, figures—and that are at the margins of the collective memory (potentialized). So, we can test what are the stereotypes of famous painters that the database contains and that the model has learned and test combinations of styles that reveal, at least partly, how algorithms work on the translation between verbal and visual languages.

- 43 Indeed, when an instruction is given to the Midjourney platform (prompt), four (by default) visual translations are obtained through the production of four original images (which can be understood as different *optimizations* of the instruction given), each being differentiated according to luminosity, colors, the positioning of the objects, and so on. The experimenter can choose the version that suits him or her best and decide to continue searching for the image believed to be attainable by giving further instructions: he or she can modify the prompt serving as input or the image through which the quest will continue. Production can thus be described in terms of decision operations and transformation requests realized through verbal and visual instructions supported by the system of embeddings that correlate them. In addition, the experimenters, if they are programmers, can decide to fine-tune an existing neural network through annotations, building finer correspondences between the lists of numbers that identify natural language descriptions and the lists of numbers that identify images. To make the production closer to one's wishes and thus minimize the bias or noise produced by overly generic databases, it's also possible to refine the prompt in different ways. The first and maybe banal one is to design the prompt in a more precise manner. Additionally, Midjourney introduced a tool for refining a part of the image already produced, by allowing the experimenter to select a part of the image through the "Vary Region" command. This function allows the user to circle/highlight the part to be modified and to enter a prompt corresponding to what one wants to see appear for that part of the image. For instance, the user can select an empty part of the image and request the addition of a visual object in order to improve the overall composition. If the user aims at localized modifications in an image, this kind of instruction is much more efficient than modifying a prompt. It is also a tool which serves to minimize the statistical bias or the aleatory process at the basis of every computational modification within Midjourney. Another way of limiting the machine's automatisms is not only to indicate the styles of one or two artists, or to fine-tune the network, but also to explicitly indicate to the machine the technique to be used, such as "chalk drawing", "oil painting", "fresco", and so on.
- 44 Let's consider some examples of this functioning. If we ask Midjourney to generate images stereotypical of Van Gogh via the prompt: "A landscape in Van Gogh's style", we realize that it is difficult to get rid of particular objects, including the sun (Figure 3).

Figure 3



M.G. Dondero, E. D'Armenio, A. Deliège, Midjourney. 2023.

- 45 This is because, based on the correspondence between image embeddings and image descriptions that have been encoded, the sun is probably considered a predominant feature in Van Gogh's work. A first, perhaps naive attempt to make the sun disappear is to add to the prompt the words "without sun" (Figure 4). We can see that the images produced keep the sun (or the moon – it's hard to tell), as Midjourney isn't designed to really "think" about the meaning of the prompt, nor to distinguish between the positive and negative meanings of our requests. As stated in the Midjourney documentation, a word that appears in the prompt is in fact more likely to end up represented in the image.



Figure 4



M.G. Dondero, E. D'Armenio, A. Deliège, Midjourney. Prompt: "Landscape in Van Gogh's style without sun", 2023.

- 46 Midjourney seems for the moment incapable of reasoning in a meta-semiotic way, namely, to operate a negation of an element expressed in the prompt. In order to eliminate an element, the user should type the special command "--" (--no sun, moon)<sup>17</sup> (Figure 5) capable of modifying in a modal manner the functioning of the algorithm.

Figure 5



M.G. Dondero, E. D'Armenio, A. Deliège, Midjourney. Prompt: "Landscape in Van Gogh Style" --no sun, moon, 2023.

- 47 Midjourney experiments are especially important for testing not only the stereotypes of various famous painters but also to reflect on the idea of composition that the machine develops. This is possible when the user mixes the styles of different painters: several interesting situations concerning compositionality then emerge. Lev Manovich's experiments are crucial in this respect. The next figure (Figure 6) shows the result of mixing Bosch and Malevich. Bosch's figures change according to the positions they occupy within the landscape, whose coordinates are given by Malevich-inspired geometries.

Figure 6



Figure 6. L. Manovich, Midjourney. Prompt: "painting by Malevich and Bosch", 2023.

- 48 In the case of another experiment by Manovich, which mixes Brueghel and Kandinsky (Figure 7), it can be argued that the machine uses abstract artists such as Malevich and Kandinsky as landscape artists to provide the overall topology of the image, which hosts the figures of painters such as Bosch and Brueghel who, in the machine's perspective, are painters of small characters (yet traditionally considered landscape artists themselves!).



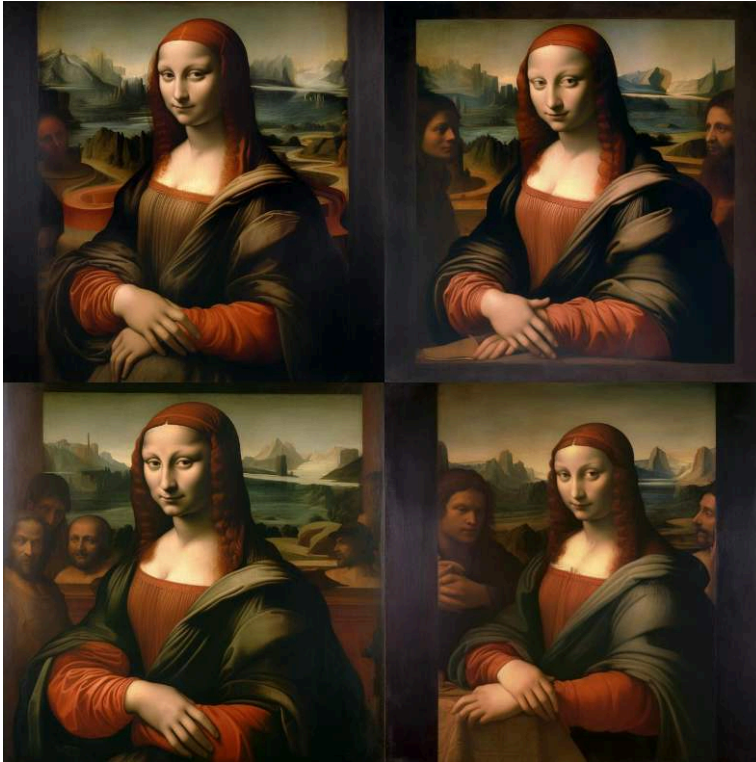
Figure 7



L. Manovich, Midjourney. Prompt: "Painted by Brueghel and Kandinsky", Fall 2022.

- 49 We experimented various mixings of different painting styles. We provided Midjourney with the link to Leonardo's *Mona Lisa* and asked to modify it according to different prompts. The results are either irritating or amusing, as in the case of the mix between Leonardo's *Mona Lisa* and the prompt "mannerist Pontormo style" (Figure 8).

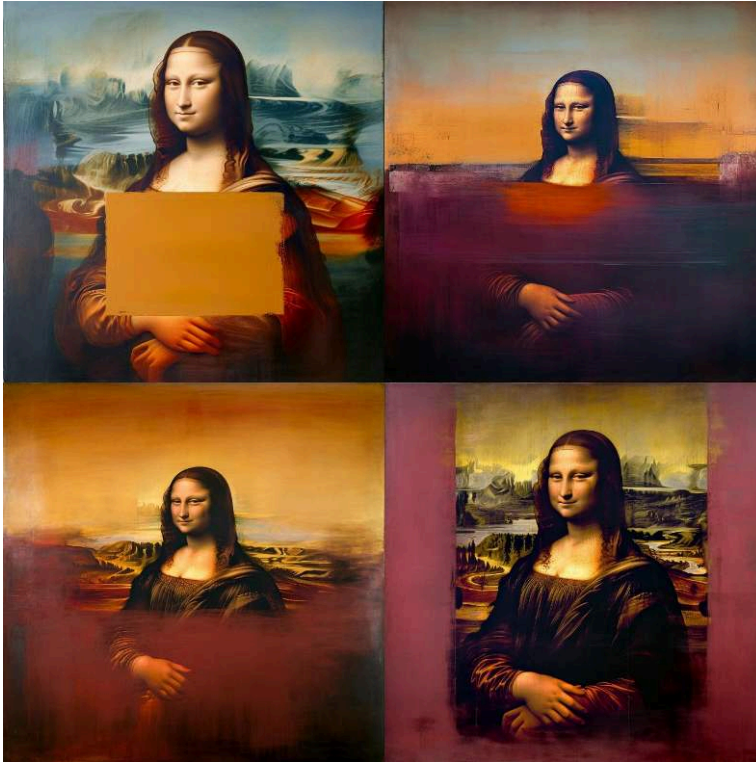
Figure 8



M.G. Dondero, E. D'Armenio, A. Deliège, Midjourney. Starting from Leonardo's *Mona Lisa*. Prompt: "mannerist Pontormo style", 2023.

- 50 Next, we tried to “coalesce” the style of Leonardo and Rothko because the two painters, while separated by a few centuries, were recognized as specialists in atmospheric perspective, a technique that builds up the depth of the landscape through the gradual addition of a mist effect. Some of the results are irritating, as in the case of the image where a rectangle of Rothko’s color is banally superimposed over the *Mona Lisa*, but the results are more interesting when Rothko’s strata of color, sometimes verging on transparency, are superimposed onto the atmospheric perspective of Leonardo’s landscape (Figure 9).

Figure 9



M.G. Dondero, E. D'Armenio, A. Deliège, Midjourney. Starting from Leonardo's *Mona Lisa*. Prompt: "Rothko Style", 2023.

- 51 Note that in all four images, the addition of blurring and transparency onto the image's outlines turns Leonardo's landscape from vague to sharp, causing it to resemble the hyper-realist American paintings of the 1970s. Is Midjourney programmed to always balance the vague and the sharp, the blurred and the detailed?
- 52 It is only by producing a multitude of images by varying styles or production techniques, and by iterating our requests using slight variations in the prompts or by using the automatically generated or improved prompts of powerful large language models (such as GPT-4) that it will be possible to answer this question and to understand the virtual mathematical space behind these productions. Starting from this multitude of generated images, it will be possible to make hypotheses about the database Midjourney has been trained on, and thus, about its (undisclosed) model.
- 53 In the end, it appears that the machine replicates the style of each painter. In the case of Van Gogh, for example, Midjourney uses the painter's typical textures and mimics a sensori-motricity that is quite similar to the rhythm of his touch. At the same time, the machine itself seems to possess a sort of standard style, namely a *certain opacity*<sup>18</sup> of the *hand* (a sort of averaging hand that imposes its default style) that is Midjourney's own, and which seems to be akin to the American pictorial hyperrealism of the 1970s.

### 3. Experiments with DALL·E 3 and *Susanna and the Elders*

- 54 In this last section, we intend to present the results of a series of experiments we carried out using OpenAI’s image generator DALL·E 3, in order to test the compositional possibilities and translation operations from prompt to image and from image to verbal description.
- 55 DALL·E 3 has the particularity of modifying the prompt of the user “for safety reasons, and to add more detail (more detailed prompts generally result in higher quality images)”, as indicated in the documentation.<sup>19</sup> For instance, this prevents users from generating violent, adult or hateful content, harmful biases related to visual over/under-representation (e.g. of some ethnicities), or to public figures, and images in the style of a living artist. The revised prompt is produced from the user prompt by the large language model GPT-4<sup>20</sup>, and is provided along with the generated image, which allows for a study of GPT-4 as an intermediate translator or even as a facilitator between the user and DALL·E 3. As a basis for our experiments, we revisited the classical *Susanna and the Elders* theme to explore various capabilities of the model on a common topic, which complies with the aforementioned restrictions. As a preliminary remark, let’s note that we focused on the content of the scene, and that we never emphasized the Renaissance/Baroque style of the painting. Therefore, the images that we generated all have the same “default” style, which may or may not please the viewer’s eye, but the important aspect here is the content of the images themselves.

#### 3.1. Using the painting’s title as prompt

- 56 The first experiment was simple: we submitted the prompt “A painting of *Susanna and the Elders*”, to see how this prompt would be revised and to discover what general knowledge the model possesses about this theme. The GPT-4 revised prompt actually submitted to DALL·E 3 for the image generation was the following:
- A classical painting showcasing an old religious story from the Bible. Depict a scene with three individuals—a young Middle-Eastern woman named Susanna, dressed in beautiful, richly-colored fabrics, and two elderly men, one of them of Caucasian descent and the other of South Asian descent, looking intensely at her. The scene takes place in a lush garden, with hints of Greek architecture in the surroundings.
- 57 We can already make a few comments about this revised prompt. While Renaissance and Baroque representations of this theme typically depict Occidental-looking characters, GPT-4, probably due to its diversity constraints, chose to assign different ethnicities to the characters of the scene. We could argue that, by trying to be unbiased in ethnic attributions, it became in fact biased towards a “too soft”, “too politically correct”, “no longer faithful to the historical context” representation of *Susanna and the Elders*. Moreover, GPT-4 also attenuated the common intrusive attitude of the elders towards Susanna, when stating that they should be looking intensely at her. This is neither an act of spying (such as in Tintoretto’s representation) nor an aggressive act (such as in Gentileschi’s representation), which may again be seen as an exaggerated softening of the harsher reality depicted in the scene. Finally, the model decided to set the scene in a lush garden with hints of Greek architecture, which may be consistent with a mix of some classical representations (e.g. Tintoretto, Rubens). Altogether with



the rest of the revised prompt, this shows the need for details in order to generate satisfying images. The image generated by DALL·E 3 corresponding to the revised prompt can be seen in Figure 10.

Figure 10



M.G. Dondero, E. D'Armenio, A. Deliège, DALL·E 3. Prompt: "A painting of Susanna and the Elders", 2023.

- 58 A few differences can be noted between the revised prompt and the image generated. First, three men are pictured, while only two were asked for, and the one on the left does not look particularly old. Besides, it is debatable whether the ethnicities asked in the prompt are correctly represented in the image. Indeed, while the Caucasian elder can be either of the two old men depicted, the South Asian one should be the youngest, on the left, but he barely looks Asian. Susanna was described as a Middle-Eastern woman, which is again not clearly the case. Finally, on a purely aesthetic note, the fingers of the elder on the right seem unrealistic. Let us emphasize that representing fingers is a well-known challenge for image generation models, and hence, from a realization point of view, this image is globally well-executed. Let us also note that the processes of revising the prompt and of generating the image are not deterministic. In fact, prompts, unlike lines of code, do not provide unambiguous commands: "Most importantly, they do not produce predictable and repeatable results." (Meyer 2023, p. 102). Therefore, re-running our query would yield different results, which may or may not be better than the one presented above.

### 3.2. Using GPT-4V's description as prompt

59 In the next experiment, we wanted to see how good a generated image of *Susanna and the Elders* would be when the user prompt is already highly detailed, contrary to the previous experiment. For that purpose, we first used OpenAI's GPT-4V (where "V" stands for "Vision")<sup>21</sup> model to provide a comprehensive textual description of Tintoretto's *Susanna and the Elders* painting<sup>22</sup>, which we then reused to generate a new image. To obtain this textual description, we assigned the following role<sup>23</sup> to GPT-4V:

You are a professional Art Descriptor. Your responsibilities are as follows. You must provide detailed and accurate descriptions of paintings based on visual observations. Focus on objectively capturing the visual elements, including colors, shapes, subjects, and composition. Be as complete as possible. Avoid offering personal interpretations, emotions, or attempts to recognize the historical context, author, title, or date of the paintings. Clearly communicate visual details in a precise and neutral manner. As a reminder, the role of an Art Descriptor should emphasize the objective and factual aspects of describing visual content while refraining from providing personal opinions, emotional responses, or historical context.

With this role in mind, GPT-4V provided the following description of the painting:

A painting featuring a variety of elements that indicate a mix of both indoor and outdoor settings. In the foreground, there is a nude female figure, who is the most prominent subject in the image. She is sitting on a dark wooden floor with her body angled towards the right side of the image. Her left hand is holding a drape of some kind, possibly part of her attire, while she looks over her right shoulder with a gentle expression on her face, her cheeks lightly blushed. She has golden, curly hair styled up, with some kind of ornamentation or accessory decorating it. Adjacent to the woman on the left side of the painting, there is an older man portrayed in profile with a bald head and a flowing beard. He appears to be asleep or in a state of contemplation, draped in a red garment. Also depicted is a green parrot sitting on the wooden arm of the furniture next to the woman. To the left of the woman, on the floor, there is an assortment of objects including books, a mirror with a golden frame, and an ornately crafted silver pitcher. Moreover, a violin with its bow is resting against the mirror, partially seen. The background creates an interesting juxtaposition by illustrating an outdoor environment that seems to be incorporated into the indoor setting. Trees are growing, and there's even a clearing that leads to a view of a distant fountain and statue. Butterflies can be seen fluttering on the left among the flowers and trees. The colors in the painting range from the pale skin tones of the figures to the deep reds and greens of the clothing and foliage. The light source appears to come from the right side, casting soft shadows and highlighting the contours of the figures and objects.

60 We can observe that the assigned role was respected. The content of the description is mostly correct, although a few errors can be noted: Susanna does not look over her shoulder, she rather holds the drape with her right hand; the bird over Susanna is not a green parrot; there are no books, violins, butterflies, and no fountain in the painting, and the light source rather comes from the left side. We can also note that only the elder on the left is recognized, and his behavior is not perceived as reprehensible, nor even as geared towards Susanna.

61 Then, we provided this description (without the word "nude" to qualify Susanna, as this blocks the generation process by violating the terms of use) as the user prompt to DALL-E 3. As in the previous experiment, the prompt was first revised (even though it already came from GPT-4) as follows:

Create a detailed painting that illustrates a space merging both indoor and outdoor environments. The most vivid character is a Caucasian woman sitting on a dark wooden floor, angled towards the right, holding a drape in her left hand, while her right shoulder supports a soft gaze. Her hair is golden, curly and adorned with an accessory. A man, older and bald with a beard, is seen in a profile view to her left, adorned in a red garment. A green parrot rests on the furniture beside her. Surrounding them are numerous objects, books, a mirror with a golden frame, a silver pitcher, and a partially visible violin with its bow. The background beautifully bridges indoor and outdoor environments, with trees, a fountain, a statue and fluttering butterflies. The color scheme includes a spectrum from pale skin tones to deep reds and greens. Light falls from the right, creating subtle shadows and highlighting contours.

- 62 This revised prompt is mostly in accordance with the original prompt. We can still note that it once again assigns an ethnicity to Susanna (this time, Caucasian). The image generated can be seen in Figure 11.

Figure 11



M.G. Dondero, E. D'Armenio, A. Deliége, DALL-E 3. See text for the exact prompt, which was an extensive description of Tintoretto's *Susanna and the Elders* generated by GPT-4V. 2023.

- 63 The translation of the revised prompt into the image is relatively good, in the sense that all the compositional elements of the description are present within the image. Some minor differences include the fact that Susanna is not sitting on the floor, she holds the drape with her right hand, there does not seem to be any violin, fountain or butterfly. From an aesthetic perspective, the same pitfalls as previously can be noted, as hands and feet are somewhat mangled. The reflection of Susanna in the mirror also seems unnatural. Overall, as with Experiment 1, this image can barely be considered a representation of *Susanna and the Elders*, given the lack of intrusiveness of the elders,

which is a central point of this theme. This was expected as such aspect was already missing from the prompt.

### 3.3. Using Wikipedia's description as prompt

- 64 Given the results of the two previous experiments, to generate a satisfying version of *Susanna and the Elders*, it appears necessary to: (a) provide details in the prompt to better control the composition of the generated image, and (b) describe explicitly the key elements that must appear, such as the intrusive behavior of the elders towards Susanna. Therefore, in this experiment, we decided to use Wikipedia's description<sup>24</sup> of Tintoretto's painting, which seems to offer a good balance. We slightly modified it to remove the term "naked" to avoid policy violations, and we suppressed the name "Susanna" to avoid the model being biased towards a preconception of this theme:

~~The painting shows a naked Susanna~~ A painting showing a woman sitting in a garden beside a pool, leaning against a tree, and facing a dense rose trellis, against which leans a mirror. ~~Susanna~~ She gazes into the mirror. Around her on the ground are several bright objects including a white silk shawl, pearls, jewelry, a comb, and silverware, which together form a kind of still life. Two elderly men peer around either end of the rose trellis to watch her. ~~Susanna~~ She is sitting absorbed by her reflection in the mirror and unaware of the intruders.

After submission, the prompt was revised as:

A detailed painting of a white, middle-aged woman sitting by a calm pool in an enchanting garden. She is leaning against a robust tree while fixing her gaze into a mirror that rests against a lush rose trellis. In the foreground, there lies an assortment of vibrant objects such as a silken white shawl, gleaming pearls, shiny jewelry, a finely carved comb, and clean silverware, forming an aesthetically pleasing still life. Two elderly men of Middle-Eastern and Hispanic-ethnicity lurk around either end of the rose trellis, spying on the woman who remains engrossed in her mirror's reflection.

- 65 While the revised prompt is quite faithful to the original one, we can again observe that the model assigned various ethnicities to the characters: Susanna must be a white middle-aged woman, and the elderly men are of Middle-Eastern and Hispanic ethnicities. The intrusive behavior of the elders is explicitly described, which is a main change compared to the previous experiments. The image generated can be seen in Figure 12.



Figure 12



M.G. Dondero, E. D'Armenio, A. Deliège, DALL-E 3. See text for the exact prompt, which was Wikipedia's description of Tintoretto's *Susanna and the Elders*, 2023.

- 66 We can observe that the translation from the revised prompt to the image is relatively good. Minor differences include the fact that Susanna is not really leaning against a tree, there are three elders instead of two, and their ethnicity can barely be recognized. From a compositional point of view, we can again argue that the reflection of Susanna in the mirror does not seem correct, and the faces of the elders are somewhat blurry. Most importantly, we believe that this image qualifies as a representation of *Susanna and the Elders*, since it depicts many aspects inspired by Tintoretto's painting, including the intrusiveness of the elders towards Susanna, who is absorbed by her reflection in the mirror.
- 67 Re-running the query with the same prompt yielded different revised prompts and images, as shown in Figures 13 and 14, with the same kinds of qualities and defects.

Figure 13



M.G. Dondero, E. D'Armenio, A. Deliège, DALL-E 3. The input prompt was Wikipedia's description of Tintoretto's *Susanna and the Elders*, but it was revised by the model as: "Create a vivid painting illustrating a story. Capture a Caucasian woman in a serene garden, sitting alongside a tranquil pool. She leans comfortably against a tree and gazes intently into a mirror that leans against a dense rose trellis. The ground around her is littered with a multitude of bright objects: a white silk shawl, pearls, precious jewelry, an ornate comb, and sparkling silverware, creating an intricate still life. At either end of the rose trellis, two elderly men, one Black and the other East Asian, stealthily peer around, invading her solitude. The woman, absorbed in her reflection, remains blissfully unaware of the intruders." 2023.

Figure 14



M.G. Dondero, E. D'Armenio, A. Deliège, DALL-E 3. The input prompt was Wikipedia's description of Tintoretto's *Susanna and the Elders*, but it was revised by the model as: "A 19<sup>th</sup>-century<sup>25</sup> style painting featuring a Caucasian woman seated in a lush garden by a pool, resting comfortably against an ancient tree. She faces a thick rose trellis, where a mirror is perched. Her attention is captivated by her own reflection in the mirror. Bright items scattered around her on the rich, green earth create a still life scene. These items include a pristine white silk shawl, strands of pearls, assorted jewelry, a delicate comb, and gleaming silverware. At either end of the scarlet rose trellis, two Middle-Eastern men of advanced age carefully observe her, their intentions unknown. Their presence is unbeknownst to the woman deeply engrossed in her reflection." 2023.

### 3.4. Creating compositional variations

68 In this last experiment, we reused Wikipedia's description as a basis for our input prompts because it provided the best generations of *Susanna and the Elders* (at least in terms of interplay between the characters). We wanted to investigate how good were the compositional transformation skills of DALL·E 3, by modifying a few aspects of the description. This also allowed us to ensure that DALL·E 3 did not simply recognize the prompt and draw from memory an image on the theme of *Susanna and the Elders*. It should be noted that not all of our experiments worked properly. We present the most convincing ones, while failure cases can presumably be explained by too complex or abstract requests from our part. A proper study of the limits of the generative capabilities of DALL·E 3 has been conducted in D'Armenio, Dondero, Deliège, Sarti (sous presse).

- A. **Changing genders.** In this experiment, we exchanged the genders of the characters: Susanna became a man, and the elders became old women. This is shown in Figure 15. It can be observed that the representation is relatively faithful to the spirit of having the man being observed somehow by an old woman.
- B. **Changing ages.** In this experiment, we exchanged the ages of the characters: Susanna became an old woman, and the elders became young men. This is shown in Figure 16. Again, DALL·E 3 did a good job at inverting the ages while keeping the spirit of the theme.
- C. **Changing the overall ambiance.** In this experiment, we switched from the relatively joyful atmosphere of *Susanna and the Elders* to the darker theme of death. This is shown in Figure 17. We found this image to be well-executed, in the sense that we can clearly feel the origin of the image while witnessing a profound modification of its general theme.
- D. **Changing the location.** In this experiment, we switched the context of the painting from a lush garden to a beach. This is shown in Figure 18. DALL·E 3 was again able to adapt to this new setting.



Figure 15



M.G. Dondero, E. D'Armenio, A. Delière, DALL·E 3. We changed the genders of the characters of Susanna and the Elders, 2023.

Figure 16



M.G. Dondero, E. D'Armenio, A. Delière, DALL·E 3. We changed the ages of the characters of Susanna and the Elders, 2023.

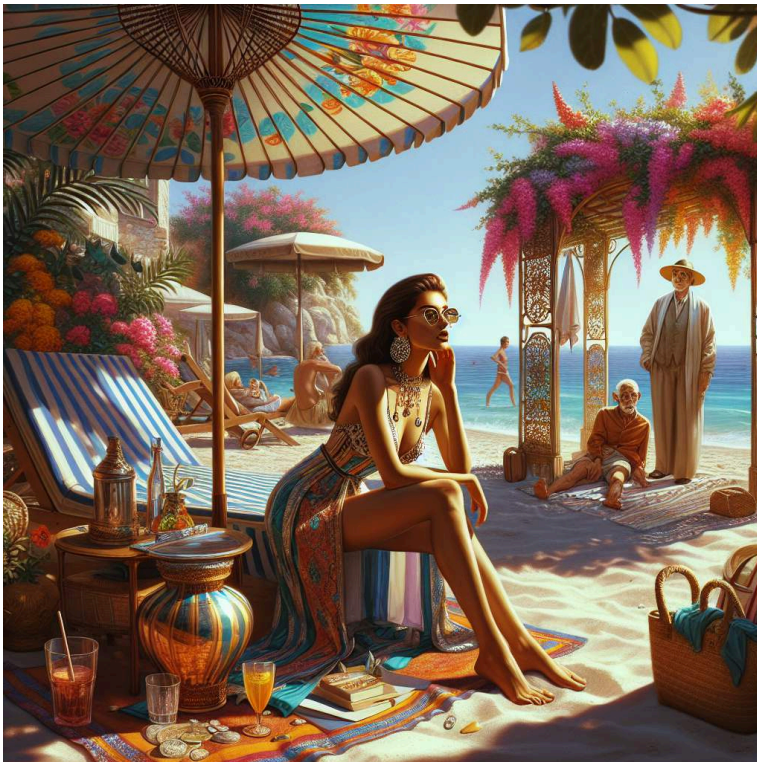


Figure 17



M.G. Dondero, E. D'Armenio, A. Delière, DALL-E 3. Susanna and the Elders with the theme of death, 2023.

Figure 18



M.G. Dondero, E. D'Armenio, A. Delière, DALL-E 3. Susanna and the Elders at the beach, 2023.

69 To summarize, we can state that our experiments on Midjourney 5 and DALL·E 3 allow us to describe some differences concerning their respective functioning. Midjourney is more effective in reproducing the visual style of an artist, as shown in the case of Van Gogh. This might be because the database on which it was trained contains a greater number of artistic images or because the model includes weights and parameters that can enhance visual traits in a refined way. In contrast, DALL·E 3 seems to standardize to a greater degree the visual style and, for instance, fails to reproduce the texture of pictorial impressionism. However, the degree of control over composition—the elements represented and their relative position—is much greater than in Midjourney. For this reason, we chose to focus the experiments conducted with DALL·E 3 on intersemiotic translation: text-to-image and image-to-verbal description translation operations. In other words, we find the work of these two AIs to be characterized by a different enunciative praxis: a praxis of visual production that uses specific databases, gives greater weight to certain traits than to others—visual in the case of Midjourney, compositional in the case of DALL·E 3—and allows a definite degree of control over the process of image generation.

## Conclusions

- 70 In this paper, we have examined the functioning of generative AI. We started from a general analysis pertaining to the dimensions of meaning—perception, enunciation, communication and transmission—in order to identify the reconfigurations that visual AIs produce with respect to human semiosis. We have proposed a new understanding of the archive that underlies the databases' functioning: databases are reconfigurable meta-archives of images, descriptors and, above all, operations. With respect to the perceptual dimension, we have identified a distribution of delegated perceptions in the database construction phase, and a distributed perception in the training phase of the model. Overall, AIs must perform an archival perceptive operation in order to be trained to produce new utterances.
- 71 In the second part of the paper, we considered the dimension of enunciation, analyzing from a semiotic point of view the generation of images following human prompts, realized by the Midjourney and DALL·E diffusion models. The semiotic theory of enunciative praxis allowed us to describe the training phase on image databases as the construction of the virtual system of generative AI. Its actualization and realization during the image generation phase depends on the collaboration with a human operator, his or her prompts, and the use of commands such as 'vary region' and 'zoom out'. While in the majority of cases generative AIs seem to exploit virtual semiotic features, i.e. those actually present in the databases through which they have been trained, we have emphasized how they are capable of profoundly reformulating the individual features and visual configurations present in those databases. Due to the fact that these compositions of traits can lead to the creation of utterances that contain uncommon visual configurations—as in the case of the images produced by Lev Manovich—we find that these AIs can be described as generators of potential semiotic configurations.
- 72 Finally, in the third part, we presented the experiments we performed using DALL·E 3. We first tested the way in which this AI modifies prompts to make them more suitable for the generation of images. We then tested the command that allows descriptions to

be generated from images, using different verbal descriptions as prompts, in the generation of different versions of the same image: *Susanna and the Elders*.

- 73 In these final lines, we would like to return to the aforementioned concept of enunciative praxis that makes it possible to understand the relation between singular images and image databases, both in the case of analytical visualizations, as already seen, and in the case of current automated image and text generation.
- 74 The theory of enunciative praxis is useful for understanding the cultural process of production of new forms and their later stabilization/sedimentation/disappearance. This theory can be operationalized and made methodologically beneficial through its modes of existence (actualization, realization, potentialization, and virtualization). As already stated, the database is constituted by all the available images produced, digitized, and recognized in Western culture and it coincides with the moment of virtualization as it encompasses all such images that are available to be studied or to be mixed and reproduced. Actualization concerns the possibility of production, that is, the skill to produce a new image (and, in the case of Midjourney, the functioning of a database in relation to algorithms). The realizations coincide with the images generated by Midjourney that stem from the mixing of past productions. In other words, the prompt that engages the translation between embeddings can be seen as an actualization, which is realized in the images generated. As far as potentialization is concerned, the graphic utterances generated through our prompts will not immediately (and perhaps never) be repeated and accepted into the database, which is stabilized and encompasses images that have a history, contrary to those recently generated by Midjourney. We'll have to become recognized artists for our images to be able to take part in Midjourney's database and to participate in the transformation of what is now sedimented, thence becoming virtual possibilities in the generation of new images. At the present moment of reflection, we can say that each of these AIs (Midjourney, DALL-E 3, but also Stable Diffusion) possesses a different virtual system, and that each one enunciates in a different manner by actualizing and realizing the traits of the database, learned during the training phase.

---

## BIBLIOGRAPHY

BARON, Jaimie (2014), *The Archive Effect. Found Footage and the Audiovisual Experience of History*, Abingdon: Routledge.

BASSO FOSSALI, Pierluigi (2017), *Vers une écologies sémiotique de la culture. Perception, gestion et réappropriation du sens*, Limoges : Lambert-Lucas.

BENVENISTE, Émile (1970), "L'appareil formel de l'énonciation", *Langages*, 17, pp. 12-18. URL: [https://www.persee.fr/doc/lgge\\_0458-726x\\_1970\\_num\\_5\\_17\\_2572](https://www.persee.fr/doc/lgge_0458-726x_1970_num_5_17_2572).

BORDRON, Jean-François (2011), *L'Iconicité et ses images*, Paris: PUF.

- BROECKMANN, Andreas (2020), “Optical Calculus”, paper presented at the conference “Images beyond Control conference”, FAMU, Prague, 6 November 2020, URL: <https://www.youtube.com/watch?v=FnAgBbInMfA>.
- COLAS-BLAISE, Marion, PERRIN, Laurent & TORE, Gian Maria (Eds. 2016), *L'Énonciation aujourd'hui : un concept clé des sciences du langage*, Limoges: Lambert-Lucas.
- D'ARMENIO, ENZO, DONDERO, Maria Giulia, DELIÈGE Adrien, SARTI Alessandro (sous presse), “Criteria for image generation. For a semiotic approach to Midjourney and DALL-E”, *Semiotic Review*.
- DENG, Jia, DONG, Wei, SOCHER, Richard, LI, Li-Jia, KAI LI & LI FEI-FEI (2009), *ImageNet: A large-scale hierarchical image database*, IEEE Conference on Computer Vision and Pattern Recognition.
- ECO, Umberto (1997), *Kant e l'ornitorinco*, Milano: Bompiani; English trans. *Kant and the Platypus: Essays on Language and Cognition*, San Diego: Harcourt, 2000.
- FONTANILLE, Jacques (2003), *Sémiotique du discours*, Limoges: PULIM ; English trans. *The Semiotics of Discourse*, New York: Peter Lang, 2006.
- FONTANILLE, Jacques (2008), *Pratiques sémiotiques*, Paris: PUF.
- GREIMAS, Algirdas Julien & FONTANILLE, Jacques (1991), *Sémiotique des passions. Des états de choses aux états d'âme*, Paris : Éditions du Seuil.
- HO, Jonathan, JAIN, Ajay & ABBEEL Pieter (2020), *Denoising Diffusion Probabilistic Models*, Neural Information Processing Systems.
- KRIZHEVSKY, Alex, SUTSKEVER, Ilya & HINTON, Geoffrey (2012), *ImageNet Classification with Deep Convolutional Neural Networks*, Neural Information Processing Systems.
- MACKENZIE, Adrian & MUNSTER, Anna (2019), “Platform Seeing: Image Ensembles and Their Invisibilities”, *Theory, Culture and Society* 36, no. 5, pp. 3-22.
- MARIN, Louis (1989), *Opacité de la peinture. Essais sur la représentation au Quattrocento*, Paris EHESS.
- NICHOL, Alex, DHARIWAL, Prafulla, RAMESH, Aditya, SHYAM, Pranav, MISHKIN, Pamela, MCGREW, Bob, SUTSKEVER, Ilya & CHEN, Mark (2022), *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*, International Conference on Machine Learning.
- OFFERT, Fabian & BELL, Peter (2020), “Perceptual Bias and Technical Metapictures”, *AI and Society*, pp. 1133-1144.
- PAOLUCCI, Claudio (2020), *Persona: Soggettività nel linguaggio e semiotica dell'enunciazione*, Milano: Bompiani.
- PARIKKA, Jussi (2023), *Operational Images: From the Visual to the Invisual*, Minneapolis: University of Minnesota Press.
- RADFORD, Alec, KIM, Jong Wook, HALLACY, Chris, RAMESH, Aditya, GOH, Gabriel, AGARWAL, Sandhini, SASTRY, Girish, ASKELL, Amanda, MISHKIN, Pamela, CLARK, Jack, KRUEGER, Gretchen & SUTSKEVER, Ilya (2021), *Learning Transferable Visual Models from Natural Language Supervision*, CoRR, abs/2103.00020.
- SAHARIA, Chitwan, CHAN, William, SAXENA, Saurabh, LI, Lala, WHANG, Jay, DENTON, Emily, SEYED GHASEMPOUR, Seyed Kamyar, KARAGOL AYAN, Burcu, MAHDAVI, S. Sara, GONTIJO-LOPES, Raphael, SALIMANS, Tim, HO, Jonathan, FLEET, David J. & NOROUZI, Mohammad (2022), *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*, Neural Information Processing Systems.
- SCHUHMANN, Christoph, BEAUMONT, Romain, VENCU, Richard, GORDON, Cade, WIGHTMAN, Ross, CHERTI, Mehdi, COOMBES, Theo, KATA, Aarush, MULLIS, Clayton, WORTSMAN, Mitchell, SCHRAMOWSKI, Patrick,



KUNDURTHY, Srivatsa, CROWSON, Katherine, SCHMIDT, Ludwig, KACZMARCZYK, Robert & JITSEV, Jenia (2022), *LAION-5B: An open large-scale dataset for training next generation image-text models*, Neural Information Processing Systems, Datasets and Benchmarks Track.

SILVER, David, HUANG, Aja, MADDISON, Christopher J., GUEZ, Arthur, SIFRE, Laurent, VAN DEN DRIESSCHE, George, SCHRITTWIESER, Julian, ANTONOGLU, Ioannis, PANNEERSHELVAM, Veda, LANCTOT, Marc, DIELEMAN, Sander, GREWE, Dominik, NHAM, John, KALCHBRENNER, Nal, SUTSKEVER, Ilya, LILLICRAP, Timothy, LEACH, Madeleine, KAVUKCUOGLU, Koray, GRAEPEL, Thore & HASSABIS, Demis (2016), “Mastering the game of Go with deep neural networks and tree search”, *Nature* 529, pp. 484–489.

MEYER, Roland (2023), “The New Value of the Archive: AI Image Generation and the Visual Economy of ‘Style’”, in *IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft*, Jg. 19, Nr. 1, S. 100–111. DOI: <http://dx.doi.org/10.25969/mediarep/22314>.

SOMAINI, Antonio (2023), “Algorithmic Images: Artificial Intelligence and Visual Culture”, *Grey Room* 93, Fall 2023, pp. 74–115.

TRELEANI, Matteo (2017), *Qu'est-ce que le patrimoine numérique ? Une sémiologie de la circulation des archives*, Lormont: Le Bord de l'eau.

## NOTES

1. In this paper, we refer specifically to DALL-E 3 (<https://openai.com/index/dall-e-3/>) and Midjourney 5 (<https://www.midjourney.com>). During the paper's preparation, Midjourney, Inc. released its version 6 and added several new features. Among these new features, one that stands out is the Style Reference tool, which allows the user to indicate the URL of an image from which aesthetically similar images will be generated. Our paper aims first of all to propose a theoretical and methodological examination of the semiotic functioning of visual generative AI, and we consider that our theoretical and analytical perspective will not be affected by slight changes or updated versions.

2. In Meyer's (2023) interesting view, what counts in this new visual economy is not the final product of the translation between prompt and image but the fact that at every command we take through a prompt an exploration among images and words is launched: “In the new paradigm, however, the relationship between description and image seems to be less one of instruction and interpretation than one of *navigation and matching*: Verbal description does not determine what is to be produced, but *functions as a means of narrowing down selections in a space of possibilities not yet realized*. (pp. 103–104, emphasis added).

3. <https://stability.ai/stable-image>.

4. <https://commoncrawl.org>.

5. <https://www.mturk.com>.

6. In addition, databases are also built through security processes, the CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart): the way computer systems ask us to recognize certain objects (bicycles, motorcycles, cars, crosswalks) for “security reasons” on certain websites to make sure we are not bots.

7. <https://github.com/ryankiros/neural-storyteller>.

8. <https://chatgpt.com>.

9. Adrian MacKenzie and Anna Munster (2019) refer to this process as “platform seeing”.

10. For instance, the AlphaGo AI (Silver *et al.* 2016), capable of beating the champion of Go, a game that requires a higher level of strategic intelligence than the game of chess, was trained on a database of images with a resolution of 19 X 19 pixels. This is because it's the layout of the game pieces that counts in a topological space, not their visual quality or size.

11. We use the terminology of the computer vision community. The “first layers” are those that are not so “deep”, they are the closest to the input of the model and thus of the original image submitted to the model. The “last layers”, or the “deep layers”, are the closest to the output of the model, of the result produced, thus the furthest from the input image. The notion of “inferior” or “superior” layer might be found in some literature but is ambiguous in our opinion, and will not be used in this text.

12. The term perception remains strong, but formulas such as “perceptual topology” (Offert and Bell, 2021) and “optical computation” (Broeckmann, 2020), are already used in the humanities literature concerned with computer vision.

13. The generative models Midjourney and DALL·E 3 used in this article are not open source, and no technical paper is currently available. Therefore, their exact functioning is not fully known. The explanations that we provide are based on the most commonly suggested hypotheses and on the main principles underlying state-of-the-art published methods.

14. <https://www.wikiart.org>.

15. <https://www.artsy.net>.

16. <https://artsandculture.google.com>.

17. This way of eliminating an element is specific to Midjourney. Other generative models have different strategies, or no strategy at all. For example, for DALL·E 3, there is no specific command to avoid an element. Instead, it is advised to reformulate the prompt. For example, instead of saying “the sun is not visible”, it is advised to use “reverse psychology” and to instead mention what should indeed be represented, in positive terms, such as “it is dark” or “it is heavily cloudy”. For a sunny image, it might be better to use a two-step approach: first generate an image with a potential sun, then edit the image locally to remove the sun, without affecting the rest of the image.

18. We use the term of Louis Marin that in his theoretic system is opposed to transparency. Opacity is the enunciative layer of an image which works as a sort of filter of the represented visual theme and that makes a “hand” recognizable.

19. <https://platform.openai.com/docs/guides/images/prompting>.

20. <https://openai.com/index/gpt-4/>.

21. <https://openai.com/index/gpt-4v-system-card/>.

22. [https://en.wikipedia.org/wiki/Susanna\\_and\\_the\\_Elders\\_in\\_art#/media/File:Jacopo\\_Tintoretto\\_-\\_Susanna\\_and\\_the\\_Elders\\_-\\_WGA22656.jpg](https://en.wikipedia.org/wiki/Susanna_and_the_Elders_in_art#/media/File:Jacopo_Tintoretto_-_Susanna_and_the_Elders_-_WGA22656.jpg).

23. We found it necessary to assign such a role, because without it, GPT-4V tried (wrongly) to recognize the painting, as its output began with: “This image depicts a painting known as “Vanity,” created by the Baroque artist Antonio de Pereda in the 17th century. [...]”. The rest of the description was also sometimes a presumably memorized interpretation of Pereda’s painting.

24. [https://en.wikipedia.org/wiki/Susanna\\_and\\_the\\_Elders\\_\(Tintoretto\)](https://en.wikipedia.org/wiki/Susanna_and_the_Elders_(Tintoretto)).

25. Anchoring the scene to a specific century was not always done by DALL·E. The reasons for this remain unclear to us. It may contribute to increasing the overall consistency and “authenticity” of the image generated. However, we did not observe significant benefits with this kind of add-on.

---

## ABSTRACTS

In this paper, we propose a semiotic study on generative artificial intelligences, considering the work of Midjourney and DALL·E, that are computational devices capable of producing original images on the basis of the training they have received on large databases of visual, verbal and multimodal documents. The aim is to frame their functioning from a semiotic point of view and to describe the operations that can be performed during image composition, through the options available on the two platforms. To achieve this objective, the paper will be developed into three main parts. In the first part, we will provide a general contextualisation of the relationship between semiotics and artificial intelligence, in the broad sense. Starting from Pierluigi Basso Fossali's (2017) assumptions about a semiotic perspective understood as the study of the social organization of meaning, AIs reconfigure the thresholds between the four dimensions he defined (perception, enunciation, communication and transmission). By considering the transmission and perceptual dimension, we will define the phases of the database construction and AI model training as pertaining to an archival (distributed) perception. In the second part, we will deal with the dimension of enunciation. On the one hand, we will describe the functioning of the diffusion models guided by human prompts. On the other hand, we will show the limits and potential of these AIs, through a discussion of the operable commands and the results obtained, in light of the experiments we have carried out over the past months (August 2023-May 2024). In particular, we will test how generative AIs produce images on the basis of prompts containing the styles of specific artists, how they fuse different styles together, and how they work on visual stereotypes. In the third part of the paper, we will focus on the relationship between verbal description and visual generation in order to provide, in accordance with the perspective of the intersemiotic translation, a new research object and a new methodology.

Cet article propose une étude sémiotique des intelligences artificielles génératives, et notamment du travail de Midjourney et DALL-E, qui sont des modèles génératifs capables de produire des images originales sur la base de l'entraînement reçu à partir de grandes bases de données de documents visuels, verbaux et multimodaux. L'objectif est d'étudier leur fonctionnement d'un point de vue sémiotique et de décrire les opérations qui peuvent être effectuées dans les phases de composition de l'image. Pour atteindre cet objectif, l'article se compose de trois parties. Dans la première partie, nous fournirons une contextualisation générale de la relation entre la sémiotique et l'intelligence artificielle. À partir des hypothèses de Pierluigi Basso Fossali (2017) concernant une perspective sémiotique entendue comme l'étude de l'organisation sociale du sens, nous étudieront la manière dont les IA reconfigurent les seuils entre les quatre dimensions prises en compte dans son livre (perception, énonciation, communication et transmission). Du point de vue de la transmission et de la perception, nous définirons les phases de construction de la base de données et de l'entraînement des modèles génératifs comme relevant d'une perception archivistique (distribuée). Dans la deuxième partie, nous aborderons la dimension de l'énonciation. D'une part, nous décrirons le fonctionnement des modèles de diffusion guidés par des « prompts ». D'autre part, nous montrerons les limites et le potentiel de ces IA, à travers une discussion sur les commandes et les résultats obtenus, à la lumière des expériences que nous avons menées au cours de ces derniers mois (août 2023-mai 2024). En particulier, nous testerons comment les IA génératives produisent des images sur la base de prompts mentionnant les styles de divers artistes, la manière dont elles fusionnent différents styles et comment elles travaillent sur les stéréotypes visuels. Dans la troisième partie de l'article, nous nous concentrerons sur la

relation entre la description verbale et la génération visuelle afin de fournir, conformément à la perspective de la traduction intersémiotique, un nouvel objet de recherche et une nouvelle méthodologie.

## INDEX

**Mots-clés:** énonciation, archive, intelligence artificielle générative, image, traduction

**Keywords:** enunciation, archive, generative artificial intelligence, image, translation

## AUTHORS

### ENZO D'ARMENIO

Enzo D'Armenio is a F.R.S.-FNRS post-doctoral researcher at the University of Liège where he conducts the project “KineticEgo—Les performances identitaires dans les jeux vidéo et la réalité virtuelle. Une généalogie des médias visuels fondée sur le concept de mouvement”, dedicated to the interactive images of video games and virtual reality. He has been the principal investigator of the Marie Curie (Individual Fellowship) project “IMACTIS—Fostering Critical Identities Through Social Media Archival Images” ([www.imactis.eu](http://www.imactis.eu)), in which he analysed identity images on social networks. He has published articles for international journals such as *Visual Communication*, *Games and Culture* and *Semiotica*. He is also the author of the monograph *Mondi paralleli. Ripensare l'interattività nei videogiochi* (Unicopli, 2014).

Email: Enzo.DArmenio[at]uliege.be

### ADRIEN DELIÈGE

Adrien Delière is a postdoctoral researcher at the University of Liège, working on M.G. Dondero's P.D.R. F.N.R.S. project “Towards a Genealogy of Visual Forms”. His current research focuses on the analysis of the poses of characters and edges directions depicted in paintings, their evolution over time, the movements they induce, and their reuse in modern artistic photography. He initially studied mathematics (bachelor and master 2008-2013, then PhD 2013-2017), then he specialized in artificial intelligence at the Montefiore Institute (ULiège), particularly in deep learning for computer vision, during his postdoctoral position on Prof. M. Van Droogenbroeck's DeepSport project funded by the Walloon Region (2017-2022). He also won the “Ma thèse en 180 secondes” science vulgarization competition in 2015, and 5 “Best paper awards” at the CVPR conference's CVSports workshop for his work on automating the analysis of soccer videos.

Email: Adrien.Deliege[at]uliege.be

### MARIA GIULIA DONDERO

Maria Giulia Dondero, PhD, is a Research Director of the National Belgian Fund for Scientific Research (F.R.S.-FNRS), and Professor at the University of Liège. She is the President of the International Association for Visual Semiotics (IAVS/AISV) since 2023 and Delegate to International Affairs of the French Association for Semiotics (AFS). Dondero is Co-founder and Editor-in-Chief of the peer-reviewed Journal *Signata Annales des Sémiotiques / Annals of Semiotics* (SCOPUS) and Co-director of the collection ‘Sigilla’ at Presses universitaires de Liège. Her main research fields are the relation between semiotics and art history, the theory of photography and intermedial translation, the diagram in scientific discourse, and big data.

Dondero is the author of four books: *Les langages de l'image. De la peinture aux Big Visual Data*, Paris,



Éditions Hermann, 2020 (English augmented and revised version: *The Language of Images. The Forms and the Forces*, Springer, 2020); *Des images à problèmes. Le sens du visuel à l'épreuve de l'image scientifique*, with J. Fontanille, Limoges, Pulim, 2012 (Eng. trans. *The Semiotic Challenge of Scientific Images. A Test Case for Visual Meaning*, Ottawa, Legas, 2014); *Sémiotique de la photographie*, with P. Basso (Limoges, Pulim, 2011) (Italian and Portuguese versions available); *Le sacré dans l'image photographique* (Paris, Hermès, 2009). She has published around 80 peer-reviewed articles in French, Italian, English, some of which have been translated into Spanish, Portuguese, and Polish. She has directed 25 collective works and special issues on photography, scientific and artistic images and on the semiotic theory of visual language. She has been Visiting Professor at the University of Manouba, Tunisia (2012 and 2013); at UNESP-Araraquara University, Brazil (2014, 2016, 2019, 2023, 2024), at the National Institute of Anthropology and History (INAH), Mexico (2017), at Paris 2 Panthéon-Assas, France (2019-2020), at Celsa Sorbonne University (2020-2021), at the University of Turin (2021-2022), and at La Sapienza, Rome, Italy (2023). She has been Invited Visiting Researcher at the University of Southern California (2020) and at Purdue University (2022). Academia.edu: <https://frs-fnrs.academia.edu/MariaGiuliaDondero>. Email: MariaGiulia.Dondero[at]uliege.be