A THEORY OF LATENT SPACES

Antonio Somaini

A theory of images and visual culture, today, needs a theory of latent spaces. In a historical phase in which images are more and more generated, modified, circulated, seen, and described by or with the help of different kinds of AI models, we need to understand the crucial role played by an abstract, mathematical construct whose cultural and political implications could hardly be overestimated.[1]

"Latent space"—a term with significant metaphorical connotations to which we will return later—is a foundational concept in machine learning and artificial intelligence. It refers to the abstract space within which complex, high-dimensional data structures (such as images, texts, and sounds, or whatever entity that may be translated into a digital form) are represented in a more simplified, lower-dimensional form, in order to be processed through different mathematical operations. Latent spaces are made of mathematical objects called vectors (lists of numbers, arranged in a specific order) which represent data points in a multi-dimensional space. Each vector, with its n number of dimensions, represents a specific data point with its n number of coordinates: these capture some of the specific features or attributes of the data point, and determine its position in relation to other data points.

1 I would like to thank for the discussions on latent spaces and for their comments on different versions of this text Ada Ackerman, Hannes Bajohr, Francesco Casetti, Grégory Chatonsky, Cécile Chièze, Kate Crawford, Noam Elcott, Alexandre Gefen, Alexandra Gilliams, Red Giuliano, Alban Leveau-Vallier, Lev Manovich, Fabian Offert, Pia Viewing.

As an integral component of deep learning neural networks —the most common and widely used paradigm in AI models, at this present moment, in the constantly shifting field of so-called "artificial intelligence"—latent spaces are also having a profound impact on contemporary visual culture. Their role is complex and multifaceted, and touches all aspects of our relationship with images.

To begin with, latent spaces are a key component of the machine vision systems that during the last few years have turned the entire realm of digital images into a vast field for data mining and aggregation: they determine the epistemological field of these systems, what they can and what they cannot "see" (i.e. detect, recognize, and classify). In the domain of generative AI, latent spaces enable the generation and the modification of still and moving images, starting from textual prompts and/or from other images, as well as the generation of captions and texts starting from images, thereby reconfiguring profoundly the relations between images and words, between the visible and the sayable. Latent spaces are also crucial to the functioning of recommendation systems, targeted advertising, and social media algorithms, heavily influencing the circulation and reception of images across networks and platforms, channeling cultural consumption and contributing to the formation of tastes, trends, and behaviors. More broadly, as vast ensembles of data points within which billions of connected images and texts have been encoded and out of which new images and new texts may emerge, latent spaces play a key role in the processing and transformation of the massive quantities of visual and textual content that are stored on the internet. As AI models become more and more pervasive, and as internet contents keep growing exponentially, latent spaces become a way of ordering, processing, and activating a hypertrophic accumulation of cultural memory that has become unmanageable and disorienting.

Given these premises, it is not surprising that latent spaces also play a central role in contemporary artistic practices that engage with AI: whether to critically respond to its increasing presence in every aspect of culture, society, politics, and economics, or to use it as a series of new media for artistic production. For a few years now, artists have developed different strategies to explore or modify the

existing, dominant latent spaces, or to produce their own alternative, antagonist, counter-hegemonic ones. Several of these strategies are documented in the exhibition *The World through AI*: taken together, they show the awareness with which the field of contemporary art is tackling the presence and the agency of this hidden layer of mathematical abstraction that is profoundly transforming the status of images and vision, as well as the relations between images and other media.

But what are latent spaces exactly? And how can we describe their origins, their structure, their limits and potentialities, their aesthetic, epistemic, and political implications?

I. MATHEMATICAL SPACES OF COMPRESSION, REPRESENTATION, AND MAPPING

To begin with, since they assign specific coordinates to each of their data points, latent spaces may be considered to be a form of *mapping*. Even though we can neither perceive nor imagine a space with hundreds or thousands of dimensions, latent spaces are "spaces"—more precisely, "vector spaces"—with their own structure, coordinates, and positions. Within them, one finds relations of proximity and distance: objects that are somehow *similar* (for example, the words "puppy" and "dogs" in the English language) (see fig. 1) are positioned *close* to one another, while objects that are somehow *different* are positioned *far* from each other.

Just like maps, latent spaces are the result of an operation of *compression* and *dimensionality reduction*. These are meant to reduce the computational complexity and the memory requirements of the mathematical operations performed in latent spaces (see fig. 2). By discarding features that an AI model considers to be irrelevant or redundant, latent spaces capture and preserve other underlying features in the original, higher-dimensional data structures. This allows for different kinds of processing: for example, analyzing, classifying, clustering, and visualizing the original data according to different criteria, or generating new data (new images, texts, or sounds) based on the features that have been identified in the original ones.

This process of compression, reduction, and abstraction, though, is highly problematic. By turning images into vectors, AI algorithms focus on certain features—for example: shape, color, lighting, camera lens, distinction between figure and background—while ignoring others. The logic behind this selection, this filtering, remains mostly inaccessible. It is a logic that is only partially based on human decisions: the other reasons behind it have to do with the structure and operations of deep learning algorithms that are mostly opaque and undecipherable by humans. Understanding why only certain features are preserved within the vast system of vectors that is a latent space is often impossible; Still, these selected features are the ones that will determine which kinds of images can be generated or modified, and which kinds cannot be.

Finally, despite their being "lower-dimensional" if compared with the complexity of the data they are meant to represent, latent spaces may have a very high number of dimensions. To give an example, a single still image with a 4K resolution (3840×2160 pixels) has approximately 8.3 million pixels, each of which has three color values (red, green, and blue in the RGB system). In its original "pixel space," this image has, therefore, approximately 8.3×3=24.9 million dimensions or parameters, which would be represented by a vector with 24.9 million values. Even though compressed to a lower number of dimensions, the latent space representation (the latent vector) of this single image may still have hundreds or thousands of dimensions.

Therefore, while it is true that latent spaces are maps, they are maps that no human eye could ever explore in their totality.

II. GENEALOGIES

A long history of latent spaces—written retroactively, from the point of view of the current AI models—could include all those techniques and forms of representation that reduce the number of dimensions of the objects or phenomena they represent, in order to position them in a new space and perform on them or with them different kinds of operations. Geographical maps should definitely be part of this history,

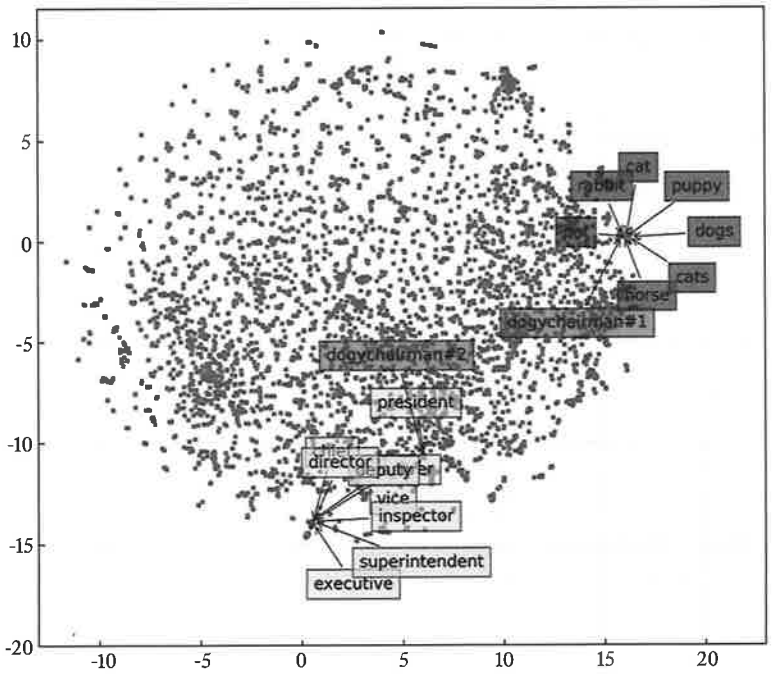


Fig. 1: T-SNE projections of word embeddings

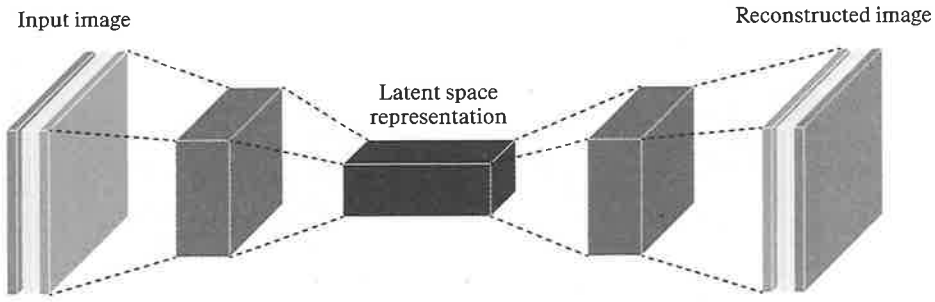


Fig. 2: Latent space as compression in a Variational Autoencoder

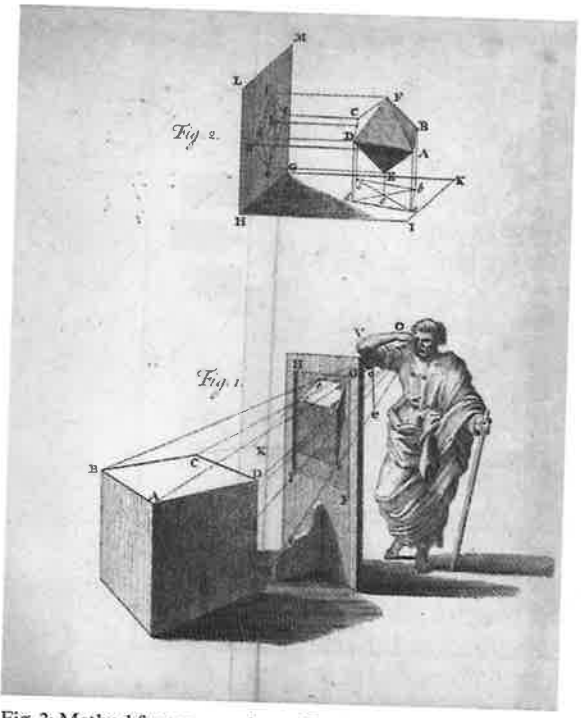


Fig. 3: Method for transposing a 3D object into a 2D representation according to the laws of perspective. Illustration taken from Brook Taylor, *New Principles of Linear Perspective: Or the Art of Designing on a Plane the Representations of All Sorts of Objects, in a More General and Simple Method than Has Been Done Before*, London, John Ward, 1719, pl. I

Fig. 4: Albrecht Dürer, *Man Drawing a Female Nude*, 1525. Wood engraving, 21.5×75 cm. Taken from Albrecht Dürer, *Underweysung der Messung* [Instructions on the way to measure], Nuremberg, 1525. Musée du Louvre, Département des Arts Graphiques, collection Rothschild. Inv.: L 37 LR/316 Recto

as should 2D geometric projections of 3D objects (for example, perspectival, orthogonal, or axonometric projections) (see fig. 3), as well as 3D models of larger objects, all of which capture certain features of what they represent while discarding others. Grid-based systems of representation could also be included, with their various applications in terms of reproduction, classification, and planning (see fig. 4).[2] Game boards, such as the ones used for chess and Go, could also be included, as complex, structured, strategic spaces for rule-based operations. We could then add catalogs, indexes, and card files, as systems used to describe, classify, search, and retrieve different kinds of objects and data (see fig. 5), as well as picture atlases, arranging constellations of images on bidimensional plates in order to discover and visualize connections, similarities, migrations of forms, gestures, motifs (see fig. 6).[3] More broadly, a long history of latent spaces could include all the attempts to find the latent, underlying structures that may explain complex observable phenomena and that allow us to operate on them.

The more direct sources of the concept of latent space currently used in the field of machine learning derive instead from a series of fields in the history of mathematics, statistics, and psychology. During the first decades of the twentieth century, psychologists such as Charles Spearman and Louis Thurstone, both pioneers in the field of psychometrics and factor analysis, introduced the idea of "latent variables" (variables that cannot be directly observed but are rather inferred from other observable variables) in their studies on human intelligence and personality traits.[4] Later, statistical techniques such as Principal Component Analysis (PCA) (introduced by Karl Pearson in 1901, and developed independently by Harold Hotelling in the 1930s) and Multidimensional Scaling (MDS) (introduced in the 1930s but then further developed in the 1960s) formalized the process of representing high-dimensional data in lower-dimensional spaces, typically 2D or 3D.[5] The 1980s and 1990s saw the rise of artificial neural networks, such as Autoencoders, which were meant to learn how to compress given data into lower-dimensional latent spaces and then reconstruct them in their original form. Finally, the rapid development of deep learning techniques during the 2000s and 2010s led to the widespread presence of the term "latent space" in the field of AI.

2 On latent spaces and the cultural technique of the grid, see Noam Elcott's response to the "Questionnaire on Art and Machine Learning," in *October*, 189, summer 2024, pp. 41–45. On grids as cultural techniques, see Bernhard Siegert, "(Not) in Place: The Grid, or, Cultural Techniques of Ruling Spaces," in *Cultural Techniques: Grids, Filters, Doors, and Other Articulations of the Real* (New York: Fordham University Press, 2015), pp. 97–120.

3 In his contribution to this volume, Fabian Offert develops a very interesting comparison between the relationships of similarity and difference between images in the latent spaces of AI models and the same relationships in the plates of Aby Warburg's *Mnemosyne Atlas*.

4 See Charles Spearman, "General Intelligence, objectively determined and measured," *The American Journal of Psychology*, 15, pp. 201–292, 1904; Louis Thurstone, "The Vectors of the Mind," *Psychological Review*, 41, pp. 1–32, 1934.

Today, latent spaces are a key component of all the deep learning models currently used to generate, modify, circulate, see, and describe images. We find them in the Convolutional Neural Networks (CNNs) used by the systems of machine vision, as well as in the various neural networks, algorithms, methods, and models that, beginning with the mid-2010s, have contributed to the rapid development of generative AI: Variational Autoencoders (VAEs) (introduced in 2013),[6] Generative Adversarial Networks (GANs) (2014),[7] Transformers (2017),[8] Contrastive Language-Image Pre-training (CLIP) (2021),[9] as well as the Latent Diffusion Models (LDMs) which are used by programs such as Stable Diffusion, DALL-E, and Midjourney (all released in their first versions in 2022), or text-to-video models such as Sora and Gen-3 (2023).

In the case of all these models, as we will see more in detail later, images and fragments of images are represented in latent spaces in a compressed form as n-dimensional vectors. This allows for a wide array of image operations —such as morphing, blending, outpainting, inpainting, upscaling, denoising, deblurring, restoring, and style transfer—which were not possible, or not possible in this way, before the recent developments in the field of AI.

One of the consequences of the pervasive presence of AI models with their latent spaces across the entire field of contemporary visual culture, is the need to rethink, from this perspective, many of the concepts that have been traditionally used to describe and analyze images: concepts such as resemblance, imitation, reproduction, style, referent, index, photorealism, description, *ekphrasis*.[10] How can we discuss what is the "referent" of an AI-generated image without mentioning the fact that this image stems out of a vector in latent space? How do we understand how images may be generated through texts, and texts be generated through images, without studying the latent spaces in which billions of connected words and images have been encoded? How do we justify the photorealistic aspect of an image generated with an AI model such as Stable Diffusion, DALL-E, or Midjourney without knowing that this aspect depends on the way in which the term "photorealistic" is connected in latent space with certain images and image features rather than others?

5 See Karl Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2, pp.559–572, 1901; Harold Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, 24, pp. 417–441, 498–520, 1933.

6 The term "variational autoencoder" was first coined by D.P. Kingma and M. Welling in their paper entitled "Auto-Encoding Variational Bayes" (2013), arXiv:1312.6114.

7 For the first presentation of Generative Adversarial Networks, see Ian Goodfellow et al., "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 27, ed. Z. Ghahramani et al. (San Diego: NeurIPS, 2014), 2672–2680.

8 The first paper introducing Transformer models was A. Vaswani, N. Sdhazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems*, 30 (2017).

Fig. 5: Card catalogs room, New York Public Library, 1911–1930. The New York Public Library Archives. Inv.: 1153322

Fig. 6: Aby Warburg, *Mnemosyne Atlas*, panel 48, final version, October 1929. The Warburg Institute, School of Advanced Study, University of London

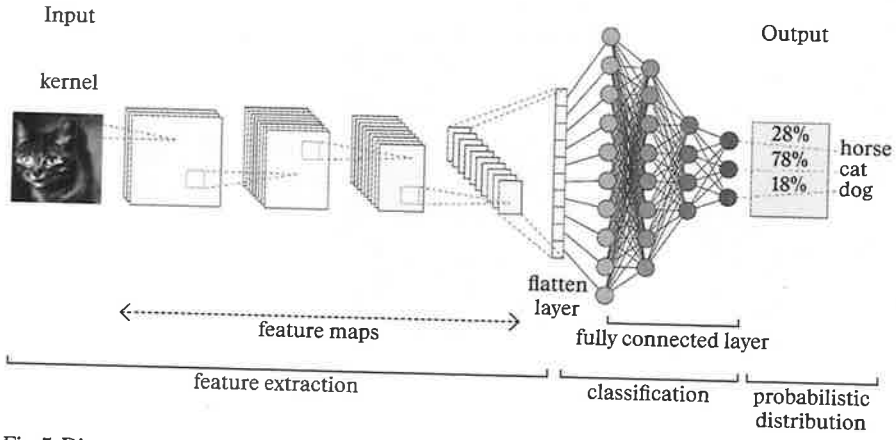


Fig. 7: Diagram of a Convolutional Neural Network (CNN)

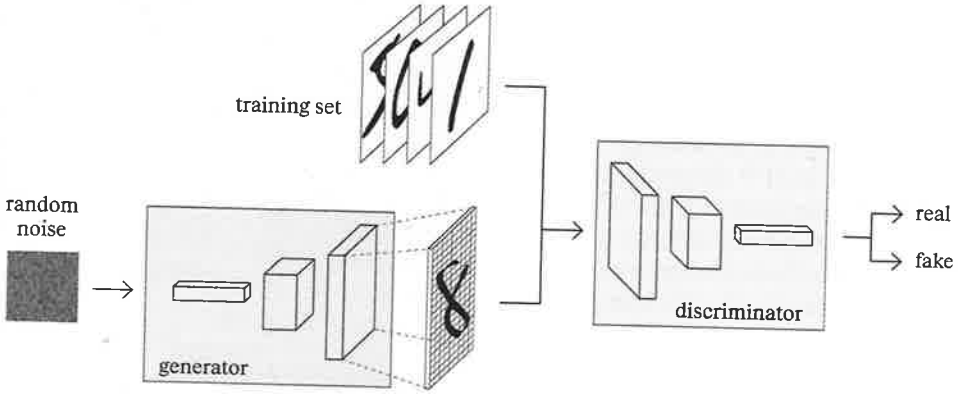


Fig. 8: Diagram of a Generative Adversarial Network (GAN)

Processes and dynamics within visual culture—such as the circulation of styles and motifs, or the dynamics of influence and reception—also need to be tackled having latent spaces in mind, since they heavily condition, with their tendency towards probabilistic averaging, the new forms of image making, even in the initial phase of "photographic" image capture.[11]

Broadly speaking, all the AI-generated and AI-modified images that are currently flooding visual culture are strictly dependent on the latent spaces of AI models and on the training sets of which such latent spaces are a representation. They are the product of the specific architecture of the latent spaces out of which they emerge, and of the specific statistical and predictive operations that are performed within it. This is why we cannot limit ourselves to speaking of "AI" or "AI models" in general, but rather need to try to understand them with their differences and specificities. This is also what artists working with AI are doing, when they engage with different AI models and latent spaces, treating them as artistic media, exploring their possibilities and limitations.

III. LATENT SPACES IN ANALYTIC AI AND GENERATIVE AI: FROM MACHINE VISION TO MULTI-MODAL MODELS

Latent spaces are learned by deep learning models during their training phase. Given an initial dataset, these models learn how to compress and represent the data into latent space in order to then perform various mathematical operations. Within contemporary visual culture, this happens both in the models used in the field of analytic AI, and in the ones used for generative AI: both in the AI models used to detect, recognize, and classify objects, bodies, and faces represented in images, and in the ones used to generate or modify images.

Convolutional Neural Networks (CNNs)—the deep learning architecture that contributed to the fast development of machine vision systems during the early 2010s—consist of multiple layers of artificial neurons, each of which transforms the image the network is meant to analyze into increasingly abstract representations (see fig. 7). In its

9 For the first presentation of the CLIP model, see A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 2021.

10 See Antonio Somaini, "Algorithmic Images. Artificial Intelligence and Visual Culture," *Grey Room*, 93 (fall 2023), pp. 74–115.

11 On the role of the embedded AI algorithms in recent cameras, and on the broader question of the relations between AI and photography, see Estelle Blaschke, Max Bonhomme, Christian Joschke, Antonio Somaini (eds.), *Photography and Algorithms, Transbordeur*, 9 (2025).

superficial layers, the network captures simpler features such as edges, textures, and basic shapes. The deeper layers, instead, capture more complex features such as object parts or even whole objects. Taken together, these representations, especially the ones of the deeper layers, constitute the CNN's latent space, and play a key role in the ways in which the model detects, recognizes, and classifies. Latent spaces, in other words, are a key component of machine vision systems: those systems that, for a few years now, have introduced in contemporary visual culture a new kind of nonhuman, algorithmic, automated visual perception that is playing an increasingly important role in the ways in which images are analyzed and activated for purposes of data extraction, control, and surveillance. The so-called "faceprints" generated by AI algorithms—those digitally encoded representations of a person's facial features that are generated by face recognition technologies—are vectors with dozens or hundreds of numerical values: compact, abstract latent space representations which capture what the algorithm identifies as the unique, essential features of a given face, and position it in relation to the "faceprints" of other faces in latent space, allowing for different operations of comparing, matching, or morphing different faces with one another.

If we move to the field of generative AI, we find a series of models, each of which constructs and activates its latent spaces in a different way.

Variational Autoencoders (VAEs) (see again fig. 2) learn how to encode images from their initial pixel space into latent space representations, and then decode them back to pixel space. Through this process—which involves two networks, an encoder and a decoder—they learn how to separate features that the model considers to be meaningful from features that are considered to be less relevant. Once the model is trained, new images may be generated from the model's latent space, by running the latent space representations through the decoder.

Generative Adversarial Networks (GANs) (see fig. 8), widely used by artists during the second half of the 2010s, also use two networks—called Generator and Discriminator—in order to learn how to generate from random vectors in a pre-defined latent space (usually a high-dimensional vector

space with a simple, uniform distribution) images that are highly similar to the ones of the initial training set. For example, if the training set is made of photographs of human faces, the model learns to generate photorealistic images of human faces[12] (see fig. 9).

At the end of training, even though the structure of the pre-defined latent space has remained unchanged, GANs can be used to generate different kinds of images. This allows users to manipulate a series of parameters in order to explore different areas of the model's latent space, generating new images that are "interpolations" in latent space: images that visualize intermediate data points (represented by vectors) which lie between other data points. Depending on the training set and on the parameters that are activated, these new images may be photorealistic, hybrid, or abstract. Each of them, though, has exactly the same status: it is the visualization of a data point in the model's latent space. Users can also generate moving images that connect different interpolations and therefore visualize different trajectories within latent space, moving from data point to data point. Transitions between data points in a GAN's latent space often look like a form of morphing, which can be smooth or abrupt depending on the number of images that compose the sequence.

Introduced for the first time in 2017, Transformers ushered in a new phase in AI's impact on images and visual culture: a new phase characterized by new algorithmic connections between images and words, and by the possibility of using AI models to perform both *text-to-image* and *image-to-text* operations.[13] Transformers were initially used in the field of natural language processing (NLP), and are at the basis of large language models such as the different versions of GPT (Generative Pre-Trained Transformers), including their popular chatbot version, ChatGPT.[14] One of their specificities, thanks to a feature called "self-attention mechanism," is the capacity of processing simultaneously *sequential data*, such as texts, by converting each element of a text into latent space representations (vectors) that capture the semantic meanings of words and their relations of similarity and difference with all the other words of a given language. This embedding of words and texts in latent space can then be combined with other deep learning processes

12 In 2019, the website thispersondoesnotexist.com, created by the software engineer Phillip Wang using StyleGAN2, an algorithm invented by three Nvidia computer scientists (Teo Karras, Samuli Laine, and Timo Aila), went viral for its capacity, each time the web page is refreshed, to generate highly photorealistic portraits of non-existing people. For the paper presenting StyleGAN2, see https://arxiv.org/abs/1812.04948.

13 On the new algorithmic relationships between images and words, the visible and the sayable, see my response to the already mentioned "Questionnaire on Art and Machine learning," in *October*, 189, summer 2024, pp. 112–120.

14 On ChatGPT, see Alexandre Gefen, *Vivre avec ChatGPT. L'intelligence artificielle aura-t-elle réponse à tout ?* (Paris : L'Observatoire, 2023).

that are used to generate images from texts or texts from images, as it happens with CLIP and with the Latent Diffusion Models.

CLIP (Contrastive Language-Image Pre-training) is entirely based on a Transformer architecture. Trained with large quantities of connected texts and images (so-called "text-image pairs") scraped from the internet, it learns how to encode and position them in a common, shared latent space where their relations can be mapped and operationalized. Two kinds of neural networks—an image encoder and a text encoder—convert the images and their textual descriptions into latent space representations made of vectors. Both neural networks are Transformers: the text encoder is similar to those used in natural language processing, while the image encoder uses a so-called Vision Transformer (ViT) which divides images in patches and treats them as if they were tokens in a sequence. Once both words and image patches have been positioned in the common latent space, CLIP learns through a "contrastive" process how to bring the embeddings of matching text-image pairs closer together, while pushing non-matching pairs further apart. The goal is to allow the model to find, by searching in its own latent space, images that match a given textual description (a prompt), or textual descriptions that match a given image.

It is important to underline that CLIP is not a generative model: it does not generate images nor text, but rather connects them. For this reason, it has often been used to explore in new, unprecedented ways, vast ensembles of images contained in the databases of archives and collections. In these cases, starting from a given prompt and from the ways in which the prompt is connected to other images in the model's latent space, CLIP searches in the database for images that somehow match the prompt, even if such images are "unlabeled," that is, not connected to captions or to other types of textual metadata. Reversing the process, CLIP may start from an image which is part of the database, and then generate one or more captions or descriptive texts by analyzing the content of the image and comparing it to a large corpus of images and associated descriptions. A third kind of search consists in starting from an image, and searching in the database for images that CLIP considers to be somehow visually and semantically similar.[15]

15 On this third kind of search, see Leonardo Impett and Fabian Offert, "There Is a Digital Art History," *Visual Resources*, 38 (2) (2022), pp. 186–209.

Fig. 9: Photorealistic portrait generated with the model StyleGAN2. Taken from the site https://thispersondoesnotexist.com/

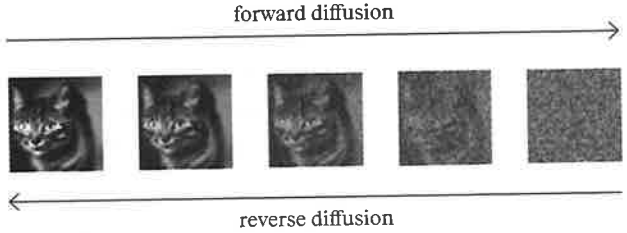


Fig. 10: Noising and denoising (forward and reverse diffusion) in Latent Diffusion Models

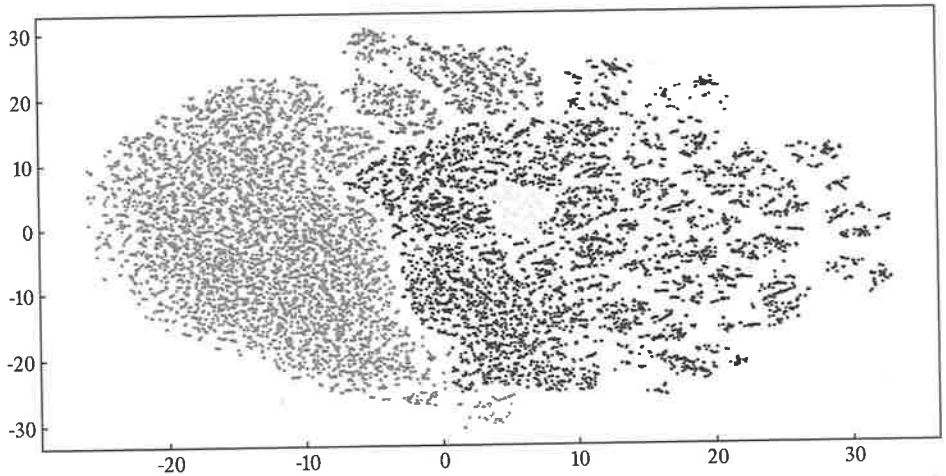


Fig. 11: Projection in 2D of a latent space via t-SNE

By enabling these three kinds of processes—from texts to images, from images to texts, from images to images—CLIP is an example of how certain so-called "foundation models," through their latent spaces, are reorganizing the relations between images and words and those between images and images, while at the same time transforming the ways in which knowledge *about* images and knowledge *through* images is constituted.

This is also the case of the Latent Diffusion Models (LDMs) that, after having been released in their first versions in 2022, are now currently used for generating or modifying still and moving images. LDMs generate images by denoising an initial image made of pure noise (random pixels), and they learn to do so after having been trained with vast quantities of images. During the training phase, they gradually add noise to a given image, and then learn to predict how much noise has to be removed, step by step, in order to find again the initial image. This noising and denoising process, which is repeated a very high number of times in the training phase, happens in a latent space in which images have been encoded from their initial high-dimensional pixel space into a lower-dimensional latent space representation (see fig. 10).

Once the model is trained, it has learned not only how to find again an image to which noise has been added, but also how to modify a given image, or to generate new images that were not part of the training set. When these operations of image modification and image generation are activated by textual prompts, as it happens with all text-to-image and text-to-video models, the words of the prompts themselves are encoded into latent space representations by a text encoder, which is a pre-trained transformer model (GPT) similar to the ones used for natural language processing. This representation then steers the diffusion process in order to generate an image that somehow matches the textual description.

Both largely relying on Transformer architectures, CLIP and LDMs are examples of how latent spaces are used by "multimodal" or "cross-modal" AI models, i.e. models capable of bridging the gaps between different kinds of data, such as images, texts, and sounds. With these models, what we are used to call "intermediality" is more and more

dependent on latent spaces. Further deepening the process of convergence that was already promoted by the diffusion of digital media, latent spaces are the abstract realm in which objects and phenomena existing in different kinds of media are turned into one single medium—latent space representations made of vectors—in order to be processed through mathematical operations and then re-materialized in different media.

Latent spaces, in other words, are spaces of *transformation*: vast matrices of numbers within which media may be transformed into other media.

4. THE "LATENCY" OF LATENT SPACES

As vast, organized structures of multi-dimensional vectors, latent spaces are *invisible* and *unimaginable*. They are not visible by human eyes, even though techniques such as t-Distributed Stochastic Neighbor Embedding (t-SNE) (see fig. 11), Uniform Manifold Approximation and Projection (UMAP), or Self-Organizing Maps (SOMs) can be used to project their multiple dimensions into lower, 2D or 3D representations. These techniques, which bring latent spaces into a field of visibility that is accessible by humans, are used to visualize the *positions* of the different data points and the spatial relations of distance and proximity between them.

Generating still and moving images out of latent spaces is of course another way of somehow visualizing them, even though only fragmentarily. Users of image-generating AI models often describe the images they generate as a way of exploring the models' latent spaces. Considered in their totality, though, latent spaces remain radically invisible and inaccessible: they are not meant to be perceivable by humans, but rather operationalized by machines.

But what does it mean, exactly, to consider these abstract vector spaces as "latent"?

Taken literally, the adjective "latent" refers to the fact that they consist of vectors representing "latent variables": series of numbers which capture underlying features that are not immediately observable in the higher-dimensional input data.

Fig. 12: A latent photographic image

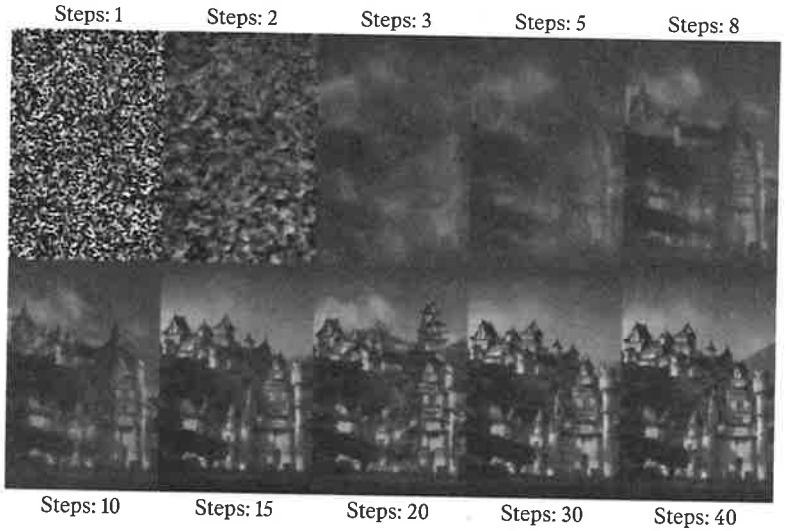


Fig. 13: An AI-generated image in the process of being denoised

Broadly speaking, though, "latent" has a number of significant connotations. "Latent" is what is present but not immediately visible or perceivable; what is hidden but could become manifest; what is inaccessible but nevertheless active. The state of latency, in other words, is not static, but rather active and dynamic. It is a state of virtuality: the state of virtualities that might or might not be actualized.[16]

If we keep these connotations in mind and come back to the field of generative AI, we may establish a series of analogies and connections.

To begin with, we can say that the term "latent space" underlines the "black box" nature of AI models: their radical inaccessibility for human subjects, and at the same time the human need for strategies and techniques capable of making them more understandable and interpretable.[17] Latent spaces are an almost emblematic example of the "encrypted" nature of many digital technologies of which only the interface, together with the inputs and the outputs, are accessible.[18]

Then, if considered from the point of view of image and media theories, the term "latent space" invites us to establish comparisons with the concept of "latent image" in analog photography: the invisible indexical trace that has formed on a photographic film or paper *after* it has been exposed to light but *before* it has been developed into a visible image (see fig. 12). Just as the latent images of analog photography need a chemical process to become visible, the multiple data points of the latent space of AI models need a complex generative process in order to become visible. And just as the latent image takes time to form and appear, so do the AI-generated images: through the interfaces of Latent Diffusion Models that allow users to see the gradual process of denoising, as it happens with Midjourney and certain versions of Stable Diffusion, we see the images emerge slowly out of a pixelated blur, as if through a process of algorithmic *pareidolia* (see fig. 13).

The analogies between latent spaces and latent photographic images, though, have their limits. Latent photographic images are direct, indexical traces of reality: traces of the light reflected from objects onto a photosensitive film. The relation between the latent spaces of AI models and whatever exterior "reality" they represent is not direct but

16 On the relationship between the virtual and the actual, see Gilles Deleuze, "The Actual and the Virtual," in Gilles Deleuze and Claire Parnet, *Dialogues II*, revised edition, (New York, N.Y.: Columbia University Press, 2007), pp. 148–159.

17 On the "black boxing" of the deep neural networks used in contemporary AI and its historical and political implications, see the dialogue with Kate Crawford in this volume.

18 On the way in which contemporary artists are reacting to the "encrypted" nature of contemporary technologies, see, Nadim Samman, *Poetics of Encryption. Art and the Technocene* (Berlin: Hatje Cantz, 2023), a book that was the basis for the exhibition presented at KW Institute for Contemporary Art in Berlin in 2023–24.

rather highly mediated since it implies the process of encoding objects and phenomena into numerical vectors.

Finally, the term "latent" may also be interpreted as referring to the fact that what is encoded in latent spaces are underlying features of cultural objects that are not fully perceivable nor understandable, but that play a key role in further processes of cultural production. In this sense, a latent space's "latency" is its capacity of capturing, representing, and operationalizing aspects of culture that are not immediately perceivable by human subjects.

The agency of latent spaces, in other words, increases the nonhuman share in human culture.

5. PRODUCING, EXPLORING, AND MODIFYING LATENT SPACES

What we find now across the current visual culture landscape, is a series of different latent spaces: each one with its own structure, size, content, possibilities, and limitations. As documented by many of the artworks presented in *The World through AI*, contemporary artists have experimented with many ways of engaging with latent spaces: either by producing them themselves (starting from carefully selected training sets and using AI models such as GANs), or by exploring or modifying existing latent spaces (such as the ones of the currently most popular text-to-image models, such as Stable Diffusion, DALL-E and Midjourney), or by trying to influence future ones (finding ways to release across the internet large quantities of carefully chosen images that might end up in future training sets).

Beginning with the second half of the 2010s, artists used different kinds of GANs to begin exploring for the first time the latent spaces of generative AI models. By training these models with specifically selected sets of images, they were able to control, at least in part, the images that could be generated, while still being exposed to the unpredictability that the use of any AI model implies.

Trevor Paglen's *Adversarially Evolved Hallucinations* (2017), for example, are meant to allow viewers a glimpse into the inner workings and the alien logics of the deep neural networks that are transforming both images and vision in

contemporary visual culture. After gathering images from various sources in a series of training sets with titles such as *Interpretations of Dreams*, *Monsters of Capitalism*, *American Predators*, *Omens and Portents,* and *Things That Exist Negatively* (see fig. 14), Paglen used GANs to generate out of latent space images which visualize the model's "hallucinations": hybrid, blurry images that are result of the model's repeated attempts to distribute pixels in order to replicate as closely as possible the images of the training set.[19]

Hito Steyerl also used different kinds of GANs in installations such as *Power Plants* (2018), *This Is the Future* (2019), *SocialSim* (2020), and *Animal Spirits* (2022), a video in which she describes her role in the credit sequence as "latent space architecture and pathmaking." In *This Is he Future*, in particular, she used a next-frame prediction algorithm in order to generate images that she presents as located "0.04 seconds in the future," thus emphasizing and questioning the *predictive* nature of AI systems (see fig. 15).[20]

The release, in 2022, of diffusion models trained with massive datasets and through quantities of computational resources to which no single individual could ever have access, opened a new phase for artists, in which generating latent spaces of similar dimension and complexity has become impossible: what is possible, instead, is either to explore in different ways the existing latent spaces, or to modify them through various techniques of so-called "fine-tuning."

The latent spaces of the most commonly available diffusion models may be explored in different ways: starting from textual prompts, from combinations of prompts and images, or just from images.

When these models are used to perform text-to-image or text-to-video operations, prompts point to specific areas in latent space and lead to the generation of certain images rather than others. The process, though, is not straightforward: the data points are so many and so dense, and the stochastic dimension of the diffusion process so important, that the same prompt, if repeated n number of times, can generate long series of slightly different images.
With prompts, language becomes a new medium for image production, in a completely unprecedented way. Prompts act as a new kind of "speech acts"[21] and as a form of "operative

19 On this series of works, see Trevor Paglen, *Adversarially Evolved Hallucinations*, ed. by Anthony Downey (Berlin: Sternberg Press, 2024).

20 On Steyerl's recent work, see Florian Ebner, Doris Krystof, and Marcella Lista, eds., *Hito Steyerl: I Will Survive*, exh. cat. (Leipzig: Spector, 2020); and Bae Myungji, ed., *Hito Steyerl: A Sea of Data*, exh. cat. (Seoul: National Museum of Modern and Contemporary Art, 2022).

21 On speech acts, see John L. Austin, *How to Do Things with Words* (Cambridge, Mass: Harvard University Press, 1975); John Searle, *Speech Acts: An Essay in the Philosophy of Language* (Cambridge: Cambridge University Press, 1969).

ekphrasis"[22] which does not describe pre-existing images, but rather generates images by pre-describing them. Prompts are also a form of *remediation*, which turns the entire history of visual media—with their material supports, techniques, operations, as well as their protagonists, their styles, their traditions, their different historical phases, the theoretical discourses surrounding them—into a wide array of nouns, adjectives, verbs, adverbs, as well as proper names, that may be used to probe into latent space.[23]

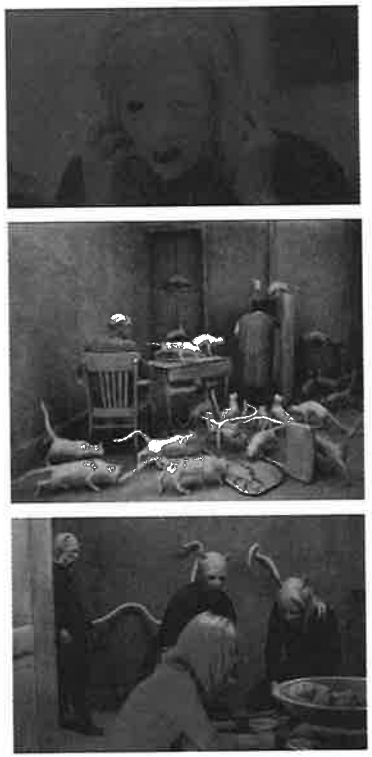
When diffusion models are instead used to perform *image-to-image* operations, as in the case of the "/blend" command in Midjourney (which allows users to combine several images in a new image), what happens is that the model analyzes and extracts key features (shapes, colors, textures, patterns, and other visual elements) from each of the input images, in order to then merge them into a new output image (see figs. 16a, 16b and 16c). In this case, the input images uploaded by the users are converted into vectors in the model's latent space, with the various numbers composing the vectors representing key features of the images. The process activated by the "/blend" command, then, interpolates between the latent vectors of the input images: it finds intermediate data points that combine different features of the input images and allow for the generation of a new image. Textual prompts may also be used to steer the blending process in certain direction rather than others.

Diffusion models can also be used to perform *image-to-text* operations: that is, to generate captions and more elaborate descriptions from still and moving images, as it happens now with ChatGPT and with the command "/describe" in Midjourney. In this case, the input image (which has been uploaded by the user or previously generated by the model) is encoded in latent space by a neural network, often a Convolutional Neural Network (CNNs) such as the ones used for machine vision. Once turned into a latent space representation, another model like CLIP, whose latent space contains data points that represent both texts and images, is used to compare the latent vector of the image with a database of latent vectors corresponding to textual descriptions. The model finds the text vectors that are the

22 On *ekphrasis* and text-to-image models, see Hannes Bajor, "Operative *ekphrasis*: the collapse of the text/image distinction in multimodal AI," *Word & Image. A Journal of Verbal/Visual Enquiry*, 40, 2024 (2), pp. 77–90.

23 On the concept of remediation, see Jay D. Bolter, Richard Grusin, *Remediation: Understanding New Media* (Cambridge, Mass.: MIT Press, 1999).

Fig. 14: Trevor Paglen, *Shadow (Corpus: Things That Exist Negatively)*, 2017. From the series *Adversarially Evolved Hallucinations*, 2017. Dye sublimation print, 121.9×152.4 cm. Courtesy of the artist, Altman Siegel, San Francisco and the Pace Gallery, London



Figs. 16a, 16b, 16c: *Blending* (bottom) produced by the artist Gwenola Wagon, 2024, from a filmstill of *Possessor* (2020) by Brandon Cronenberg (top) and the photograph *Radioactive Cats* (1980) by Sandy Skoglund (center). Courtesy of the artist

Fig. 15: Hito Steyerl, *This Is the Future*, 2019. HD video, color, sound, 16 min. Courtesy of the artist, the Andrew Kreps Gallery, New York and Esther Schipper, Berlin/Paris/Seoul

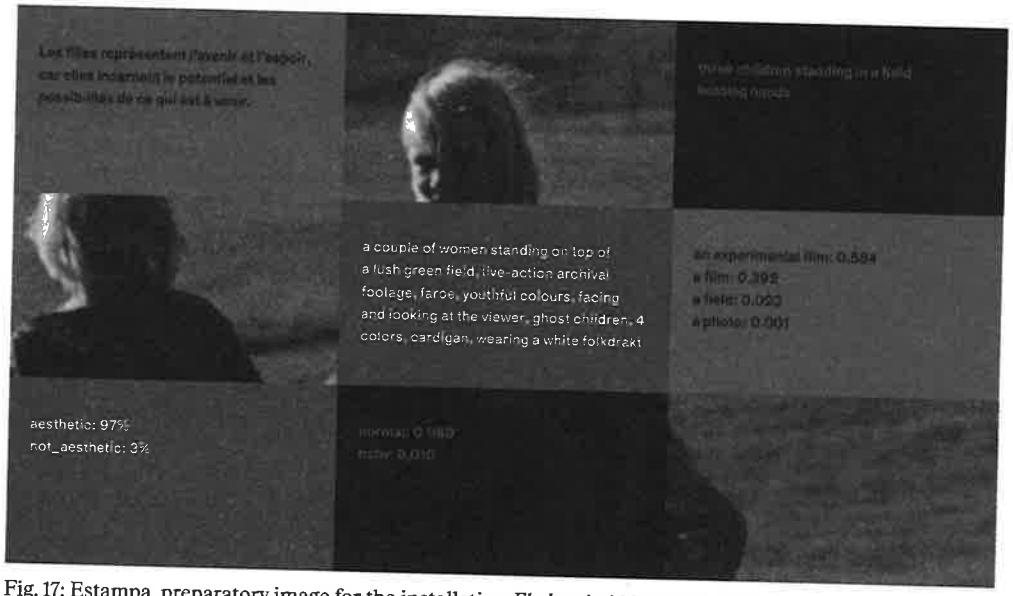


Fig. 17: Estampa, preparatory image for the installation *Ekphrasis*, 2025. Courtesy of the artists

"closest" to the image vector in latent space, and generates textual descriptions out of these vectors (see fig. 17).

Latent spaces of diffusion models can also be *modified* through different forms of fine-tuning, such as the technique called LoRA (Low-Rank Adaption): by adjusting some of the model's parameters and weights, by forcing the model to learn new patterns and create new pathways without changing the entire model's latent space, LoRAs may allow users to steer the process of image generation in certain directions rather than others, or to introduce in the model new entities—new faces, bodies, gestures, objects, spaces, materials, textures, styles, and atmospheres—that were not present before. LoRAs, in other words, intervene directly in the ontology of latent spaces, even though modifying only certain limited areas.[24]

Finally, one can try to influence future latent spaces, by trying to determine at least in part the content of the datasets that will be used to train the future AI models. An example of this strategy can be found in a recent work by Holly Herndon and Mat Dryhurst, *Xhairymutants*, presented for the first time at the 2024 Whitney Biennial. Having analyzed the way in which Herndon's image is embedded in the latent spaces of the most common current AI models, and after noticing that such compressed latent-space representation seems to focus on Herndon's distinctive red hair and blunt-cut side bangs, the artists decided to train a new text-to-image model that allows museum visitors to generate large quantities of images that amplify this cliché while at the same time themselves introducing a wide array of variations. The thousands of resulting images, stored in a source as trusted as the website of the Whitney Museum, will probably become part of the datasets used for the training of future text-to-image models and thus will influence future latent-space representations of Herndon. In this way, the artists raise the question of the limits of self-determination in relation to generative AI models, and advocate for the need to regain agency over the ways in which we are represented in latent spaces.

24 On the use of LoRAs and the idea of a "latent specificity" in AI imaging, see the essay by Noam M. Elcott and Tim Trombley in this volume.

VI. BOUNDARIES AND LIMITATIONS

As users of programs such as Stable Diffusion, DALL-E, and Midjourney know very well, when working with widely available AI models one faces all sorts of boundaries and limitations.

These depend on the AI model that is used to produce the latent space or to interact with it, and on the nature of the datasets that were used to train the model. As the researchers of the project Knowing Machines have shown, the criteria behind the selection of the connected words and images that were included in the LAION-5B dataset—used to train Stable Diffusion, the only diffusion model that was released as open source—played a key role in determining the content and the formal features of the images that the model could generate.[25] Kate Crawford adds: "All the content in latent space comes from training data, so that data becomes the *Weltanschauung* of the model: it sets the parameters of the possible."[26]

The structure of the largest latent spaces currently available in the field of generative AI is largely dependent on the commercial logics that were behind the assembling of the vast datasets used to train them, such as LAION-5B. By privileging certain sources rather than others, by focusing on the ALT tags used by commercial websites to describe images of products and make them easily accessible through search engines, by including massive quantities of generic stock photographs, training sets such as LAION-5B are "far from a neutral or representative collection of all human visual and written culture."[27] On the contrary, they are heavily rooted in specific cultural contexts and specific economic strategies.

The commercially available latent spaces are also the objects of all forms of policing and control. By tweaking the parameters or "weights" of the models, by introducing "hidden prompts" that are added to the ones provided by the users, or by establishing lists of "banned prompts" that cannot be used, those who own and control the models can heavily condition their outputs.[28] They can make it easier or more likely to generate certain images rather than others, they can facilitate the emergence and the circulation of

25 See Christo Buschek and Jer Thorp, "Models All The Way Down," in https://knowingmachines.org/models-all-the-way.

26 Kate Crawford, response to the "Questionnaire on Art and Machine Learning," in *October*, 189, Summer 2024, p. 22.

27 On this see Fabian Offert, Thao Phan A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space," *Journal of Digital Social Research*, forthcoming). Preprint: https://arxiv.org/abs/2211.06323

28 Ibid.

certain features and styles, and they can make it simply impossible to visualize certain areas of latent spaces.

Controlling latent spaces, in other words, is a way of controlling the meaning and the agency that images may have within a specific cultural context, as well as the relations between images and words, between the visible and the sayable. It means controlling the possibilities of visualization, the lines that separate what can from what cannot be seen. It gives the possibility of imposing dominant visual styles, making it difficult for users to avoid them.

Countering the tendencies inherent in latent spaces is not easy, because latent spaces are highly abstract and non-transparent: as we have seen, they are spaces of which there is no complete cartography. Those who explore them are forced to advance blindly, through trials and errors, endless variations, adjustments, and serendipitous discoveries.

What they explore in this way, are vector spaces which are like a sort of "vector imaginary": *vast repositories of possible images*.[29] Each data point within them can potentially be visualized as a still image, and each interpolation between data points may be visualized as a sequence of moving images. Only some of these images, though, can be visualized. While being extraordinarily large and vast enough to be considered nearly infinite, the number of images that can be generated out of a latent space is not infinite.

Latent spaces are therefore not limitless. On the contrary, they are full of boundaries, blind spots, "no-go" areas, as well as clichés, stereotypes, and default styles. Still, they enable a vast spectrum of operations, many of which, as we will now see, have to do with *the processing of the past*.

29 The term "vector imaginary" is used by Leonardo Impett and Fabian Offert in their essay "There Is a Digital Art History," *Visual Resources*, 38 (2) (2022), pp. 186–209. In his essay in this volume, Fabian Offert describes latent space as "an entirely virtual space, which does not contain any data. Instead, it is the space of all potential transformations of all potential images when processed by the model."

VII. LATENT SPACES AS META-ARCHIVES: POSSIBLE PASTS, COUNTERFACTUAL HISTORIES, FICTIONAL MEMORIES.

What we find in the training sets that are represented in the latent spaces of the current generative AI models are vast swaths of internet contents, and since over thirty years of internet have turned it into a massive, disorderly archive in which all kinds of cultural objects are uploaded, positioned, and preserved, latent spaces have now become a sort of

meta-archives: a key factor in the processes through which the past is recorded, interpreted and transformed.

Despite their abstract, mathematical nature, latent spaces have some elements in common with the institutions and the apparatuses that have traditionally played a crucial role in the storing, ordering, processing, and accessing of past cultural objects: we can think of libraries, archives, and collections, together with their organizing systems, such as catalogs, indexes, and card files. Interpreted in this sense, latent spaces are part of the long history of cultural techniques of encoding, organizing, and mapping, even though in the case of latent spaces these techniques operate in an abstract space that is not entirely perceivable by humans.

Their role in the processing of the past is complex and multifaceted.

When they are used for purposes of so-called "cultural analytics"—the computational analysis of cultural data—latent spaces may be used to produce forms of data visualization that reveal analogies and differences, proximities and distances between vast quantities of cultural objects that could never be apprehended in such synthetic way (for example, hundreds or thousands of artworks from a given art historical corpus, or the billions of images that are shared every day on social media platforms). The intrinsic limitation of such a process being, of course, the fact that the variety and the complexity of these cultural objects need to be reduced to data points in two- or three-dimensional spaces that can be visualized and navigated.[30]

When instead they are used for generative purposes, latent spaces have a crucial difference with libraries, archives, and collections: they are not meant for *preservation* and *retrieval*, but rather for *transformation*. They do not allow users to find the original cultural objects that were stored and positioned in them, nor exact copies of these objects. What they allow, instead, is to generate new cultural objects or endless variations of existing ones. The past that is disassembled and stored in latent spaces in the forms of vectors, in other words, is a past that is not fixed and crystallized, but rather unstable, metamorphic, subject to a wide variety of transformations[31].

This is why more and more artists are using generative AI models to visualize missing images, possible pasts, counterfactual histories, imaginary archaeologies, fictional memories, responding to a new kind of "archival impulse."[32] Searching through latent spaces for images of a past that *could have been*,[33] they treat latent spaces as a kind of "speculative archive"[34]: a vast, plastic archive in which massive quantities of cultural objects stemming from different temporal layers have been reduced to a timeless matrix of vectors and out of which new, possible, temporally ambiguous cultural objects may emerge. In the attempt to describe the peculiar status of such AI-generated images, German filmmaker Alexander Kluge uses the expression "*Konjunktiv der Bilder*," "visual subjunctive" or "subjunctive of images": "the grammatical mood of our sense of what is possible," the specific modality of images used to visualize "hypotheticals and heterotopias."[35] (See fig. 18)

Examples of this approach are presented throughout the exhibition *The World through AI*. We find them in the 'possible' cave paintings displayed in Justine Emard's *Hyperphantasia* (2022), a work that explores the latent space of a GAN model trained with a dataset of photographic images documenting the paleolithic cave of Chauvet-Pont-d'Arc, and in the "reverse archaeology" of Egor Kraft's *Content Aware Studies I* (2018–present), which use generative AI models to try to reconstruct missing fragments of statues from classical antiquity, or to visualize statues and low-reliefs that *could have been* created but never were. We find them in the GAN-generated, "hypothetic portraits" of Julien Prévieux's *Les Inconnus connus inconnus* (2018), in the speculative, decolonial reappropriations of Mesopotamian objects that ended in Western museums in Nora Al-Badri's *Babylonian Vision* (2020), and in the "missing images" of forgotten, anonymous workers in the archaeological excavations in Cyprus that Alexia Achilleos and Theopisti Stylianou-Lambert have generated for their *The Archive of Unnamed Workers* (2020). We find them, finally, in the AI-generated images with which Érik Bullot speculates on the unfinished film projects by Abel Gance or the utopian visions of a "*cinéma vivant*" by Saint-Pol-Roux (2023), in the AI-processed family photographs and archival images of Gwenola Wagon's dystopian *Chronicles of the Dark Sun* (2023) and in the fictional autobiographies of Grégory Chatonsky's *The Fourth Memory* (2024): a work in which the artist has embedded in latent space multiple

30 On *cultural analytics*, see Lev Manovich, *Cultural Analytics* (Cambridge, Mass.: MIT Press, 2020). See also Lev Manovich and Emanuele Arielli, *Artificial Aesthetics: Generative AI, Art and Visual Media*, available online at https://manovich.net/index.php/projects/artificial-aesthetics.

31 See for example the curious merging of Nicéphore Nièpce's *View from the Window at Le Gras* (1826 or 1827) and Daguerre's *Boulevard du Temple* (1839) in the AI-generated image commented by Joanna Zylinska in her essay in this volume. See also Joanna Zylinska, *AI Art: Machine Visions and Warped Dreams* (London: Open Humanities Press, 2020) and *Nonhuman Photography* (Cambridge: MIT Press, 2017).

32 On generative AI, missing images, archaeology and synthetic history, see Ada Ackerman's text in this volume. The idea of an "archival impulse" in contemporary art refers to Hal Foster's essay "Archival Impulse," *October*, 110, fall 2004, pp. 3–22. Susanne Østby-Sæther spoke of an "AI-rchival impulse" in a paper presented at a conference at the Centre universitaire de Norvège in Paris in October 2024.

33 According to Roland Meyer, "AI image synthesis is a backward prediction: it makes plausible guesses on what could have been" and "promises to fill the imaginary gaps in the virtual archive of past images": Roland Meyer, "Platform Realism: AI Image Synthesis and the Rise of Generic Visual Content," in *Photographie et algorithmes*, ed. by Estelle Blaschke, Max Bonhomme, Christian Joschke, Antonio Somaini, *Transbordeur*, 9 (2025), forthcoming.

34 Érik Bullot, *Cinéma vivant* (Paris: Macula, 2024). According to Bullot, the temporal status of AI-generated images is that of "a past conditional that allows us to imagine past utopias, renewing our relationship with archives. The archive has become speculative."

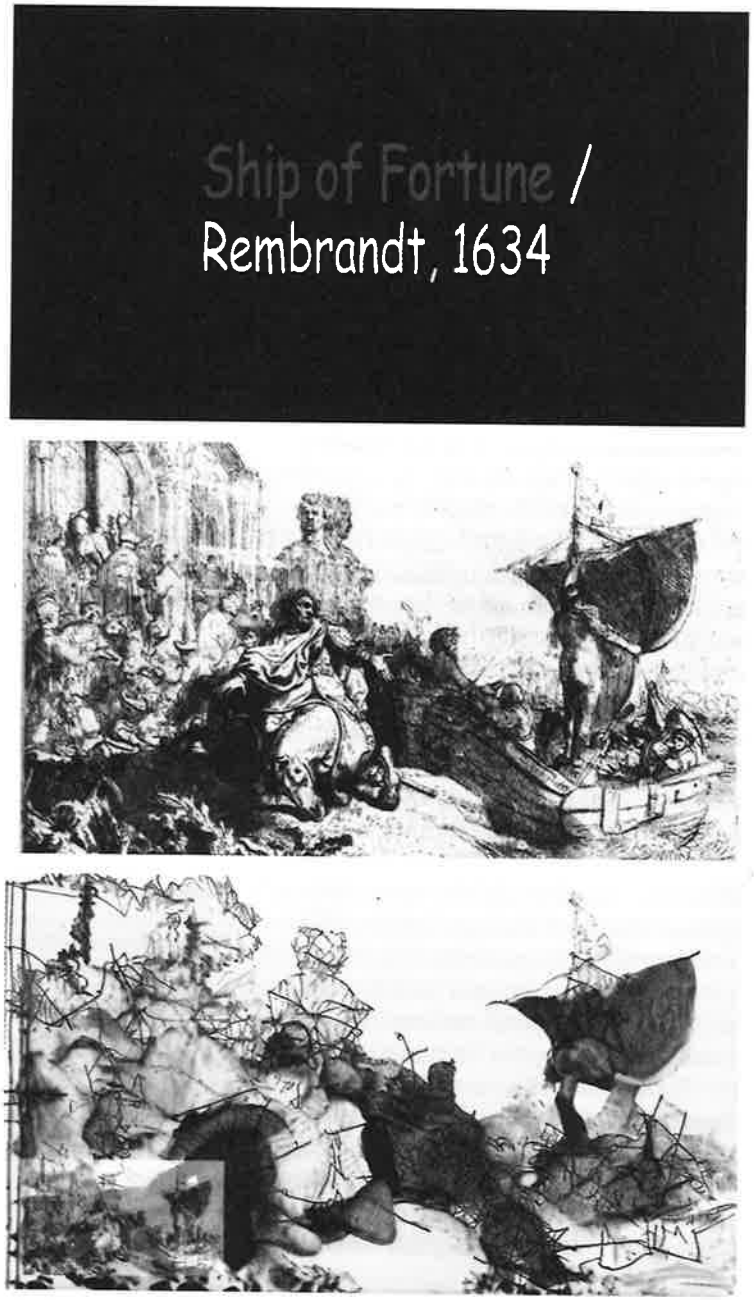


Fig. 18: Alexander Kluge, *Ship of Fortune/Rembrandt, 1634* (details), 2024. Video, color, sound, 5 min. 2 sec. Courtesy of the artist and Kairos-Film

traces of his own existence, in order to generate other, possible versions of his own past.

Chatonsky's own writings are a key contribution to a theory of latent spaces, which tackles specifically their role in the processing of the hypertrophic, disorderly mass of cultural traces that are accumulated on the web.[36] The title of the new installation he presents in *The World through AI—The Fourth Memory*—highlights the fact that what is stored and processed in the latent spaces of generative AI models are digital objects (images, texts, sounds, voices) which are already the result of technical process of recording and memorializing: more precisely, they are part of what Bernard Stiegler's called "tertiary retention," a memory that has been exteriorized and inscribed in material forms through different recording media (gramophone disks, analog photographs and films, video tapes, digital files, etc.).[37] According to Chatonsky, by encoding the traces of this "third memory" into vectors and by storing them in latent spaces in which they can be processed and transformed, AI models act as a sort of "fourth memory," a "meta-memory" that contains, in the form of vectors in latent space, a whole spectrum of possible pasts and counterfactual histories such as the ones that we encounter in his installation.[38]

"Latent space architecture and pathmaking": the phrasing used by Hito Steyerl in the credit sequence of *Animal Spirits* encapsulates one of the key questions that artists and the public at large are facing in dealing with present and future latent spaces. As the exhibition *The World through AI* aims to show, exploring them, modifying them, generating them means today operating on those abstract, invisible, active algorithmic entities that are playing an increasingly important role in the fabric and the dynamics of contemporary culture.

35 Alexander Kluge, *Der Konjunktiv der Bilder. Meine virtuelle Kamera (K.I.)*; English translation *The Dragonfly Eyes: My Virtual Camera (A.I.)* (Leipzig: Spector Book, 2024), pp. 33 and 396.

36 See, for example, Grégory Chatonsky, "The Imagination OF the Latent Space": http://chatonsky.net/de-of/.

37 See Bernard Stiegler, *La technique et le temps* (1994) (Paris: Fayard, 2018).

38 Grégory Chatonsky and Yves Citton, "La quatrième mémoire," *Multitudes*, 96 (2024), pp. 186–189.

ISBN: 978-2-36568-108-7

THE WORLD THROUGH AI

● JEU DE PAUME JBE BOOKS

EXPLORING LATENT SPACES