Semiotics of artificial intelligence: a computational analysis of big datasets and automatic image generation[a, b]

Semiótica da inteligência artificial: análise computacional de grandes bases de dados e geração automática de imagens

MARIA GIULIA DONDERO[c]

Fonds National de la Recherche Scientifique. Liège, Bélgica

GUSTAVO H. R. DE CASTRO[d]

São Paulo State University. Araraquara – SP, Brazil

MATHEUS NOGUEIRA SCHWARTZMANN[e]

São Paulo State University. Araraquara – SP, Brazil

DANIERVELIN PEREIRA[f]

Federal University of Minas Gerais. Belo Horizonte – MG, Brazil

ABSTRACT

Artificial intelligence today simulates the complexity of language and human actions in an increasingly satisfactory waymanner. In this paper I discuss artificial intelligences using semiotic tools, assuming as a theoretical standpoint É. Benveniste's theory of enunciation in post-Greimasian semiotics, notably Jacques Fontanille's concept of enunciative praxis, applied to the study of artificial intelligence. This theoretical basis will allow us to address the relation between image databases and algorithms in analyzing of large image collections through computer vision, as well as user's communication modes with the Midjourney generative artificial intelligence model, focusing on machine creativity.

Keywords: Artificial intelligence, generative model, enunciative práxis, image generation, Midjourney

[a] This paper is partly based on the research presented in the article "Inteligência artificial e enunciação: análise de grandes coleções de imagens e geração automática via Midjourney" (Dondero et al, 2024). This research is developed, updated, and improved here.

[b] We thank A. Deliège for his explanations of methods of analysis in machine learning. We also thank Enzo D'Armenio for his comments on image generation using Midjourney.

[c] Director of research at the Fonds National de la Recherche Scientifique (F.R.S. – FNRS), in Belgium, and professor at the University of Liège. Orcid: [0000-0003-2320-8130]. Email: mariagiulia.dondero@uliege.be

[d] PhD student (Fapesp 2019/27000-7) of the Postgraduate Program in Linguistics and Portuguese Language at the Universidade Estadual Paulista (Unesp). Translation, review and notes. Orcid: 0000-0003-4486-9579. Email: g.castro@unesp.br

[e] Associate Professor in Discourse Semiotics at Universidade Estadual Paulista (Unesp), currently Coordinator of the Postgraduate Program in Linguistics and Portuguese Language at Unesp in Araraquara (SP) and member of the Permanent Committee of the Coordination of Affirmative Actions, Diversity and Equity (CAADI) from Unesp. Rereading. Orcid: 0000-0002-2887-3570. Email: matheus.schwartzmann@unesp.br

[f] Professor at the Faculty of Languages, Literature, and Linguistics at the Federal University of Minas Gerais. In 2024, he will carry out a post-doctorate at the Federal Fluminense University and the Université de Liège with Capes PrInt funding. Final rereading. Orcid: 0000-0003-1861-3609. Email: daniervelin@gmail.com

DOI: http://dx.doi.org/10.11606/issn.1982-8160.v18i3p29-54

RESUMO

A inteligência artificial simula hoje, de maneira cada vez mais satisfatória, a complexidade da linguagem e das ações humanas. Neste artigo abordamos as inteligências artificiais com instrumentos semióticos. Aqui, assumiremos o ponto de vista da teoria da enunciação de É. Benveniste, especialmente dos desenvolvimentos em semiótica pós-greimasiana, e sobretudo de Jacques Fontanille aplicados ao estudo da inteligência artificial. Essa base teórica nos possibilitará discutir, primeiramente, a relação entre banco de dados de imagens e algoritmos na análise de grandes coleções de imagens por meio da *computer vision*, além, dos modos de diálogo do usuário com o modelo de inteligência artificial generativa Midjourney, que nos permitirá tratar a criatividade da máquina.

Palavras-chave: Inteligência artificial, modelo generativo, práxis enunciativa, geração de imagens, Midjourney

THE HISTORY OF artificial intelligence (AI) goes back to the 1950s and the Turing[g] machine, which remains the fundamental theoretical model for all computing today: this was the beginning of digitization and the automation of calculations. Let us leap back in time to more recent history. We can see that the automation of computing underpins the operation of many everyday utilities: search engines, product recommendation and navigation systems, strategy games, chatbots, and, more recently, automatic image[h] generation models.

In general, artificial intelligences offer tools that try to simulate in an increasingly convincing and powerful way, the particularity of human language and its practices, including thinking processes[1]. For this reason, it is essential that post-structuralist semiotics deal with artificial languages as well as the technologies and practices that automate human actions.

This paper is divided into two sections. In the first, we look at the approach to database *analysis* to examine how Computer Vision, in conjunction with other disciplines such as art history, enables the *analysis of large amounts of data* (Big Visual Data) using appropriate algorithms that transform statistical analysis into image visualizations (meta-images).

Next, we will study the *automatic generation of visual statements*, i.e. the large collections stored in databases, used to produce new statements from old texts, already sedimented in the collective memory, through operations or even instructions (prompts)[i]. Particularly in the case of generating new utterances, we will study some interactions and some of their textual products obtained

[g] (N.T.) The Turing Machine consists of the conceptual metaphor of an infinite tape, which acts as long-term memory, on which symbols can be read and written; and of a read/write head that moves along the tape, according to a table of instructions responsible for determining operations.

[h] (N.T.) The first attempts to generate images automatically date back to the 1960s-1970s, with Harold Cohen's AARON program. Subsequently, a series of technologies were developed. Just as an example, we can mention a few milestones: in 2018, Progressive GANs appeared, followed by Google's BigGAN, which made it possible to generate images by gradually improving them in terms of resolution. In 2021, OpenAI introduced DALL·E, inaugurating the generation of images from textual descriptions.

[1] In the 1950s, Turing asked a crucial question in his famous article "Computing Machinery and Intelligence" (1950): "Can the machine think?". For a philosophical discussion of the origin, history, and developments of the Turing machine, see Jean Lassègue's book: Turing (2017).

[i] (N.T.) In the context of computing, a prompt consists of instructions or stimuli given, for example, to AI systems to generate responses or perform specific tasks, directing the model in the production of content.

through Midjourney. The models used by Midjourney (or even by DALL·E, to mention another example) translate verbal statements (prompts) into visual ones, or vice versa: they produce verbal statements with the aim of, for example, describing an image that the user proposes to Midjourney.

We would dare to say that the generation of visual texts using this model interests us more than the experiments with ChatGPT. This is because, especially in the case of Midjourney, translation does not just take place between machine language and human language[2]. It mainly takes place between the verbal language of the prompt (the command given) and the visual language (the product generated). The instructions are applied to verbal and visual databases, which play a fundamental role in these operations of analysis, translation, and production of utterances.

A database can, in semiotic terms, be seen according to the notion of an encyclopedia proposed by Umberto Eco (1984), or even in Greimasian and post-Greimasian terms, as *the place where verbal and visual discursive forms are sedimented*, thinking now of the mechanism for the renewal of human culture formalized by Jacques Fontanille (1999) in the theory of enunciative praxis. We will use this theory in our paper, considering that the database would be the place of virtualization, i.e. cultural objects, and sedimented speeches in collective memory, and archives, from which new *creations/performances can be produced (updating/realization)*, in this case, "automatically", since we are dealing with artificial languages. The theory of enunciative praxis will therefore help us to study the dynamics *between innovation and sedimentation* in the context of databases understood as archives and as places where the new is generated.

ANALYSIS OF SIMILARITIES/DISSIMILARITIES BETWEEN IMAGES IN DATABASES

Briefly, we could define AI as a tool dedicated to performing tasks in place of a human who has previously trained it. Teaching a machine essentially consists of enabling it to learn to perform a task from an appropriate database. To do this, the programmer must first choose the type of learning algorithm (random forest, svm, etc.), which means choosing a strategy according to the task to be performed and the nature of the data provided (images, spreadsheets, etc.).

In the case of analyzing large collections of images, we will consider two strategies. The first, *feature extraction*, is the strategy used by Lev Manovich (2020b). It consists of a method that draws resources from the content of

[2] For a comparison of all generative models, see Santaella and Kaufman (2024).

databases based on rules defined previously and "manually" by the researcher, who sets the computational instructions to be followed to carry out the task. This is an example of the choice of features to be extracted, such as the gradients of luminosity in paintings by abstractionist artists from the early 20th century[3].

The second strategy is *deep learning*[j], which consists of an algorithm responsible for providing the machine with a set of data through which and in which it must detect similarities/dissimilarities.

When we use a deep learning algorithm, we are no longer in *the eye of the researcher* who decides what the machine should find in the collection of images (as was the case with feature extraction). It is another situation, which puts us in the extension of the database used to train the algorithm. Now, by adopting deep learning as a strategy, we are letting the algorithm itself decide what it has to calculate to perform its task satisfactorily. In this case, all the researcher has to do is give the model an opinion on its results, allowing it to correct itself without, however, telling it exactly what calculations it should have made. In fact, *it is the quality of the database that will determine the model's ability to learn to perform its task more or less correctly*.

Obviously, if the algorithm has been trained on a data set made up of common images, representing everyday objects, for example, it will be very difficult to obtain good results in the context of searching for artistic images[4]. In other words, the data set on which the algorithm is trained must have sufficient affinities with the database that will be presented later: only then can it analyze it relatively satisfactorily, in terms of similarities and/or differences.

The tasks performed by the machine "in our place" - and which we have been studying for some years (Dondero, 2020) - are mainly related to image analysis. It is obvious that, especially when what is at stake is the organization of large collections of visual data (thousands of images), digitized according to their similarities/dissimilarities, the machine is being asked to perform a task that goes beyond the purely human capacity for analysis[5]·

Therefore, we can see that the production of this massive data has made it possible to analyze large collections of images, even reopening the ground for research projects that were not even conjectured before. We refer here in particular to the project by art historian Aby Warburg. In his work *Atlas Mnemosyne* (1924-1929), Warburg (2012) studied the image through the image, using visualization as his research method, so that images with common characteristics[6] were arranged next to each other on large black panels, based on similarities in composition.

[3] See Lev Manovich's website in this regard, and especially the section on the spaces of styles (Manovich, 2011).

[j] (N.T.) Deep learning (or deep learning) consists of an algorithm that defines a model, often a neural network, based on an initial set of parameters, gradually optimizing (deepening) the variables to perform the desired task.

[4] In general, you start by compiling the complete set of data to be analyzed. We then extract part of it, make notes, and use it for training. The rest will be used to verify the results, which are, in fact, what we want to study. If this distribution of the initial data set is successful, we have a better chance of having a model that more accurately generalizes the data of interest.

[5] In this case, it is important to note that we are talking about *digitized* data, not *digital* data: we intend to focus on the analysis of large collections of images belonging to Western artistic heritage, and not on the data produced by digital technology itself, such as those we generate daily on social networks, emails, etc.

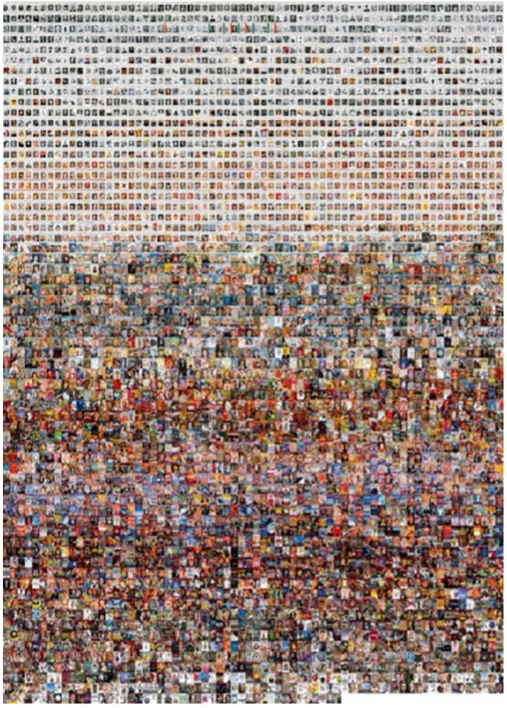
[6] On the problem of style in Warburg, see Pinotti (2001).

Faced with the multiple questions that each artistic image poses for the observer and the historian, Warburg chose a formula as an answer and a general explanation, so to speak: bring an image closer to "its nearest neighbor", in aesthetic terms, and gradually distance it from those it opposes or conflicts with. This is exactly what the researchers who, like Lev Manovich, followed this proposal did, ensuring that databases and their algorithms could group data according to their similarities and differences, according to Warburg's "nearest neighbor" rule.

A few years ago, we studied two visualization models (Dondero, 2017b, 2019b, 2020) that exemplify the possibilities of this proposal. The first is a classic montage of around 4,500 images (Figure 1).

Figure 1

Montage with 4,535 Time Magazine covers (1923-2009)



Note. Manovich and Douglass (2009).

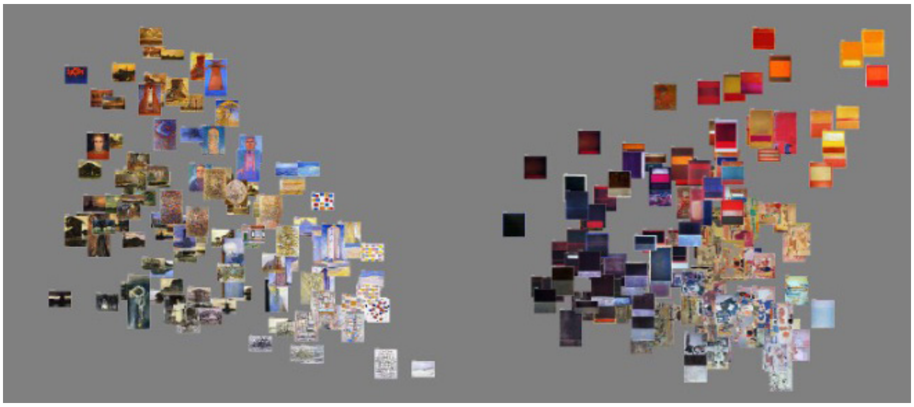
The second, more interesting, are the visualizations we call image diagrams. The montage seems to us to be less interesting for a perhaps obvious reason: it follows an organization determined by a metadata, in this case, the date of production[7].

[7] This is a strategy that we have already criticized in several publications (Dondero, 2017, 2019b, 2020), in which we explained that organizing collections through metadata falls into the same error that Roland Barthes is criticized for with translinguistics, that is, the attempt to reduce the image to what can be lexicalized.

In the case of the diagram, the arrangement of the images depends solely on the instructions that the researcher gives to the machine - and not on the metadata, as is the case with the montage. These instructions aim to measure the visual similarity between the aesthetic characteristics of the images contained in the database[8] (Figure 2).

Figure 2

Comparison of 128 paintings by Piet Mondrian (1905-1917) and 151 paintings by Mark Rothko (1944-1957). The two image views are placed side by side so that they share the same X-axis. X-axis: average brightness. Y-axis: average saturation



Note. Manovich et al. (2011).

Among the aesthetic qualities, the chromatic category is easy for the machine to work with because it is a quantitative character category, just like light intensities. In fact, in the case of the "feature extraction" method discussed here, the aim is to extract from images the aesthetic characteristics which, in Computer Vision, are called "low-level features"[k].These are qualities that are not directly linked to figuration. However, this task involves measuring the averages of each characteristic distributed on the surface of each image and is therefore not to be confused with the identification of *visual or figurative formants* (Greimas, 1984).

In some of our previously published works (Dondero, 2017a, 2019b, 2020), we criticized this methodology for analyzing large collections of images: the extraction procedure works with *average*[9] aesthetic characteristics, failing to focus on the *distribution* of these characteristics within the artistic image, understood as a totality[10].

However, despite the criticisms that can be made of this or that statistical method, numerous reasons justify the semiotic interest in these analyses, among which we list two: these visualizations develop one of Warburg's questions (that

[8] For an enunciative analysis of these two types of image visualization, see the third chapter of The Language of Images (Dondero, 2020), in which we distinguish between the relevant focal points for montage and those relevant to diagrams, adopting Fontanille's classification proposed in Sémiotique et littérature (1998).

[k] (N.T.) The term refers to basic and primitive characteristics of images, extracted without the need for semantic modulation, which can include colors, textures, edges, shapes, histograms, gradients, among others.

[9] There is, however, one observation: when we use a network like ResNet to extract/calculate an embedding from an image, which we call feature extraction (at least in the case of Computer Vision), this embedding still contains information related to the distribution of features and not just the average.

[10] On the image as a whole, see Goodman (1976), Thom (1983) and Dondero (2020).

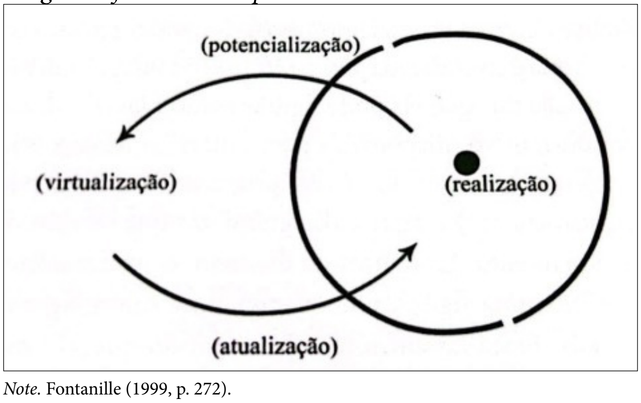
of images and their "nearest neighbors"), through a controllable methodology, and also these visualizations reveal work that can be understood as structuralist in two senses:

(i) these visualizations contrast groups of images with gradually similar or opposite aesthetic characteristics and organize the individual aspects of each image *gradually*, within a controlled space (a tensive perspective of the structure)[11]; these image visualizations present the analysis carried out (in the sense of division, grouping, reconstruction of the relationship) and allow diagrammatic reasoning to be carried out, bringing statistical and perceptual aspects into play via *semi-symbolism*, according to the similarity/dissimilarity parameter. We can use these visualizations to carry out statistical-perceptual experiments on a collection based on various parameters relevant to each database (which are not limited to chromatic or luminous characteristics but also include aspects of the geometry of the shapes, the length, and typology of the lines drawn);

(ii) the collection of images can be studied as a system in which the machine has been trained to work with differences and similarities. As we have already indicated, the collection system functions as an encyclopedia, in other words, as a system of "co-texts", to use a term from U. Eco (1984). Alternatively, even as the place where the discursive strategies of artistic forms (for example, pictorial forms) of all times have been sedimented. We can therefore situate them in the diagram of enunciative praxis in the *space of virtualization*:

Figure 3

Diagram of enunciative praxis



Note. Fontanille (1999, p. 272).

[11] See Fontanille and Zilberberg (1998).

Many of the questions asked in visual semiology and semiotics since the 1960s have still not been resolved - for example, “Is there a visual language?” However, at least now these questions have been concretely posed and, in part, answered thanks to the algorithmic operations carried out on image databases. A database is not exactly comparable to the Saussurian *langue*, which is made up of virtualities, but it has a conceptual surface very similar to that of the *locus* of virtualization: the images it contains are not mere pictorial virtualities, but images that have been produced historically and, in a way, are co-present in the database. After digitization, they share a common digital substance that makes them *commensurable* and available for algorithmic analysis. This commensurability means that, in a database, images can be manipulated and measured until their specificity/difference stands out from the rest.

A research project on the genealogy of gestures in painting[12]

My research *Em direção a uma genealogia das formas visuais—Semiótica e abordagens computacionais para grandes coleções de imagens (Towards a Genealogy of Visual Forms: Semiotic and Computer-Assisted Approaches to Large Image Collections)* (2022-2025, F.R.S-FNRS) is another way—more complex, we hope—of continuing the theorization of the genealogy of forms that was done by art historian Henri Focillon (1934) in the book *Vie de formes (Life of forms)* and in particular the forms of *pathos* (Warburg’s *pathosformeln*[1]) that we can see in *the poses and gestures of the figures* portrayed in paintings throughout history.

However, that is not all. It’s also about advancing the project of visual semiotics itself, in particular, a specific issue that we have already addressed above in our book *The Language of Images* (Dondero, 2020): the study of movement, temporality, and narrativity in the static image, based on how temporal and aspectual enunciation is signified in a fixed substance such as painting. The enunciation of the category of person has been extensively developed and studied in works on the face and profile (Beyaert-Geslin, 2017; Dondero, 2023, 2024b). The same has happened with spatial enunciation - in the latter case, in particular, thanks to various works on perspective, such as those by L. Marin (1993) and J. Fontanille (1989). However, temporal enunciation (the before and after within an action represented in an image), aspectual enunciation (the moment of the action, focused on by the producer), and the rhythm of the unfolding of the action have not been sufficiently investigated. There are very few studies on this issue: we could mention the

[12] This research project, which began in 2022, will end at the end of 2025. It is entitled “Towards a Genealogy of Visual Forms: Semiotic and Computer-Assisted Approaches to Large Image Collections” and is funded by the Fonds de la Recherche Scientifique (F.R.S.-FNRS-Belgium). More information at Dondero (2022).

[1] A concept whose literal translation would be “forms of pathos”. It refers to the feeling conveyed (culturally) by gestures, poses, and facial expressions. In the context of Computer Vision, which the author of this paper is involved in, the processing of poses is crucial for improving human-machine interaction, for example.

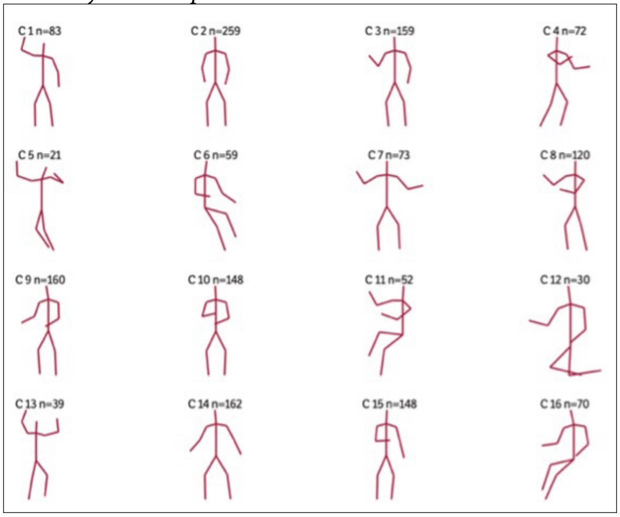
work of J. Petitot (2004) on *Laocoon*, an article by the µ Group (1998), and another by M. Colas-Blaise (2019), for example[13].

Taking into account this state of the art in semiotics, our project aims to analyze the representation of body gestures in a *corpus* of millions of images. Why poses and body gestures? Because they are the site of a *figurative dynamic*, i.e. a *continuum*, and because our challenge is to study *movement in static images automatically*.

On this topic, before us, Impett and Moretti (2017) formalized the gestures of the bodies found in the panels of Warburg's *Atlas Mnemosyne* (Figure 4).

Figure 4

Mnemosyne Atlas panels.



Note. Impett & e Moretti (2017).

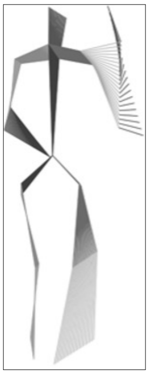
My critical observation regarding this method is that Impett and Moretti's (2017) modeling reduces the body to a skeleton, a geometric figure made of line segments, while the body has a volume made of internal forces that occupy space and a silhouette that is involved in every gestural dynamic and plays a crucial role in directionality within a landscape.

Leonard Impett, in a 2020 article published in the *Routledge Companion to Digital Humanities and Art History,* entitled "Analyzing Gesture in Digital Art History", tries to complexify the model of the body by inserting the parameters of directionality and the rhythm of movement (Figure 5). From this diagram, we can see that not only has the skeleton been complexified, but also that Impett tries to calculate the rhythm of movement and the displacement of the body.

[13] For a more recent discussion of this issue, taking into account the authors cited and the book by Basso Fossali (2017), see Dondero (2024b).

Figure 5

Principal component analysis on the poses of Atlas Panel 46, capturing the strongest morphological feature of the panel: the nymph



Note. Impett (2020).

Similarly, our current research project aims to trace poses, gestures, and other types of movement and force dynamics in static images, such as paintings and photographs spanning a period from Baroque painting to contemporary fashion photography, extracting poses and grouping them.

What do I mean by "the dynamics of forces in a still image"? The forces can be partially identified with directionality: the direction of a gaze, a raised hand, a pointed finger, but also with the directions given by components of the image that are not figurative, but formal/visual: the change in luminosity in a painting acts as a kind of arrow, the change in saturation is capable of producing a lifting or falling force. The geometry of a gesture also counts: a gesture that composes a circular figure on the plane of expression reflects stability and calm on the plane of content; on the contrary, a gesture that composes an irregular triangle will reflect disturbing directions, a conflict on the plane of content.

In this context, Adrien Deliège and I took some examples from a corpus produced from the complete collection of paintings available on *Wikiart* - sorted, of course, according to our needs. We arrived at a group of 5,000 religious images, containing 8,599 individual poses. Each individual pose is redrawn into a separate image, with key point coordinates normalized to allow meaningful comparisons between the images (Figure 6).

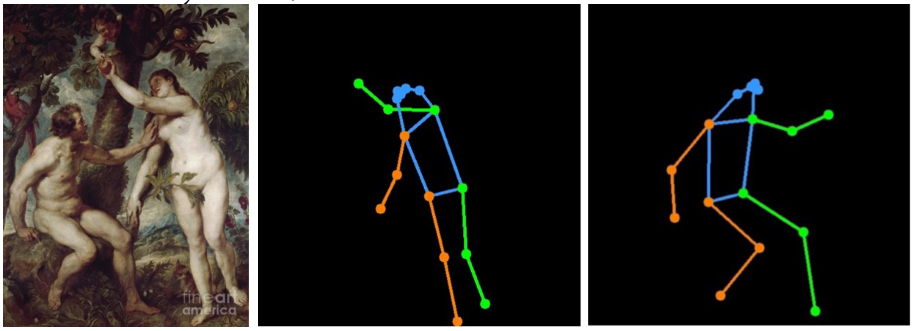
We used MMPose[m] for mapping and Pixplot[n] for visualizing the poses. In MMPose, the poses are mapped using 17 key points, which attempt to cover the entire *body*. When the 17 points are not covered, we exclude the image from the corpus because the algorithm has not identified an entire body.

[m] (N.T.) MMPose is a framework (a set of tools, libraries, and conventions that provide a basis for the development of open-source software) designed to map human poses in images and videos by tracking joint angulation and direction, as shown in figure 2.

[n] (N.T.)PixPlot is an interactive visualization tool designed to explore large collections of images in a two-dimensional way. It allows you to identify visual patterns, groupings, and relationships between images intuitively or automatically.

Figure 6

Example of individual poses taken from an original painting, Peter Paul Rubens, Adam and Eve in Earthly Paradise, circa 1628

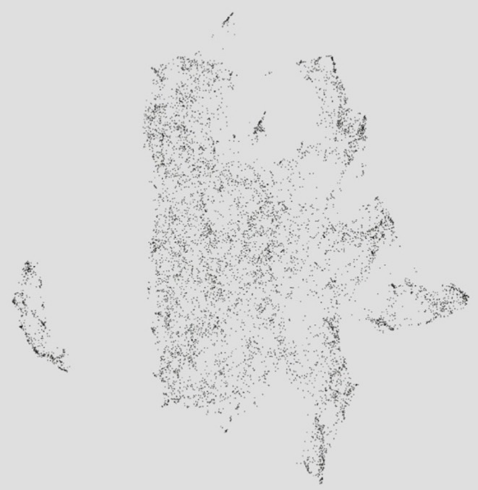


Note. Deliège and Dondero (in press).

We first analyze all these individual poses before moving on to the collective ones. We defined the distance between two poses as the sum of the distances between their corresponding key points. We then used this distance metric in the PixPlot software's UMAP dimensionality reduction module to produce a visualization of the way the individual poses are arranged (Figure 7), i.e. the similar poses are placed close to each other and far away from the different poses. The following example shows this organization.

Figure 7

Visualization of 8,599 individual poses from 5,269 religious paintings contained in the Wikiart database and organized by pose similarity[14]



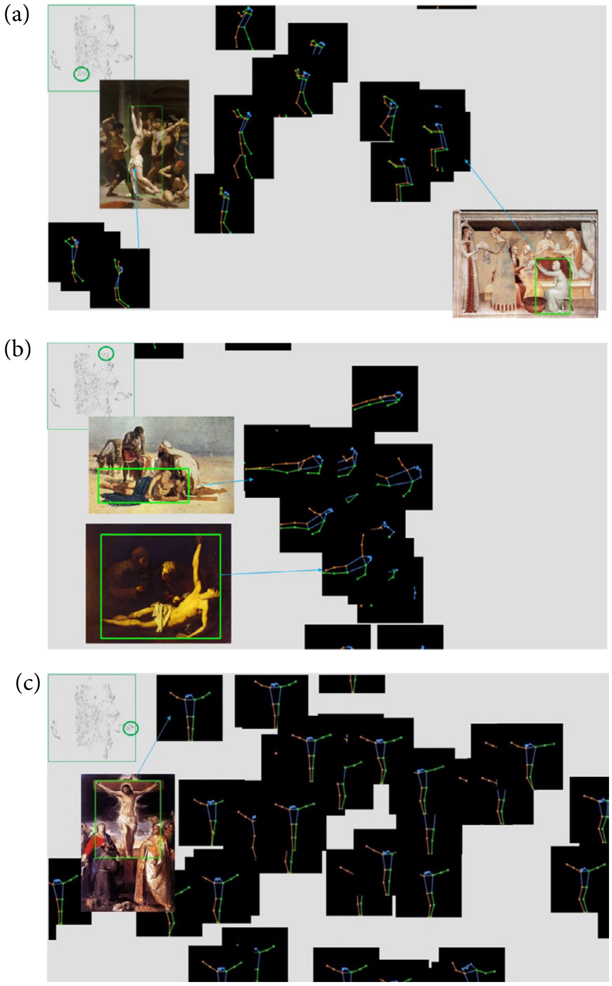
Note. Deliège and Dondero (in press).

[14] An interactive web application that allows you to browse this "image cloud" (with zoom in and out functionalities) is available at: https://adriendeliege.z6.web.core.windows.net/outputs/WikiArt_religious_painting_solo_poses/index.html. Adrien Deliège and Maria Giulia Dondero (2023).

Within this collection, it is possible to follow the variety of poses according to the organization established by the algorithms. For example, each green circle in the general diagram of the collection (top left of each subfigure in Figure 8a,b,c) refers to a group of paintings that the machine recognizes as belonging to a specific pose, which is enlarged to show the validity of the approach.

Figure 8

Examples of specific poses shared by certain paintings are outlined in green (top left) in the overview and zoomed in on these delimited regions to show the similarity and variations of similar poses belonging to a group of images—visualization generated via MMPose and Pixplot.



Note. Deliège and Dondero, (in press).

We can go through the entire collection—that we call the *reference corpus*, as it is comprehensive—and select the *working corpus*, which, also according to our terminology, comprises several groups of poses (outlined in green).

You can see that the center of the visualization tends to group relatively neutral poses, representing a person standing, and facing the viewer. As we move away from the center, the poses vary continuously, reaching completely different poses in the corners of the meta-image, such as characters lying down, sitting, falling, etc. A group of representations of Jesus on his cross is also clearly visible, as this pose is common in religious paintings. We can also identify a group of images in which the character is seen from the back, which is completely, and rightfully, dissociated from the rest of the images (on the far left of the view). We should also note that a body lying down with the head to the left is a completely different pose (according to the metric used) than if the head is to the right of the image. Opposite poses are represented in opposite parts of the visualization, i.e. at the top and bottom in this case. Finally, as with all large-scale automated analysis, some unfiltered errors have gone unnoticed by this visualization; in this case, such as a small group of poses with legs cut off at the knees, corresponding to characters who are not completely shown in the images (in the top central part surrounded by an empty neighborhood in the visualization).

These poses can, for example, be used in research into the relationship between the expression of these images and their content, following a *semi-symbolic analysis* that establishes that an opposition on the plane of expression is correlated to an opposition on the plane of the images' content. Two simple examples could be the following: arms up *vs.* arms down = prayer *vs.* rest, or standing body *vs.* reclining body = arrogant pose *vs.* pious pose.

There are still several issues to be considered in collective poses: we need to decide whether to create a genealogy of the most similar collective poses or the most similar *forms of poses* (some collective poses can form triangles, squares, etc.), or whether to take into account the location of the poses and their scale on the surface of the image. Some of the developments in this work can be found in other texts by Adrien Deliège[15].

CREATING NEW IMAGES

After the computational analysis of big data, we come to the second part of this paper, on the *automatic and algorithmic creation of new images*. In this case, it is the machine that enunciates our instructions, translating them from verbal language into visual language. Nevertheless, it can also do the opposite, as we have already said: describe an image in verbal language. In this respect,

[15] See Deliège (2024).

the enunciative possibilities of generators such as Midjourney are quite broad for each translation direction (verbal ↔ visual): it is possible to change the style of a photo, mix several of them, or merge images by different artists. However, each new image begins with another, more fundamental type of translation: the translation of the image and verbal text into numbers. The ability to manipulate the image, acquired in this way, allows generative AI to produce new images automatically, using databases and machine learning methods. These new images are generated by operations carried out on all the images already produced, which are stored and annotated according to style, author, and genre in the available databases (Wikiart, Artsy, Google Art and Culture, etc.).

Image generation models use a Large Language Model[°] component, or at least a model that understands natural language (e.g. CLIP – Contrastive Language-Image Pre-training, to transform prompts (commands) into embeddings (lists of numbers) that can be used by the machine[16].The lists of numbers that describe images are linked to lists of numbers that identify texts in natural language. These learning models, responsible for translating between verbal and visual languages, are determined by the organization of the database content.

Producing a series of images requires the user to perform various operations[17]. When an instruction is given to Midjourney, four versions of this verbal instruction are obtained by default, which differ from each other in terms of light intensity, positioning of objects, etc. The experimenter must choose the best one and decide—or not—to continue searching for the ideal image, giving additional instructions to modify the chosen version. It is possible to transform the four versions produced (which can be understood as different *optimizations of the instruction given*), choosing one in each series of four until the result corresponds to the image desired by the experimenter.

It is also necessary to remember that each image produced, or each set of images produced, is, from a scientific point of view (which interests me), more interesting as *samples of the dataset's areas than as isolated images tout court*. In other words: the images produced by generative models count more as extractions of typical characteristics of the *dataset region*, that is, as examples of patterns produced by the work of algorithms that explore (Meyer, 2023) certain domains of the dataset decided by the notes and operationalized by the embeddings, than as stabilized images and definitive correspondences between certain words and certain shapes.

We will now examine the process that leads us to generate images.

Allow us to look, for example, at the stereotypes that the machine presented to us from the extensive databases on the Renaissance, Baroque, Mannerist,

[°] (N.T.) Artificial intelligence model developed by OpenAI, designed to understand both text and images in an integrated way. This allows the model to associate words and phrases with visual content, without the need for supervised training.

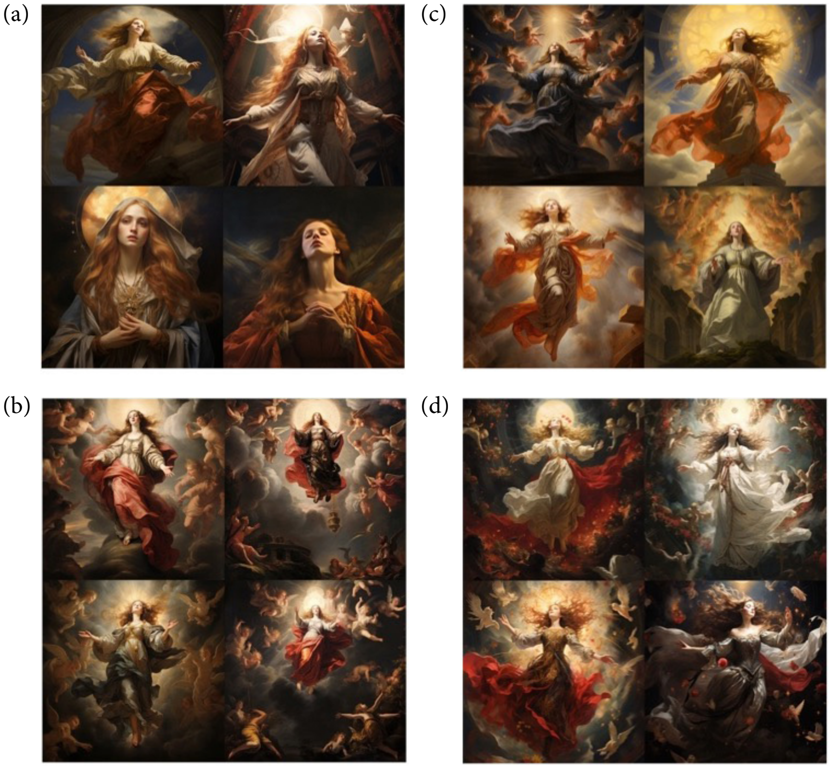
[16] This is the case with models for generating verbal texts, such as GPT-3.5, GPT-4, Llama, Claude, PaLM.

[17] Enzo D'Armenio and Adrien Deliège were important players in the development of these reflections, having actively participated in the experiments.

and Rococo styles: the machine produces an *average* of all the paintings in the Renaissance, Baroque, Mannerist, and Rococo styles as shown in Figure 9.

Figure 9

Prompt: /Ascension of Mary Magdalene in Renaissance, Baroque, Mannerist and Rococo styles/



Note. Experiment carried out by M. G. Dondero, Enzo D'Armenio, & Adrien Deliège &e M. G. Dondero (2024), via Midjourney.

There are two things that this process of transforming styles into averages of several singular images does not prevent: the first is that several averages can produce new forms, as demonstrated by many competitions won by AI[18]-generated images (which also allows us to appreciate the part of randomness that follow all generations of images); the second is that, although the machine can extract and imitate various styles, the "hand" of the machine always remains visible. Thus, we can study the stylistic singularity—what in semiotics we call *enunciative opacity* (Marin, 1993)—of the machine's hand since we know the reference styles based on which the machine works[19].

[18] Some images produced by machines have even won prizes in competitions for images produced by humans. This is the case at the "Colorado State Fair", where Jason Allen won the art competition thanks to his work entitled Théâtre d'Opéra Spatial, produced with Midjourney (Geoffre-Rouland, 2022). Another example of a work of art produced by artificial intelligence is Unsupervised, 2022, by Refik Anadol Studio. This project uses neural networks trained on a database of 10 thousand works of art from the collection of MOMA - Museum of Modern Art. This collection includes art from 1870 to 1970, as well as works from later decades. On this subject, we recommend the work by Manovich & Arielli (2021-2024) and the paradox it highlights between the modernism movement - to which the works in the database belong, which aims for the new and the destruction of the old - and the algorithms that rework them.

[19] On the difference between Midjourney and DALL·E based on their respective results, see D'Armenio, Dondero, Deliège and Sarti (in press). This paper compares various parameters: the way the two models respond to requests about aesthetic categories (shape, color, topology), the generality/specificity relationship, temporality in the static image, and so on. For a comparison of the Midjourney and DALL·E databases using tests on pictorial iconographies, such as "Suzanna and the Elders", see D'Armenio, Deliège, and Dondero (2024).

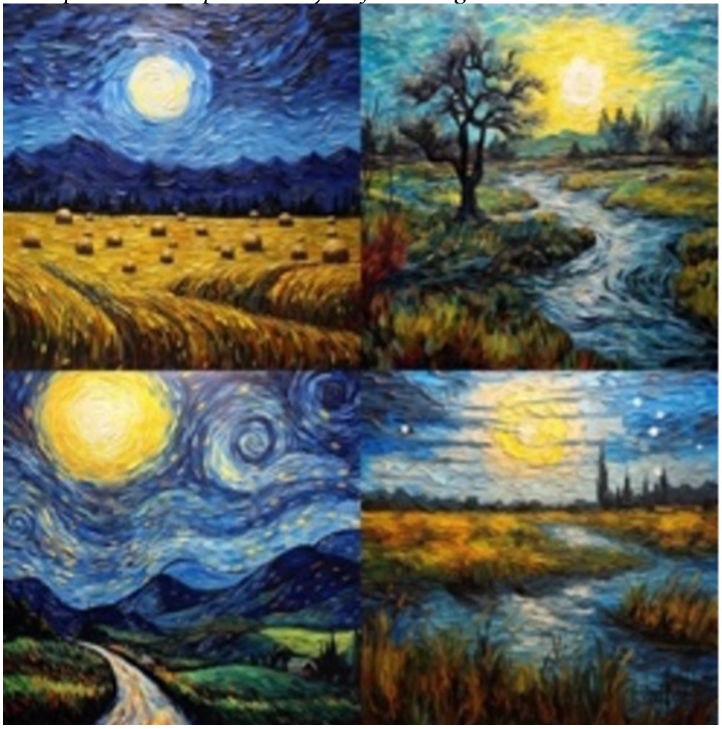
The experimenter, if he or she is a programmer, can decide to refine (adjust) a neural network using notes, building more precise correspondences between the lists of numbers that identify the descriptions in natural language and the lists of numbers that identify the images.

From our side, to make the production of images closer to our objectives and thus minimize the bias or noise generated by excessively generic databases, we can at most refine our prompt by providing it with more indications. Another way of creating restrictions that limit the generality of the results is to explicitly indicate the technique to be used, such as /chalk drawing/, /oil painting/, / fresco[20] / etc., as well as, of course, specifying one or more pictorial styles.

We also asked Midjourney to generate typical images of Van Gogh, using the prompt: /a landscape in the style of Van Gogh/ (Figure 10). We quickly realized that it would be difficult to get rid of certain objects, especially the sun.

Figure 10

Prompt: /a landscape in the style of Van Gogh/

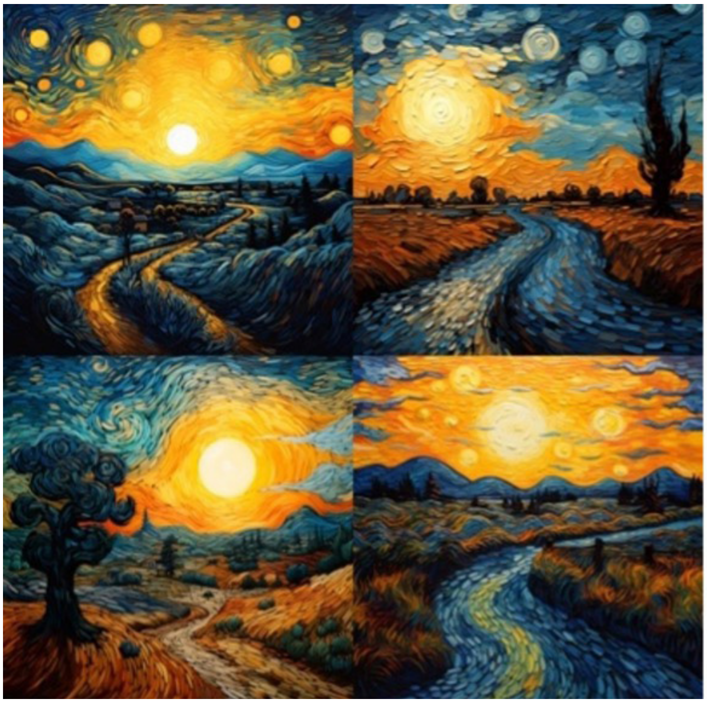


Note. Experiment by Enzo D'Armenio, Adrien Deliège & M. G. Dondero (2024), via Midjourney.

[20] This dimension is apparently very important for the machine, but it was relatively ignored in semiotics until the theoretical-methodological developments relating to media.

This is probably because this figure is considered a predominant feature in Van Gogh's work (depending on the correspondence between the image embeddings and the descriptions of the images that have been coded). A first, perhaps "naive" attempt to make the sun disappear was to add /without sun/ to the prompt:

Figure 11
Prompt: /a landscape in the style of Van Gogh without sun, without moon/

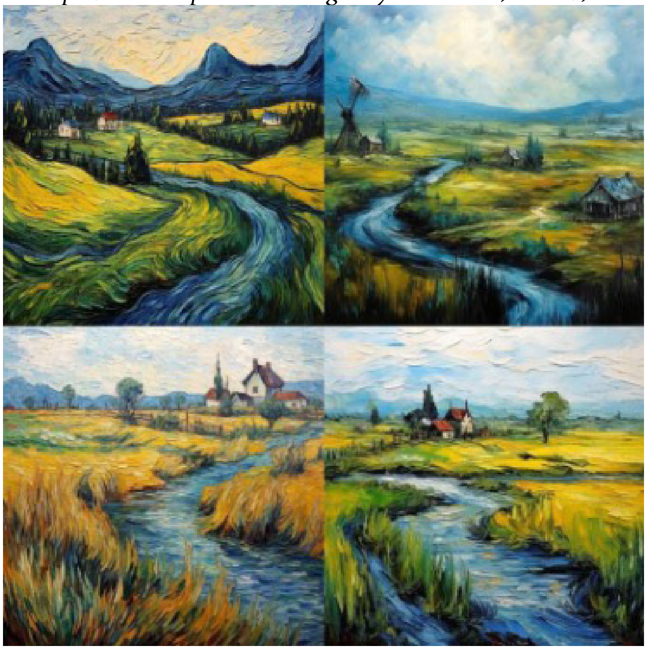


Note. Experiment by Enzo D'Armenio, Adrien Deliège & M. G. Dondero (2024), via Midjourney.

We can see that the images keep the sun (or a moon? It is hard to say) because Midjourney was not designed to distinguish between the positive and negative meanings of our requests. As indicated in the Midjourney documentation, a word that appears in the prompt is more likely to be represented in the image. We discovered that to get rid of an element, the user needs to use the *special command minus minus* (--), in other words: "/--no sun --no moon/" (Figure 12) without changing the prompt.

Figure 12

Prompt: /Landscape in Van Gogh Style --no sun, moon/, 2023



Note. Experiment by Enzo D'Armenio, Adrien Deliège & M. G. Dondero (2024), via Midjourney.

We consider this example to be very significant because we understand that denial in images is produced exclusively by going beyond the prompt and the level of translation that the machine can currently provide between the prompt and the visual form. Therefore, we need to use commands that allow us to act directly on the image without going through the translation process[21].

From the point of view of the enunciation enunciated, i.e. how the act of production is reflected in the enunciation produced, Midjourney can use a style for each painter and for each painting, and aims to develop it. In the case of Van Gogh, for example, the Midjourney uses the painter's typical texture and imitates a sensory motricity that is very similar to the rhythm of his touch. However, the machine has *its style*, which is close, in our view, to the American pictorial expressionism of the 1970s.

Mixing styles to test art history

In our case, what is particularly important is to test the mixing of different painters' styles according to their characteristics and to reflect on various

[21] Midjourney has also recently introduced a tool for modifying only a part or region of the image produced, previously selected by the experimenter (vary region): simply outline the part to be modified and insert a prompt that meets the experimenter's needs. This is a breakthrough, as modifications using this command are much more efficient than modifying a prompt directly, of course, if you want to make changes that are specific to you. For example, if we have already generated a man holding a ping-pong racket in his left hand, and we want him to hold a tennis racket, it will be (in our opinion, but this is open to testing) more efficient to use the new functionality by selecting the racket and entering the /tennis racket/ prompt, than to redo an entire prompt specifying all of this. What's more, redoing an entire prompt could change the image more than desired.

interesting situations that arise in composition. The computer-generated images allow us to understand how great works by artists of the past can be mixed and, in some cases, point out the most common stereotypes of each artist or artistic movement. Nevertheless, we might ask: why mix styles? What is the purpose of this mixture? Therefore, we can test which stereotypes of famous painters the database has learned and test combinations of styles that reveal, at least in part, how algorithms work in translating verbal and visual languages. However, how do we deal with the fact that the machine can produce styles or put them together?

As Wilf (2013) stated in an article inspired by Peircean semiotics, written ten years before the diffusion of ChatGPT and Midjourney, entitled "From Media Technologies That Reproduce Seconds to Media Technologies That Reproduce Thirds: A Peircean Perspective on Stylistic Fidelity and Style-Reproducing Computerized Algorithms":

> Unlike a CD or an MP3 file, which are reproduction technologies, these generative systems do not reproduce specific texts, or Seconds, but styles, or Thirds (GENERALITY). Their object of reproduction is the principle of generativity, responsible for the production of specific texts that are the object of reproduction of the kind of media technologies that have traditionally been at the center of linguistic and semiotic anthropological research. *The style reproduction of these systems consists both of their ability to abstract a style from a corpus of Seconds and to generate new and different Seconds or texts in that style, indefinitely* [emphasis added]. (p. 186)[22]

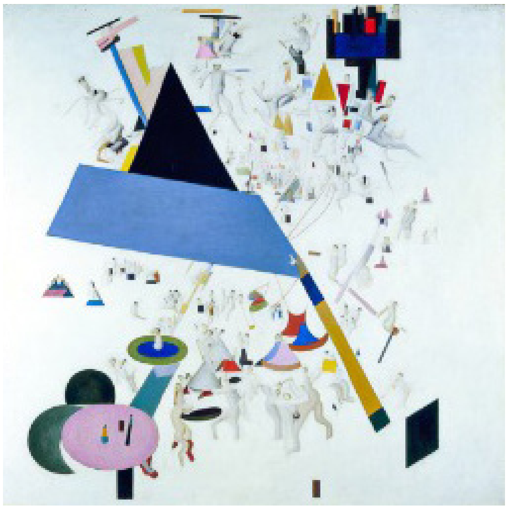
Undoubtedly, it is only through the forms already known and established in our cultural perception that it is possible to understand the work of the machine - not only the degree of its much-questioned "creativity", but also how it transforms the forms we know into averages. Lev Manovich's experiments (published on Facebook in 2023) and Manovich and Emanuele Arielli's (2021-2024) are very convenient in this sense: in them, Bosch's figures change according to the positions they occupy in the landscape, whose coordinates are given by geometric patterns inspired by Malevich, as shown in Figure 13.

If we look at another production by Manovich and Arielli (2021-2024), which mixes Brueghel and Kandinsky (Figure 14), it seems possible to argue that abstractionist artists such as Malevich and Kandinsky used the machine as landscapers. As you can see below, they end up determining the general topology of the image, which places the figures of painters such as Bosch and Brueghel, traditionally considered landscape painters. In other words, if we look

[22] On the original: "Thus, although these systems, much like a CD or an MP3 file, are technologies of reproduction, they do not reproduce specific texts, or Seconds, but styles, or Thirds. Their object of reproduction is the principle of generativity that is responsible for producing the specific texts that are the object of reproduction of the kind of media technologies that have traditionally stood at the center of linguistic and semiotic anthropological research. These systems' reproduction of style consists both in their ability to abstract a style from a corpus of Seconds and to generate new and different Seconds or texts in this style, indefinitely so".

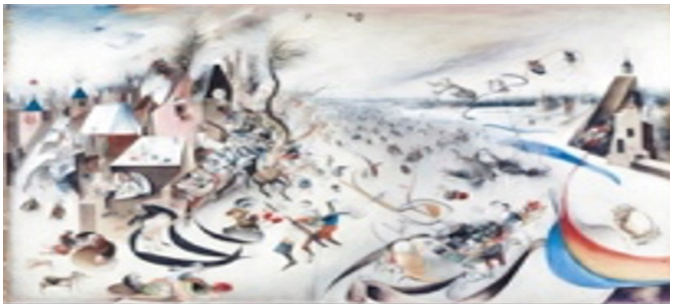
at the history of art, there is a reversal of roles: traditionally, abstract painters are not considered landscape artists, because this concept is no longer relevant in this context. However, the machine's work uses these abstract painters as frameworks that encompass figures inspired by real landscape painters such as Brueghel and Bosch.

Figura 13
Experiment with the prompt: /painted by Malevich and Bosch/



Note. Manovich and Arielli (2021-2024) via Midjourney.

Figura 14
Experiment with the prompt: */painted by Brueghel and Kandinsky/*

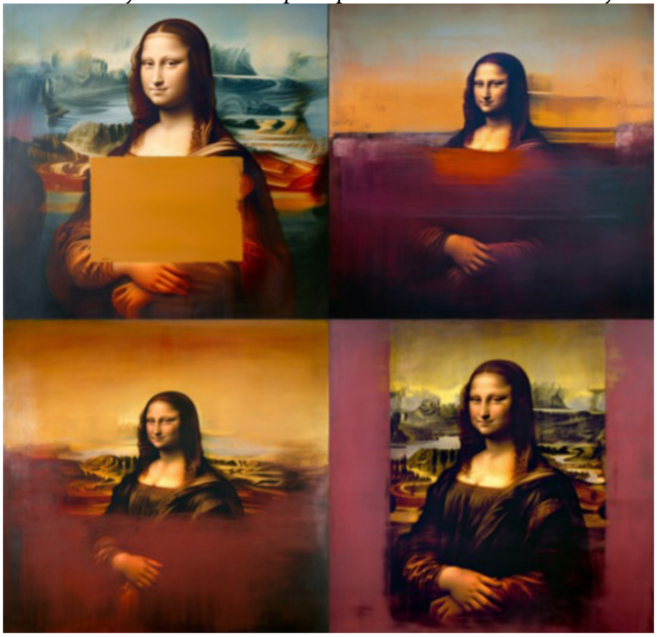


Note. Manovich & Arielli (2021-2024) via Midjourney.

We also tried to mix up some pictorial styles. The results are frustrating and, in some cases, amusing. One example is the mix between Da Vinci and Rothko. These two painters, separated by a few centuries, were recognized as specialists in atmospheric perspective and imprecise contours and layers of color, respectively. Some of the results were disappointing, for example: the machine gave us a Monalisa banally superimposed by a red-Rothko triangle. However, we got more interesting results when Rothko's layers of color, sometimes bordering on transparency, appeared superimposed on Da Vinci's atmospheric perspective (Figure 15).

Figure 15

Mona Lisa by L. Da Vinci + prompt: /Mona Lisa in Rothko style/



Note. Experiment by D'Armenio, Deliège and Dondero (2024) via Midjourney 4.

In these four images, we can see that the addition of blur and transparency transforms Da Vinci's landscape from blurry (due to the distance imposed by perspective) to sharp, reminiscent, in the latter case, of the hyper-realistic American paintings of the 1970s. Considering all these experiments, we are left with the following question: is Midjourney programmed to always achieve a balance between the blurred and the sharp, the imprecise and the detailed? Ultimately, we see that it is only through the production of a multitude of images, the mixing of styles, and production techniques—in other words, only by reiterating our requests—that we will be able to understand the space of language/virtualization that lies behind these productions. It is from an infinity of automatically generated images that we will be able to build hypotheses about the database on which Midjourney was trained and, therefore, about his model (kept secret).

From the point of view of enunciative praxis, the formal mechanism for renewing and maintaining cultures, this process is operationalized through its modes of existence. The database performs the process of virtuality/virtualization of forms, because the images of the painters it contains, in the case of Midjourney, can be seen as sedimented forms of our Western visual culture. The procedures that we trigger via prompts, on the other hand, can be seen as an updating stage carried out on the generated images. As far as potentiation is concerned, the words we produce, in other words, the statements generated through our prompts, will not—perhaps never—be immediately sedimented and accepted into the database, which is stabilized, and fixed. After all, we would have to become recognized artists to be able to do this and thus participate in the transformation of what is sedimented in the databases and, by extension, in culture itself.

CONCLUSIONS AND OPENINGS

We ended our research mainly with questions. The first concerns word-image translation: how can we explain that a generative intelligence can produce a coherent image in terms of composition from a verbal request? The second, which is closely related to the first, concerns the machine's perception: what kind of perception characterizes the image-generating machine? What kind of perception allows it to build a non-nonsensical and even coherent two-dimensional composition? Is it a perception that depends on the image datasets used in the training stage and therefore on the images that relate to the "vision" of others, a "vision" that is shared within the dataset? This is a perception that calculates the average of all the "stimuli received" by the different data sets. Like any perception, this machine perception must certainly be supported by an orientation and a perspective which, in this context, is given by the prompt, which has the task of triggering the databases. Furthermore, algorithms that direct it manipulate this type of perception that assembles the database. However, what kind of perception is this, since the machine has no body or sensations?

Although we can't give a definitive answer to these very general[23] questions, which we may be able to answer in a few years when generative artificial intelligence is more widely used and better studied, we can say a few things that seem certain to us: the results cannot be understood exclusively based on strict statistical-computational functioning, but also on other more socially pertinent factors. I will mention just two of them here, at the beginning and end of the transformation/translation chain. These two factors concern:

[23] Some provisional answers about perception can be found in D'Armenio et al. (2024) and D'Armenio, Deliège and Dondero (in press).

(i) the programmers' skills in human perception and the ideologies involved, with the programmers determining the correspondences between the numerical representations learned by the models, in the form of lists of numbers called embeddings, and the verbal and visual modalities;
(i) the objectives of the users of these models, can be aesthetic, artistic, commercial, or scientific.

In other words, and to summarize: it is necessary to study which cultural operations are at work in the transition between the word-image correspondences produced by programmers (the incorporation phase) and the circulation of automatically generated images (the implementation phase in front of an audience). M

REFERENCES

Beyaert-Geslin, A. (2017). *Sémiotique du portrait: De dibutade au selfie*. De Boeck Supérieur.

Colas-Blaise, M. (2019). Comment penser la narrativité dans l'image fixe? La "composition cinétique" chez Paul Klee. *Pratiques*, 181-182. https://doi.org/10.4000/pratiques.6097

D'Armenio, E., Deliège, A., & Dondero, M. G. (2024). Semiotics of Machinic Co-Enunciation About Generative Models (Midjourney and DALL·E), *Signata*, *15*. https://doi.org/10.4000/127x4

D'Armenio, E., Deliège, A., & Dondero, M. G. (in press). A semiotic methodology for assessing the compositional effectiveness of generative text-to-image models (Midjourney and DALL·E). In *Proceedings of the 1st workshop on critical evaluation of generative models and their impact on society, ECCV 2024*. Springer. https://orbi.uliege.be/handle/2268/321378

D'Armenio, E., Dondero, M. G., Deliège, A., & Sarti, A. (in press). Criteria for image generation. For a semiotic approach to Midjourney and DALL·E. *Semiotic Review*.

Deliège, A. (2024). Advances on the F.R.S.-FNRS research project "Towards a Genealogy of Visual Forms": On character poses in paintings. *Centre de Sémiotique et Rhetórique*. https://ceserh.hypotheses.org/3929

Deliège, A., & Dondero, M. G. (in press). The Semiotic and Computational Analysis of Represented Poses in Painting and Photography. In P. Conte, A.C. Dalmasso, M.G. Dondero & A. Pinotti (Orgs.), *Algomedia. The Image at the Time of Artificial Intelligence*. Cham; Springer.

Dondero, M. G. (2017a). Barthes entre sémiologie et sémiotique: le cas de la photographie. In J.-P. Bertrand (Org.), *Roland Barthes: Continuités* (pp. 365-393). Christian Bourgois.

Dondero, M. G. (2017b). The Semiotics of Design in Media Visualization: Mereology and Observation Strategies. *Information Design Journal*, *23*(2), 208-218. https://doi.org/10.1075/idj.23.2.09don

Dondero, M. G. (2019a). *Le travail des algorithmes. Quelques réflexions sur l'actantialité et l'énonciation* [Apresentação de trabalho]. Conferência da Associação Francesa de Semiótica. https://core.ac.uk/outputs/220155468/

Dondero, M. G. (2019b). Visual semiotics and automatic analysis of images from the Cultural Analytics Lab: how can quantitative and qualitative analysis be combined? *Semiotica*, *230*, 121-142. https://doi.org/10.1515/sem-2018-0104

Dondero, M. G. (2020). *The Language of Images. The Forms and the Forces*. Springer.

Dondero, M. G. (2022). P.D.R. F.N.R.S. Towards a genealogy of visual forms. *Centre de Sémiotique et Rhetórique*. https://ceserh.hypotheses.org/p-d-r-towards-a-genealogy-of-visual-forms

Dondero, M. G. (2023). Emerging Faces: The Figure-Ground Relation from Renaissance Painting to Deepfakes. In M. Leone (Org.), *The hybrid face: Paradoxes of the visage in the digital era* (pp. 74-86). Routledge.

Dondero, M. G. (2024). The Face: Between Background, Enunciative Temporality and Status. *Reti, Saperi, Linguaggi: The Italian Journal of Cognitive Sciences*. https://www.rivisteweb.it/doi/10.12832/113797

Dondero, M. G. (in press). Enunciación temporal en imágenes fijas. *Tópicos del seminario*.

Dondero, M. G., Rodrigues de Castro, G. H., & Schwartzmann, M. N. (2024). Inteligência artificial e enunciação: Análise de grandes coleções de imagens e geração automática via Midjourney. *Todas as Letras*, *26*(2), 1-24. https://editorarevistas.mackenzie.br/index.php/tl/article/view/17164

Eco, U. (1984). *Semiotica e filosofia del linguaggio*. Einaudi.

Focillon, H. (1934). *Vie de formes*. Presses Universitaires de France.

Fontanille, J. (1989). *Les espaces subjectifs. Introduction à la sémiotique de l'observateur*. Hachette.

Fontanille, J. (1998). *Sémiotique et littérature. Essais de méthode*. Presses Universitaires de France.

Fontanille, J. (1999). *Sémiotique du discours*. Presses Universitaires de Limoges.

Fontanille, J., & Zilberberg, C. (1998). *Tension et signification*. Mardaga.

Geoffre-Rouland, A. (2022, September 2nd). Midjourney: une œuvre d'art générée par l'IA remporte un concours et suscite l'indignation. *Tom's Guide*.

https://www.tomsguide.fr/polemique-une-oeuvre-dart-generee-par-lia-remporte-un-concours-les-artistes-sindignent/

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Hackett Publishing Company.

Greimas, A. J. (1984). Sémiotique figurative et sémiotique plastique. *Actes Sémiotiques Documents*, (60). https://www.unilim.fr/actes-semiotiques/3848

Groupe μ. (1998). L'effet de temporalité dans les images fixes. *Texte*, (21-22), 41-69.

Impett, L. (2020). Analyzing Gesture in Digital Art History. In K. Brown (Org.), *Routledge Companion to Digital Humanities and Art History* (pp. 386-406). Routledge.

Impett, L., & Moretti, F. (2017). Totentanz. Operationalizing Aby Warburg's Pathosformeln. *Literary Lab Pamphlet*, (16). https://litlab.stanford.edu/LiteraryLabPamphlet16.pdf

Lassègue, J. (2017). *Turing* (G. J. F. Teixeira, Trad.). Estação Liberdade.

Manovich, L. (2011, 4-6 de agosto). Style space: How to compare image sets and follow their evolution. *Manovich*. https://manovich.net/index.php/projects/style-space

Manovich, L. (2015). Data Science and Digital Art History. *International Journal for Digital Art History*, *1*(1), 3-35.

Manovich, L. (2017). The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. In M. K. Schäfer, K. Vanes (Orgs.), *The Datafied Society. Studying Culture through Data*. AUP.

Manovich, L. (2020a, 22 de novembro). Computer Vision, Human Senses, and Language of Art. *AI & Society*. http://manovich.net/content/04-projects/109-computer-vision-human-senses-and-language-of-art/manovich_computer_vision.pdf

Manovich, L. (2020b). *Cultural Analytics*. MIT Press.

Manovich, L., & Arielli, M. (2021-2024). *Artificial Aesthetics: Generative AI, Art and Visual Media*. http://manovich.net/index.php/projects/artificial-aesthetics

Manovich, L., & Douglass, J. (2009). Timeline. 4535 Time Magazine Covers, 1923-2009. *Cultural Analytics Lab*. http://lab.culturalanalytics.info/2016/04/timeline-4535-time-magazine-covers-1923.html

Manovich, L., Douglass, J., & Zepel, T. (2011). How to Compare One Million Images? In D. Berry (Org.), *Understanding Digital Humanities* (pp. 249-278). Palgrave Macmillan.

Marin, L. (1993). *De la représentation*. Seuil.

Meyer, R. (2023). The New Value of the Archive: AI Image Generation and the Visual Economy of 'Style'. *IMAGE. Zeitschrift für interdisziplinäre Bildwissenschaft*, *19*(1), 100-111. http://dx.doi.org/10.25969/mediarep/22314

Pinotti, A. (2001). *Il corpo dello stile: Storia dell'arte come storia dell'estetica a partire da Semper, Riegl, Wölfflin*. Mimesis.

Petitot, J. (2004). *Morphologie et esthétique*. Maisonneuve et Larose.

Santaella, L., & Kaufman, D. (2024). A Inteligência artificial generativa como quarta ferida narcísica do humano. *MATRIZes*, *18*(1), 37-53. https://doi.org/10.11606/issn.1982-8160.v18i1p37-53

Thom, R. (1983). Local et global dans l'œuvre d'art. *Le Débat*, *2*(24), 73-89.

Warburg, A. (2012). *L'Atlas Mnémosyne*. Éditions Atelier de l'écarquillé.

Wilf, E. Y. From Media Technologies That Reproduce Seconds to Media Technologies That Reproduce Thirds: A Peircean Perspective on Stylistic Fidelity and Style-Reproducing Computerized Algorithms. *Signs and Society*, *1*(2), 185-211. https://doi.org/10.1086/671751

Article received on October 17th, 2024 and approved on October 23rd, 2024.