Grégory Chatonsky.
Complétion 1.0, 2021.
Still from video installation.
© and courtesy the artist.
Commissioned by the Centre national des arts plastiques (CNAP).

https://doi.org/10.1162/grey_a_00383

Algorithmic Images: Artificial Intelligence and Visual Culture

ANTONIO SOMAINI

Since the beginning of the 2010s, the vast field of digital images has been increasingly impacted by different kinds of deep-learning algorithms—one of the forms of so-called artificial intelligence (AI)—that profoundly transform the ways in which images are captured, generated, modified, and seen. Computer vision and its application in machine-vision technologies are now capable of detecting and recognizing objects, places, bodies, and faces within the billions of images circulating across the internet, even when they do not appear on screens and are not visible to human eyes. Today's smartphones are equipped with different AI features that automatically modify and correct captured images on the basis of recurrent patterns and preferences that have been statistically detected across social-media platforms. Deep-learning models can automatically generate images from text or vice versa. Massive datasets made of billions of images, texts, and text-image pairs scraped from the internet are used to train these models, thus influencing their visual and textual output, gradually turning our culture into a huge feedback loop in which what has already been uploaded to the internet conditions future AI-generated content.

Each one of these phenomena has profound implications all across the field of contemporary visual culture. Machine vision introduces a new form of automated visual perception that decenters the human gaze and reorganizes the field of the visible, redrawing the lines that separate what can from what cannot be seen. AI features embedded in smartphones blur the distinction between image capture and image processing, and promote new forms of standardization. Deep-learning models introduce new ways of connecting images to other images, images to texts, texts to images, and, ultimately, people to people through the mediation of images and texts. Techniques of image generation, processing, and editing in fields such as illustration, graphic design, photography, video, and film are being quickly transformed and in some cases entirely replaced by new forms of so-called prompt engineering. All the research fields that deal, from different perspectives, with the study of images and visual media are suddenly faced with the possibilities opened up by algorithms capable of analyzing and classifying images within vast image databases, but also with the questions raised by the limits and the biases of these algorithms, as well as by an entirely new array of AI-generated images and image operations.

All this is happening at a pace that was hardly imaginable just a few years ago. Still, our feeling of surprise and even stupor will not last long. Deep-learning technologies operating on images will soon become the "new normal," and distinguishing the lines of continuity and the moments of rupture that define their position within the longer history of images and vision will be more difficult. In some cases, they will become standard digital tools (as, for example, has already happened with technologies such as QR code reading) and may no longer be considered "artificial intelligence," since the meaning of the term has been continually shifting.[1]

We are now poised on a threshold and have an opportunity to make sense of these developments. Before these transformations become invisible, before they sink to the deeper layers of our digital infrastructure, before they are replaced by even further transformations, we can stop for a moment and try to understand what is happening.

In what follows, I describe the main kinds of deep-learning algorithms that are behind the current transformations and explain in simple terms how they function and how they have been applied within visual culture at large and within a series of contemporary artistic practices that may be particularly relevant in helping us navigate this new landscape. The impact of these algorithms on images is so profound that it raises a series of key aesthetic, epistemological, ontological, and political questions that need to be tackled from both theoretical and media-archaeological perspectives.[2]

We need not only to understand what *properties* an image must have if it is to be analyzed, generated, or modified by deep-learning algorithms but also to map and study the various *operations* in which images processed by such algorithms are involved.

We also need an analysis of the main deep-learning algorithms dealing with images and of the datasets used to train them. On the one hand, it is important to understand the structure of these algorithms and the different intertwinings between human and nonhuman agencies that regulate their functioning, giving them different degrees of autonomy. On the other hand, we have to study the sources, the content, and the guiding criteria of the datasets that are used to train such algorithms: both the ones used for machine-vision applications (to better understand what they can and cannot "see," their biases, and their contribution to various forms of algorithmic governmentality and discrimination) and the ones used to generate or modify images (to understand how the images and the text-image pairs contained in these datasets condition their outputs).[3]

The choice, reflected in the title of this article, of the term *algorithmic images* has a specific aim: it foregrounds the fact that, within the current media landscape, the status, the agencies, and the affordances of images across contemporary visual culture—the ways in which images are captured, modified, circulated, and seen within different social and cultural contexts—are intrinsically connected to their being processed by deep-learning algorithms trained with large datasets.[4]

The study of such "algorithmic images" leads me both to introduce in a theory of images and visual culture concepts stemming from the field of machine learning, and to reconsider, from the perspective of the current transformations, a series of key concepts and issues in fields such art history, visual-culture studies, photography, film, and media theory.

Concepts from the field of machine learning that need to be introduced to a theory of images and visual culture include those of the dataset, the training set, embedding, conditioning, alignment, hallucination, and interpolation. A particularly important one is that of "latent space": the abstract, multidimensional space in which deep-learning algorithms turn digital objects (e.g., the vast quantities of images and texts that have been uploaded to the internet) into latent representations so that they can be processed and used to generate new digital objects (e.g., new images and new texts).[5] Latent representations are made of vectors; that is, long lists of numbers that define the coordinates of the digital objects encoded and embedded in latent space and their relations of distance and proximity within it, just as the three coordinates *x*, *y*, and *z* define the position of a physical object in three-dimensional space and its relations to other physical objects. In the coming years, understanding the dynamics of a culture more and more innervated by deep-learning algorithms and examining the ways in which it processes vast amounts of visual and textual traces that the past has left on the internet will become impossible without recognizing the key role of this abstract, unintuitive, multidimensional space within which preexisting images and texts are embedded, positioned, and processed, and out of which a wide spectrum of new images and new texts may emerge.

Among the concepts and issues stemming from disciplines such as art history, visual-culture studies, photography, film, and media theory that may be reconsidered from the perspective of the impact of deep-learning algorithms on images, we find not only the concepts of image and vision, but also concepts such as resemblance, imitation, original versus copy, index versus indexing, referent, objectivity, style, abstraction versus figuration, realism and photorealism, as well as the question of the nature of the artist's agency, authorship, and creativity in the context of complex interactions with algorithms that have various degrees of autonomy.

Deep-learning algorithms are also recalibrating the relations between images and words, the visible and the textual, since their functioning depends, in one way or another, on the availability of vast quantities of images that are systematically indexed, labeled, and captioned and then gathered in large datasets. At stake is therefore a future visual culture in which images and words are increasingly algorithmically connected and inseparable.

Finally, reflecting on the impact of deep-learning technologies on images can also help us understand what role AI-generated images play in making "artificial intelligence" itself *visible*: either to make it somehow more transparent and comprehensible, or to explore its image-generating potential, or to capture and divert our attention through seductive visual surfaces and different forms of "art washing" that may help disguise its most problematic implications as a technology of surveillance, prediction, discrimination, and job replacement.

Artificial Intelligence, Images, and Vision

First used in 1955 in the project proposal for the Dartmouth Summer Research Project on Artificial Intelligence (1956), the term *artificial intelligence* refers today to a field that developed over the following decades as a complex network of theories, technologies, and applications, surrounded by a spectrum of discourses, narratives, and imaginaries.[6]

As researchers in the second half of the 1950s worked to develop an intelligence that was "artificial" (i.e., technically produced, nonbiological), two main approaches emerged: a "symbolic" approach and a "subsymbolic" or "connectionist" approach. They differed not only in how each understood the very idea of an "artificial intelligence" but also for a series of technical, institutional, economic, and political reasons that, over the following years, heavily conditioned the allocation of funding.[7]

The "symbolic" approach was grounded in the traditions of mathematical logic, systems engineering, and cybernetics and based on the idea that computers could reproduce some rational aspects of human thought (e.g., solving problems, making judgments, taking decisions) through symbol-processing programs (i.e., programs based on rules and performing operations on symbols and combinations of symbols).[8] Promoted by the organizers of the Dartmouth workshop (John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon), this approach dominated research and funding from the late 1950s to the mid-1990s.

In contrast, the "subsymbolic," "connectionist" approach was grounded in a tradition that combined mathematics, statistics, cognitive psychology, and neuroscience and was based on the idea that "artificial intelligence" could be conceived as a form of "machine learning"; that is, the use of computer algorithms that are capable, through inductive sequences of trials and errors, of learning how to recognize patterns within given datasets so as to make predictions on newly given data. The structure of these algorithms imitates, in a streamlined and schematic way, the connections between biological neurons, and for this reason they are called "artificial neural networks."

Today, among the multiple approaches adopted by technologies labeled as "artificial intelligence," the dominant ones tend to be connectionist. The algorithms they employ are "deep," multilayered, artificial neural networks, and the form of "machine learning" they activate is therefore called "deep learning." Fueled by technologies that use large quantities of planetary resources, trained through various kinds of human labor and through large datasets whose structure, content, and guiding principles raise a whole series of ethical and political issues, these deep-learning algorithms are deployed to introduce various forms of *automation* and *prediction* in areas where large amounts of data need to be processed.[9]

The fields of application for deep-learning algorithms are currently as vast, diverse, and complex as the very fabric of our culture. They include auditory perception, through forms of machine listening that automatically recognize voices, sounds, and noises; "natural" (i.e., human, instead of machine code–based) language processing, through so-called large language models (LLMs) capable not only of synthesizing and translating existing texts but also of generating new texts starting from textual prompts; musical composition, with systems capable of completing unfinished works or of generating new works "in the style of" a given musician or tradition; strategic game systems, such as the ones developed to play chess, Go, or Atari videogames; advanced web search engines; recommendation systems; targeted online advertising; email spam filters; so-called virtual assistants (such as Siri and Alexa); robotics and driverless vehicle guidance (for cars, trucks, drones); the management of energy storage and consumption; automated tools for analysis, prediction, and decision-making in fields such as industrial production, supply chain logistics, finance, credit rating, medical diagnosis, pharmaceutical research, meteorology, politics, and, of course, military operations.[10]

Among the fields that are today most impacted by the use of technologies labeled as "artificial intelligence," we find the field of *visual culture*, a term I use here to refer broadly to the roles that images, visual media, and visual experience play within various technical, cultural, social, and political contexts.[11]

Three major phenomena deserve our closest attention. We can list them in the chronological order of their appearance over the last ten years.

1. Beginning around 2010 (though based on research initiated during the second half of the 1950s), deep-learning algorithms such as convolutional neural networks (CNNs) began to be systematically used to implement systems of "machine vision" that can detect, analyze, and classify entities (e.g., objects, places, bodies, faces, gestures, expressions, and actions) represented in images. Technologies of machine vision can today be applied to the billions of digital images that are accessible through the internet and to the even larger number of images that are stored within our digital devices or in offline archives, even when these images do not appear on screens and therefore are not visible to human eyes.
2. During the mid-2010s, other deep-learning algorithms, such as DeepDream and generative adversarial networks (GANs) were introduced. Their main function is not to analyze and classify images but to *modify* existing images through a series of operations or to *generate* entirely new images that can be photorealistic, hybrid, or completely abstract.
3. In early 2022, new deep-learning models that are part of the wider field of so-called generative AI became broadly available. These are capable not only of generating both *still* and *moving* images from texts (i.e., from so-called prompts, as with such text-to-image models as DALL-E 2, Stable Diffusion, and Midjourney, and with various text-to-video applications), but also of *generating texts from images* (e.g., with models that extend the task of image classification to generate captions, descriptions, and even short stories starting from a given image, answering questions about it, or performing tasks that further develop it).

These three phenomena are currently impacting not only visual culture at large but also a whole spectrum of artistic practices, some of which are particularly relevant since they tackle, using a variety of strategies, the crucial epistemological and political questions raised by these technologies.

In order to be processed by deep-learning algorithms, images must possess two key properties. The first (which has to do with the digitizing of images in general) is that images need to be *rasterized*; that is, reduced to an orthogonal grid of pixels, each with its own coordinates (row plus column) and color values.[12] Reducing an image to an orthogonal grid in which each pixel is associated with two coordinates and a set of numerical values is the fundamental condition of possibility for an image to be processed by deep-learning algorithms (this holds for many other digital image-processing applications too). The grid acts here, once more, as a "cultural technique" that allows for the *localization*, the *addressing*, and the *activation* of single elements in the image.[13] All images—whatever their original material supports, size, their format, and original contexts of production and reception—must be reduced to the technical a priori of the grid so they can be turned into data that can be processed.

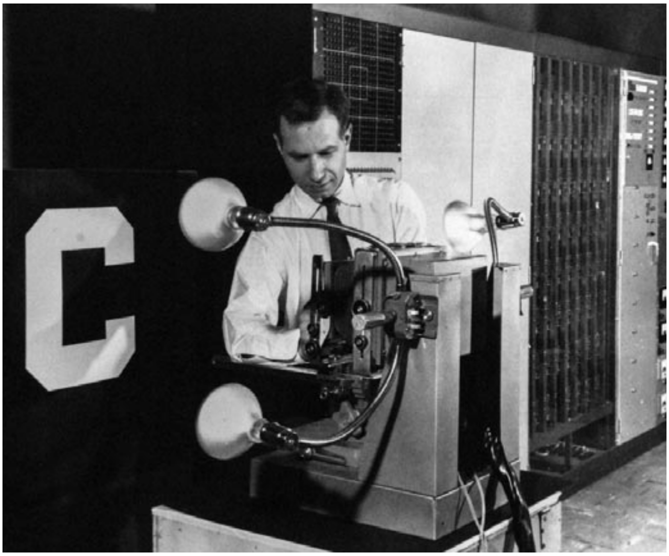
The second key property images must possess to be processed by deep-learning algorithms is that they need to be *indexed*; that is, images must be systematically linked to words (e.g., a label or a caption) that allow them to be organized in categories. The existence of vast taxonomies connecting images with words is therefore another fundamental a priori for the functioning of all these algorithms, be they CNNs used for machine-vision applications, the GANs used to modify existing images or generate new ones, or the recent text-to-image and image-to-text models.

Detecting, Recognizing, Classifying: Machine Vision and Operational Images

Technologies of "artificial intelligence" have been dealing with tasks related to images and vision since the beginning.

In 1957, a year after the Dartmouth workshop at which the term *artificial intelligence* was first used, the psychologist Frank Rosenblatt, working at the Cornell Aeronautical Laboratory, developed the Perceptron, a machine for automated image recognition.[14] A first example of a connectionist, "subsymbolic" program of AI, inspired by the studies of Warren McCulloch and Walter Pitts on "artificial neurons" (mathematical functions conceived as a model of biological neurons), the Perceptron was a single-layer artificial neural network whose task was learning how to recognize two-dimensional alphabetical characters after having captured them through an orthogonal grid of sensors composed of four hundred photocells.[15]

Fifty years after the development of the Perceptron, following several cycles of enthusiasm and disappointment, investment hype, and funding cutbacks ("AI winters") from corporate, public, and military sources (e.g., the Defense Advanced Research Projects Agency), funding and investment in AI sharply increased at the beginning of the 2010s.[16] This new "AI spring"—motivated, like the preceding ones, by a complex series of technological, institutional, and economic developments—allowed machine-vision applications to enter a new phase, thanks to the convergence of four different factors.



Frank Rosenblatt working on the camera system of the Mark 1 Perceptron, 1960.

The first factor was the development, already imagined by Rosenblatt in 1962 but materialized only during the 1970s and 1980s with the Cognitron and Neocognitron, of new, “deep” (i.e., multilayered) artificial neural networks whose performance was then highly improved by a key feature of machine learning, an algorithm for the “back-propagation of error.”[17] This algorithm makes it possible, once an error (e.g., the misidentification of a digit) has been identified in the output of a deep neural network, to trace the factors that caused the error, layer by layer, connection by connection, in order to correct it by modifying the parameters (the “weights”) that regulate the circulation of the inputs and outputs across the various connections.

The second key factor was the possibility of training such neural networks on vast quantities of images downloaded from the internet, systematically indexed and then organized in large image datasets. Some of these datasets, such as ImageNet, were, during the first half of the 2010s, the standard benchmarks for research in machine learning applied to images, before being replaced by even larger datasets such as LAION-5B.

The third key factor was the availability of new forms of distributed microlabor, whether performed through CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) or by people tagging and commenting on images circulating on social media or accessible through online labor platforms such as the crowdsourcing marketplace Amazon Mechanical Turk.[18] This so-called click-work, as it is now well known, raises serious ethical and political issues, because it involves large quantities of underpaid, underprotected, and even free labor. Beginning with the mid-2000s, it provided one of the key steps in the training of machine-vision technologies: the labeling of the images gathered in the datasets.[19]

The final key factor that allowed machine-vision technologies to enter a new phase was the introduction, during the early 1990s, of a new generation of powerful graphics processing units (GPUs): these are specialized electronic circuits, which were initially designed to accelerate operations of computer graphics and image processing in the domain of real-time digital animation in videogames but ended up being one of the key ingredients of the algorithmic processing of images in general.

Among the deep-learning algorithms that contributed to the rapid development of machine-vision technologies at the beginning of the 2010s, a key role was played by CNNs. Understanding, even if only schematically, how they operate in detecting, recognizing, and classifying the entities represented in images is crucial if we are to better grasp the possibilities and limitations of machine vision.[20]

Strictly speaking, the goal of a CNN used for a machine-vision task is to recognize a particular group of pixels in a digital image as representing a

A convolutional neural network (CNN). Diagram by the author.

particular object category, such as “human face,” “cat,” “tree,” “table,” “church,” and so on. Object categories can have multiple levels of generality (“cat,” “Siamese cat,” “female Siamese cat”) and can also refer to individual objects and individual people (“Notre-Dame Cathedral,” “Barack Obama”).

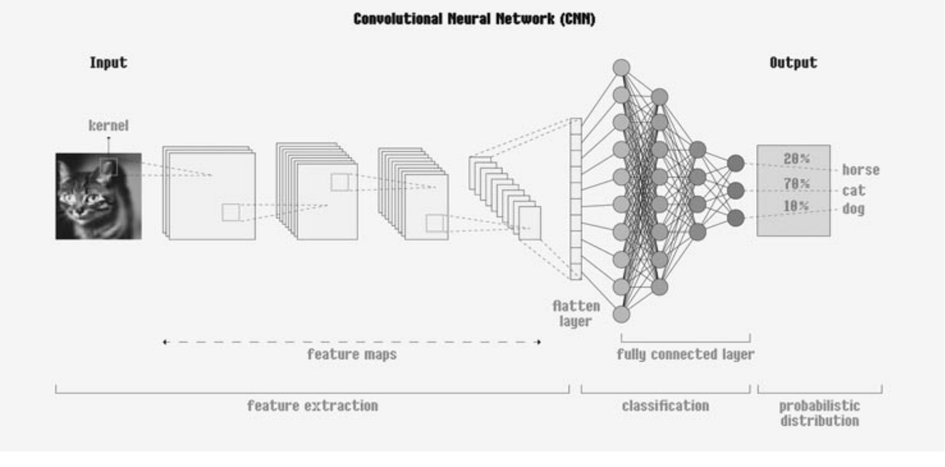
The functioning of a fully trained CNN applied to a task of machine vision unfolds in two parts: *feature extraction* and *classification*.

At the beginning of the *feature extraction* process, an initial input—a digital image (say, the image of a cat)—is organized as an orthogonal grid of pixels, each with its own coordinates and numerical values. The size of the image (i.e., the number of pixels in the orthogonal grid) must be equal to the size of the first layer of the artificial neural network (the number of inputs it can receive).

Once an image is provided, the different layers of the CNN begin to analyze it. Each feature of the given image is targeted by a specific artificial neuron within each layer of the network. The neurons of the initial (lower) layers have the task of detecting simpler features (e.g., lines, curves, corners, or edges—that is, the boundaries between two contrasting image regions). If these simple features are detected, if a certain “activation value” (expressed in numbers) is reached (by multiplying the value of each neuron of a given layer by its weight), the neurons are activated and transmit their input to the further (higher) layers. These, in turn, are in charge of detecting more and more complex features (e.g., shapes, surfaces, volumes, textures, and then entire objects, individual faces, and so on). The flow of activations between the different layers of a CNN—the result of sequences of mathematical operations called “convolutions”—proceeds both through feed-forward activations (from lower to higher layers) and through feed-backward activations (from higher to lower layers).

The second part of the process, *classification*, is performed by another kind of neural network, called a “classification module.” Based on the inputs received from the higher layers of the CNN, it produces as output a classification based on a certain percentage of confidence.

To perform the two operations (feature extraction and classification), CNNs need to be *trained* to recognize either a single object category or multiple object categories, including eventually texts, code, and mathematical formulas that might appear in an image. If, for example, “cat,” “horse,” and “dog” are among the categories the network has been trained to recognize, the (probabilistic) output from the analysis of an image of a cat might be



"cat 70%, horse 20%, dog 10%." In machine-vision applications, such training—which goes through many phases or "epochs"—is either *supervised* or *unsupervised*, depending on the role human labor plays in the various stages of the process.[21]

Understanding the inner workings of CNNs is extremely difficult if not impossible, since these deep, multilayered neural networks may have millions of "neurons" and an even larger number of interconnections between them. The forward and backward flow of information between the layers may unfold through thousands of iterations, making them even more opaque and inaccessible. CNNs are thus a typical example of "black box" deep-learning models that cannot be fully interpreted.[22]

Instead, the structure and content of the datasets used for their training may be analyzed, although this is possible only through samples, since the number of images contained in the datasets may be on the order of millions or, in the most recent datasets, even billions.

Training sets play an essential role in defining the "epistemic space" of CNNs: the series of entities they can detect and classify within images and the words the network uses for such classification. The selection, labeling, and taxonomic grouping of the large quantities of discrete images contained in the training sets play a crucial normative role in distinguishing between what can be *seen* and *named* by a CNN and what remains *unseen* and *unnamed*. In many cases, what remains unseen and unnamed may be within the image itself. A neural network, for example, may be capable of recognizing an apple in an image but not the plate it sits on nor the table on which the plate rests.

To better understand this articulation between detecting and classifying, seeing and naming, consider ImageNet, the vast image dataset that allowed CNNs to reach a near-complete dominance in the field of machine vision in the early 2010s, replacing earlier, smaller datasets that were based on a more limited number of object categories.[23]

ImageNet, created at Stanford University by a team of computer scientists led by Fei-Fei Li, was presented for the first time in 2009.[24] Its explicit aim was to "map out the entire world of objects."[25] The statement highlights a search for cartographic totality that is not only highly questionable—it implies that one can establish a complete list of all objects, that they are all visible, and that they can all be represented by an image—but is further undermined when one takes a closer look at the sources, structure, and multiple embedded biases of the dataset.

As Kate Crawford and Trevor Paglen show in their "Excavating AI: The Politics of Images in Machine Learning Training Sets," ImageNet is an emblematic example of the fact that "every layer of a given training set's architecture is infused with politics."[26] With its fourteen million images downloaded from the internet without prior clearance from their authors, then labeled by tens of thousands of clickworkers hired through Amazon Mechanical Turk and asked to use an interface that invited them to select "photos only, no painting, no drawing," and finally organized in twenty-one thousand categories and subcategories based on nouns of the English language as they appear in the WordNet hierarchy, ImageNet is far from being an objective mapping of "the entire world of objects."[27] Instead, it is the result of multiple layers of technical determinations and human evaluations, judgments, decisions, and biases.[28] The conclusion Crawford and Paglen draw from their analysis is that "understanding the politics within AI systems matters more than ever, as they are quickly moving into the architecture of social institutions" which are involved in various activities of surveillance, monitoring, control, selection, and prediction.[29]

Since the early 2010s, machine-vision technologies have been increasingly applied to the immense field of machine-readable images, a field whose dimensions may be imagined only if we understand that, potentially, *any digital image*—whether accessible through the internet or stored in our devices, whether produced through some kind of lens-based optical recording or entirely computer-generated or a mix of the two, as is often the case—may be analyzed by machine-vision technologies. All the main smartphone producers have equipped their devices with cameras and image-processing technologies that turn every photograph we take into a machine-readable image and use machine-vision applications to either search through the photographs and videos one has taken or to search the web starting from a given image. Social-media platforms use machine-vision systems to extract data from the images and videos uploaded by users, while private companies (e.g., the controversial Clearview AI) offer state agencies and private clients machine-vision and face-recognition systems capable of analyzing the immense quantity of photographic images that can be found on the internet and that continue to be uploaded every day, raising all sorts of ethical and political issues and highlighting the need for a broader legal framework that for the moment is largely missing.[30]

Considered together, machine-vision systems are turning the contemporary digital "iconosphere" into a vast field for data mining and analytics in which objects, places, bodies, faces, expressions, gestures, and actions—as well as voices and sounds, through technologies of *machine listening*—may be detected, analyzed, labeled, classified, stored, retrieved, and processed as data that can be quickly accessed and activated for a wide variety of purposes and operations: from surveillance to policing, from marketing to advertising, from the monitoring of industrial processes of manufacturing and distribution to the functioning of driverless vehicles (cars, drones, and robots), from medical diagnostics performed through the automated analysis of medical imaging all the way up to various military applications.

Machine-vision applications have also been used, for a few years now, in various fields in the humanities: from art history to the history of photography, film, and audiovisual media.[31] Scanning vast corpuses of still and moving images, they may search not only for bodies and objects but also for colors, textures, configurations of light and shadow, degrees of sharpness and blurredness, and then movements, postures, gestures, actions, facial expressions, all the way up to recurring motifs and styles.[32] In analyzing moving images, they may also detect different kinds of shots (e.g., close-up, medium, and long shots), camera movements, and editing techniques. The traditional skills of art connoisseurship are also being transformed by new forms of so-called AI art authentication that may help attribute a painting, reconstruct its date of production, and detect copies and duplicates.[33]

Considered from the point of view of an epistemology of art history and image analysis, these applications of machine vision, with their stated goal of "provid[ing] additional objectivity derived from quantitative and computational processes," raise a whole series of questions.[34]

To be studied through machine-vision technologies, artworks must undergo an ontological transformation. Their concrete materiality, their techniques, and their dimensions are all reduced to digital images, i.e., to orthogonal grids of pixels, each with a unique set of coordinates and values. Two- or three-dimensional visual forms are segmented and turned into data (i.e., numerical values that can be computed by algorithms). The highly complex notion of "style"—whether the style of a single artist, of an artistic movement, or of a historical period—is untethered from its sociohistorical determinations and from questions of technique, materiality, and facture so it can be reduced to a series of recurring pixel patterns that may become the object of some form of computational pattern recognition. More broadly, what used to be an analysis developed by a human observer—even though often through the aid of various technical tools, such as photographic reproductions, double slide projections, close-ups, and, in the case of moving images, slow motion and freeze-frame—turns into a process of formal analysis in which human vision is decentered and repositioned and in which the focus is on algorithmic operations that tend to leave out "what remains irreducible to quantification and computation, and what cannot be explained in quantitative terms."[35]

From the point of view of a theory and history of images and vision, a crucial turning point in the recent development of machine-vision technologies is the fact that they increasingly operate in circuits that connect machines directly to other machines and trigger various kinds of operations without necessarily generating images that are visualized on screens and therefore made visible for human eyes. Machine-vision technologies may operate on images generated *by machines for other machines*—without human eyes ever being involved (as in the case of a camera that monitors a production line and activates some technical intervention if it detects a mistake)—or *by humans for other humans* (as in the case of a photograph taken and then shared on social media to trigger reactions such as likes, tags, and comments). In this second case, machine-vision technologies remain active even when the images have become invisible. Images uploaded to the internet may keep fueling algorithmic systems of data collection and analytics even after they have ceased to appear on screens (as with an image on a website with no visitors or a social media post that is no longer looked at).

With the proliferation of deep-learning technologies, the vast field of so-called instrumental and operational images undergoes a profound transformation.[36] In a series of texts and video installations of the early 2000s, Harun Farocki defined operational images as "images without a social goal, not for edification, not for reflection"; that is, images that "do not represent an object, but rather are part of an operation."[37] Installations such as *Eye/Machine I, Eye/Machine II, and Eye/Machine III* (2001–2003), *Counter Music* (2004), *Deep Play* (2007), and *Serious Games I–IV* (2009–2010) introduced the viewer to a vast array of operational images circulating across various technical, industrial, scientific, entertainment, and military contexts. Farocki meticulously gathered, reorganized, and exposed these images, which still appeared on screens. Today, as Paglen underlines in an essay written shortly after Farocki's death, operational images do not need to be visible to be active.[38] Their role in the collecting and processing of visual data needs less and less to be monitored and validated by a human viewer.

This has a direct impact on the very concepts of "image" and "vision." Can we still use the term *image* for a digital file, encoded in some image format, that is machine-readable even when it is not visible by human eyes or becomes visible on a screen as a pattern of pixels only for a small fraction of time, spending the rest of its indefinite lifespan circulating across invisible digital networks? What becomes "vision" when the human psychophysiological process of seeing is converted, in the case of machine-vision technologies, to automated operations of pattern recognition and labeling, and when the various applications of such operations may be deployed across vast quantities of digital images that no human eye could ever see in their entirety? By using the term *vision* within the concept of machine vision, are we mistakenly using a metaphoric term that should be discarded in favor of a different set of technical terms, ones specifically related to the fields of deep learning and data analysis?

Authors such as Andreas Broeckmann (with his notion of "optical calculus" as "an unthinking, mindless mechanism, a calculation based on optically derived input data, abstracted into calculable values, which can become part of computational procedures and operations"), Adrian MacKenzie and Anna Munster (who describe machine vision as a form of "platform seeing" that targets vast "image ensembles" through an "invisual perception"), and Fabian Offert and Peter Bell (who underline the unique specificity of the "deeply-non-human" "perceptual topology" of machine-vision systems) have all argued for the need to move beyond anthropocentric frameworks and terms, highlighting the radical differences between machine vision and human vision.[39]

Using the term *vision* in *machine vision*, though, has an undeniable heuristic and hermeneutic value. It invites us, for example, to position machine-vision technologies in the longer history of visual media that were used to enhance, decenter, or entirely replace human vision through technical means.[40] The use of the term *vision* also underlines the importance of trying to visualize, for human eyes, the "invisual perception" of machine-vision technologies.

Some of the most interesting attempts in this direction can be found in Paglen's videos, exhibitions, and writings (often in collaboration with Crawford).[41] In a text titled "Invisible Images (Your Pictures Are Looking at You)," Paglen describes a new landscape in which "images are made by machines for other machines, with humans rarely in the loop" and draws the following conclusion: "If we want to understand the invisible world of machine-machine visual culture we need to unlearn to see like humans. We

Trevor Paglen. *Behold These Glorious Times!* 2017. Still from single-channel color video projection, stereo, 10 min.; original score by Holly Herndon. © Trevor Paglen. Courtesy the artist; Altman Siegel, San Francisco; and Pace Gallery.

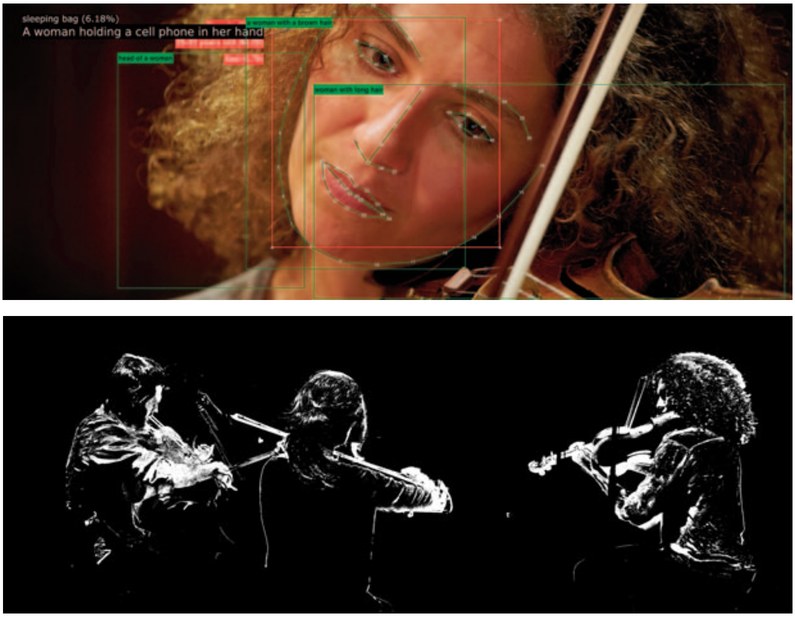
need to learn how to see a parallel universe composed of activations, keypoints, eigenfaces, feature transforms, classifiers, training sets, and the like."[42]

A video such as *Behold These Glorious Times!* (2017) brings together, through fast editing and a grid-like projection, thousands of images stemming from various machine-vision training sets, showing how, in some cases, human beings themselves had to be trained to perform in front of a camera facial expressions and bodily postures that would then be used to train algorithms.[43] Another video, titled *Image Operations. Op.10* (2017), highlights the radical difference between human sensory perception and machine vision. Faced with the performance of a string quartet, the machine-vision system, for which sound does not exist, focuses on the faces, expressions, and gestures of the musicians, trying to detect their age, their emotions, and the objects they are holding. Other images from the same video try to visualize what the CNN is "seeing"; that is, the way it begins to analyze, in a given image, simple features before gradually moving to more complex ones.[44]

Images from Images: Latent Space Visualizations

The possibility of using deep-learning algorithms to *generate* images—rather than to *analyze* and *classify* them—was highlighted in 2014 and 2015 by the introduction of two algorithms whose images began quickly to proliferate across the internet. In both cases, the new images generated by these algorithms are deeply related to the images contained in the datasets that were used to train them. For this reason, they may be considered to be *images from images*; that is, images produced through the algorithmic processing of vast quantities of other images.[45]

The first of the two algorithms, called DeepDream, was developed in 2015 by Google engineers Alexander Mordvintsev, Christopher Olah, and Mike Tyka on the basis of a deep CNN architecture called GoogLeNet (also known as Inception) that was presented for the 2014 ImageNet Large-Scale Visual Recognition Challenge.[46] The name "Inception" refers both to the 2010 science fiction film directed by Christopher Nolan—in which the


sleeping bag (6.18%)
A woman holding a cell phone in her hand
a woman with a brown hair
head of a woman
woman with long hair

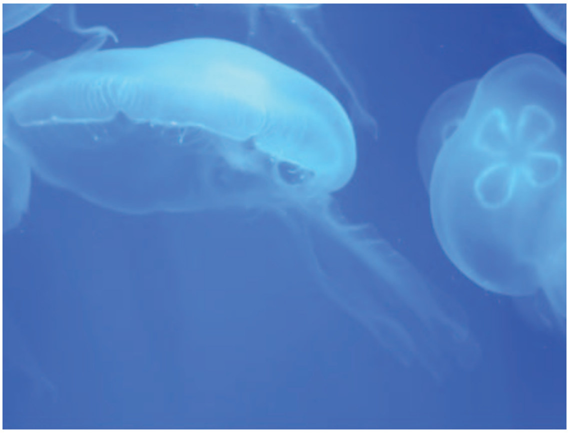
Trevor Paglen. *Image Operations. Op.10*, 2018. Still from single-channel color video projection, 5.0 surround sound, 23 min. © Trevor Paglen. Courtesy the artist; Altman Siegel, San Francisco; and Pace Gallery.

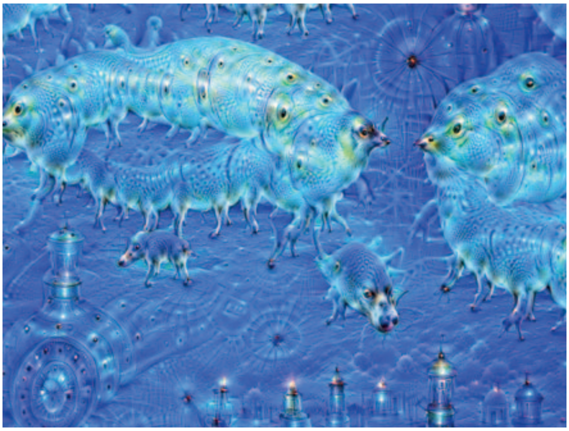
protagonists infiltrate the different layers of their targets' subconscious to plant or extract information—and to a 2014 paper, titled "Network in Network," that explores the consequences of the implementation of embedded internal complex structures within networks.[47]

Used to *find* and *enhance* faces and other patterns in images, DeepDream is a CNN that is run, in a way, in reverse order. Instead of starting with an *input*—that is, with an image that is given to a CNN to analyze and classify based on the images it has been trained with—one starts at the other end, with the *output*. After selecting a desired output—for example, the image of an animal's face—one tweaks the parameters of the neural network to force it to activate the artificial neurons that detect the various features of the animal's face. If the results of this tweaking and these activations are visualized, one sees, in the initial image (e.g., an image of a jellyfish), the desired features even if the image was completely devoid of them. If this process is reiterated enough times, these features begin to multiply, becoming more and more visible.

Another alternative to prescribing exactly which feature one wants the network to amplify is to let the network itself make the decision, based on how it has been trained. After receiving an image, the network begins to analyze it, starting from simpler features and moving gradually through the various layers of the CNN to more complex ones. If one picks a layer and asks the network to enhance what it has detected, some of the image's features, the ones that are related to images of the training set, will become more visible. If the operation is repeated, those features will begin to proliferate throughout the image.[48]

Initially produced as "technical metapictures" to better understand the structure of CNNs and the connections between patterns in the image and activations in the neurons, the images generated by DeepDream, once the algorithm was released as open source, began quickly to proliferate across the internet, popularizing the question of whether an "AI" could be visually "creative."[49] With their overprocessed, overinterpreted, hallucinatory qualities, they were seen as an example of "pareidolia": the impression of seeing a pattern (e.g., the shape of an animal, a face, or an object) emerging from a complex, confused visual stimulus such as an ink blot or a cloud formation. In





"A Sea of Data: Apophenia and Pattern (Mis-)Recognition," Hito Steyerl sees in these images "a striking visual example of pure and conscious apophenia," i.e., a heightened search for patterns within random data that inevitably leads to a form of "pattern overidentification" and in this way "reveals the networked operations of computational image creation, certain presets of machinic vision, its hardwired ideologies and preferences."[50]

From an art-historical perspective, DeepDream images may be considered in relation to the long tradition of "images hidden within images," such as images in clouds.[51] Over the following years, thanks also to a rather superficial formal similarity with the traditions of surrealist and psychedelic imagery, DeepDream images contributed to a widespread tendency to consider images generated by deep-learning algorithms as a form of "dream" or "hallucination" of the "machine," revealing, in a way, the "altered states" of AI itself. References to this idea can also be seen in the titles of extremely different artworks, such as Paglen's *Adversarially Evolved Hallucinations* (2017) and Refik Anadol's *Unsupervised: Machine Hallucinations* (2021) and *Renaissance Dreams* (2021).[52]

The idea of a "machine hallucination," though, is not confined to the realm of artworks using AI-generated images. Even though clearly metaphorical, *hallucination* is also a technical term in the language of deep learning. A "hallucination" is a situation in which a deep-learning algorithm gives a confident response that is not justified by its training data. A CNN trained to recognize cats in images that then mistakes a dog for a cat is an algorithm that "hallucinates." A chatbot that gives responses that contain invented facts or invented references is also "hallucinating." As in many other cases, complex, unintuitive mathematical processes related to deep-learning algorithms are described through metaphorical, anthropomorphic terms that make them somehow *perceivable*, thereby influencing the cultural and political reception of increasingly pervasive technologies labeled as "artificial intelligence."

The second type of deep-learning algorithms that began to be used in the mid-2010s to *generate* and *modify* images, rather than to classify them, are the so-called GANs. First introduced in 2014, their structure is made of two multilayer neural networks—called the "Generator" and "Discriminator"—that compete with each other in a zero-sum game, so that the gain of one network is the loss of the other and the sum of gains and losses is always zero.[53]

The game unfolds as follows: once the Discriminator (which is a CNN) has been trained to classify the images of a given initial training set, the other network, the Generator, starts producing images that it then submits to the Discriminator. The Discriminator then classifies the new images on

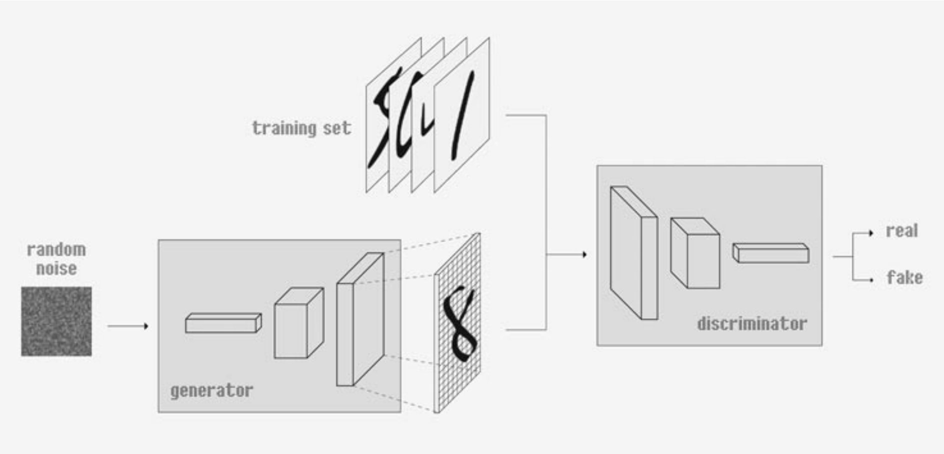
The image of a jelly fish (a) and the same image after applying several iterations of the DeepDream algorithm (b).

the basis of what it has learned from the training set, giving for each image submitted to it a numerical response between the extremes of 0 ("no," the image produced by the Generator does not belong to the initial training set; it is "false") and 1 ("yes," the image belongs to the training set; it is "true"). The closer the numerical response is to 1, the more the images produced by the Generator resemble those of the training set; the closer the numerical response is to 0, the less similar they are, while between 1 and 0 are several degrees of uncertainty.

At the beginning of the process, the Generator is given a latent space in which images have been embedded based on a series of parameters. If, for example, the images are photographs of birds, the embedding may be based on parameters that refer to the different shapes that birds may have in different positions, their different colors, and so on. From this initial latent space, the Generator begins by generating images that look like random noise; that is, random pixel configurations. Then, based on the responses received from the Discriminator, the Generator learns how to tweak its own parameters to generate from the latent space new images that look more and more like images of birds.

These exchanges take place in a process that resembles a real competition between "opponents." The Discriminator wins the game if it becomes capable of correctly determining whether the images produced by the Generator are identical or not to the ones of the initial training set. The Generator wins if it becomes capable of generating images that "fool" the Discriminator, leading it to give incorrect classifications.

The learning process of GANs is therefore interactive and relational, to some extent "social," since it involves two "agents" (two algorithms) that are in competition with and react to each other's actions.[54] Its functioning is based on a dual and competitive structure, and the speed of its learning depends on the number of exchanges between the Generator and the Discriminator. At its base, we find a statistical induction process through which the Generator learns to produce images whose patterns of pixels are more and more similar to the ones of the training set. These new images, however, are not "imitations" or "copies" of an "original" in any traditional sense of the term, since the Generator does not have direct access to the images of the training set: it receives only the responses of the Discriminator, which are "yes" or "no" formulated in numerical terms and with different percentage values.



During the training process of a GAN, the algorithm tweaks its own parameters to generate from the latent space different kinds of images. A latent space used to generate images is a multidimensional space in which each point, with its multiple coordinates described by a vector, may be visualized through an image. Taken together, the images corresponding to all the points within a latent space constitute all possible visualizations of the space. We can consider all these images as a sort of complete *cartography* or *atlas* of the latent space, the full map of all the images that a specific GAN, trained through the interactions with a specific Discriminator, can possibly generate.[55]

The content and the formal properties of an image generated by a GAN depend on the position, in latent space, of the point the image visualizes. Points that are *close* to one another in latent space generate images that are *similar* to one another; points that are *distant* from one another generate images that are *different*.

Connections between points (also called "interpolations") are instead visualized through moving images. What we see in this case, is a *trajectory* within the latent space: a gradual morphing from the image corresponding to the point from which the trajectory starts, to the image corresponding to the point where it ends. If the trajectory connects points that are close to one another in latent space, the moving images appear as a gentle, gradual morphing. If instead the trajectory connects distant points, the transitions are faster and more abrupt.

Moving images generated by GANs have their own specificities. Still, it is tempting to analyze them through concepts derived from theories of time-based visual media such as film and video. The concept of *montage*, for example—with its complex history, its different manifestations, and its ideas of "continuity," "discontinuity," "conflict," "tension," "distance," and "interval"—may be used to analyze moving images that visualize interpolations within latent space. Smooth transitions between images that visualize points that are close to one another in latent space may be considered a form of montage that emphasizes continuity, while sudden transitions that connect points that are distant from one another in latent space introduce forms of discontinuity, of conflict and tension between images.

For a few years now, GANs have been used either to *transform* existing images through a series of operations or to *generate* entirely new ones.

The use of GANs to *transform* existing images includes different kinds of so-called image-to-image translation: translation of satellite photographs to Google Maps, of photographs from day to night, of black-and-white photographs to color, and so on. Some of these translations give the impression of transposing an image from one medium to another (e.g., from a drawing

A generative adversarial network (GAN). Diagram by the author.

to a painting, or to a photograph), while others give the impression of transposing an image from one style (of a single artist, of an artistic movement, or of a historical period) to another style, through an operation called "style transfer." Other applications of GANs consist in changing photographs of human faces in order to show how an individual's appearance might change with age; animating still images (e.g., the photograph of a deceased person); allowing real-time face swapping (superimposing one person's face on another's body) and face reenactment (manipulating facial expressions), as in the case of deepfake videos; repairing a given image through a process of "inpainting," which completes or fills in parts that are damaged, deteriorated, or missing; inlaying images into other images, or, conversely, eliminating a body or an object in an image without leaving any visible trace; extending an image beyond its initial frame, through a process called "outpainting"; choosing a frame from a video and predicting the next frame; taking any given video and upscaling it by increasing its frame rate and its definition, reaching a level of so-called super-resolution.[56]

An emblematic example of this last application, which is currently proliferating across the internet and may end up altering significantly our experience of visual documents of the past, is the upscaled versions of Lumière films such as *L'arrivée d'un train en gare de La Ciotat* (*Arrival of a Train at La Ciotat*; 1896), in which the film is transposed from the original sixteen frames per second to a high frame rate of sixty frames per second, from the original 1.33:1 format to a contemporary 16:9 format (through cropping), and from the original, grainy 35 mm analog film to a 4K digital resolution.

In the case of all these *image operations* produced by GANs—translating, modeling, rejuvenating, aging, animating, simulating, transferring, inpainting, outpainting, predicting, upscaling—the new images that are produced are *images from images*: images resulting from the processing of the large quantities of images contained in the training sets. By upscaling the digital version of a Lumière film, for example, the GANs add to the initial images a series of pixels that, across various layers of algorithmic mediation, stem from many other images of other trains arriving in other stations. The new, upscaled Lumière film, in a way, inherits and absorbs within itself all these images, something that alters profoundly its temporal status. The upscaled video, which looks as if it had been shot not with a Lumière camera but with a much more recent digital camera, is a hybrid temporal object that contains, embedded in itself, pixels that are the visual traces of a series of temporal layers.

The use of GANs to *generate* new images, instead of *transforming* existing ones, has quickly changed throughout the years. GANs were initially introduced to produce images that were similar (but not identical) to the ones of a given dataset. In a 2014 paper, Ian Goodfellow and his collaborators used GANs to generate handwritten digits, new faces, and new objects that were similar to the ones of the MNIST handwritten-digit dataset, the Toronto Face Dataset, and the CIFAR-10 small-object photograph dataset, respectively. The idea was to increase the number of images that could be used for training image-recognition systems.

Shortly afterward, though, GANs began to be used to generate images that were not necessarily highly similar to the ones of the dataset used to train the Discriminator. In these cases, the GAN-generated images could be highly "photorealistic," hybrid, or entirely abstract, depending on the composition of the training sets, on the kinds of image translations that were performed, and on the points in the latent space that they visualized. Projects such as *This Person Does Not Exist* (Philip Wang, 2019) or *DoppelGANger.agency* (Mitra Azar, 2019) attracted a lot of attention when they were launched because they highlighted the capacity of new versions of GANs (such as StyleGAN and StyleGAN2, presented in the media as "artificial intelligence," as with DeepDream) to generate highly photorealistic images of faces of nonexistent people.[57] The photorealistic nature of these images—which were soon used in advertising and to create fake profiles on social media platforms—raised questions about how we ought to understand the impact of AI-generated images on the very idea of "photography" and a series of concepts traditionally related to it, such as "index" and "referent."[58] What exactly do these highly realistic images of nonexistent people represent? Do they have a referent? If so, where is it located? How should we explain their peculiar form of photorealism?

To answer these questions, we need to remember that GANs generate images out of a latent space that they have learned to explore through interactions with a Discriminator trained with a specific dataset. In the case of the images in *This Person Does Not Exist*, the dataset was composed of real photographs of real people, which means that the "photorealism" of the GAN-generated images is based on various layers of referentiality.

To begin with, these photographic "portraits" of nonexistent people refer—across multiple layers of algorithmic mediation—to the photographs of real faces that were part of the training set, and these photographs, in turn, refer to the real faces of the real people who were photographed. Then, these "portraits" refer to the categories that, within datasets such as ImageNet, have been used for their labeling. Finally, as with all images generated by GANs, they also refer to the specific points they visualize within the latent space that the GANs have learned to explore through their training. Different kinds of "referents" are therefore present at each of these layers.

The relations between the idea of "photography" and the images generated by GANs can also be analyzed from another point of view. As the artist and programmer Mario Klingemann suggests, these images may be considered a form of "cameraless neurophotography"; that is, as photographs or even "snapshots" that frame, capture, and visualize different areas of the latent space.[59] Anadol puts forth a similar idea: in commenting on his work with GANs for video installations such as *Unsupervised—Machine Hallucinations—MoMA* (2022), he talks about those moments, in the training of an algorithm, called "checkpoints," in which one has the chance "to see what the machine learns, and to take snapshots."[60]

In the field of contemporary art, images generated by various kinds of GANs appear in particularly interesting ways in the work of artists such as Paglen (*Adversarially Evolved Hallucinations*, 2017), Pierre Huyghe (*UUmwelt*, 2018), Steyerl (*Power Plants*, 2018; *This Is the Future*, 2019; *SocialSim*, 2020; *Animal Spirits*, 2022), and Grégory Chatonsky (*Second Earth*, 2019; *I Will Resemble What You Have Been*, 2020; and *Complétion 1.0*, 2021). In most cases, instead of using GANs to generate photorealistic images, these artists choose to visualize areas of the latent space in which images appear blurred, hybrid, and with various digital artifacts, as if to emphasize both the radical *otherness* of the images—the feeling of seeing something one has never seen before—and the vague, metamorphic traces of images one might partially recognize.[61]

The strategies according to which these GAN-generated images are used, though, are very different. Images from Paglen's *Adversarially Evolved Hallucinations*, for example, were generated through a process that the artist himself has openly explained, as if his aim were to render the image-generating potential of deep-learning algorithms more transparent. To produce the images in this work, Paglen began by establishing special training sets based on themes stemming from literature, philosophy, psychology, folk wisdom, and history, with titles such as "Interpretations of Dreams" (a collection of images showing symbols from Freudian psychoanalysis), "Omens and Portents" (images of comets, eclipses, and other natural events historically considered as supernatural), "American Predators" (images of predatory animals, plants, and human beings indigenous to the United States, and of military hardware such as drones and stealth bombers), or "Monsters of Capitalism" (images of monstrous creatures, such as vampires, that have at some point been associated with capitalism). Once these training sets were established (by collecting images from the ImageNet dataset),

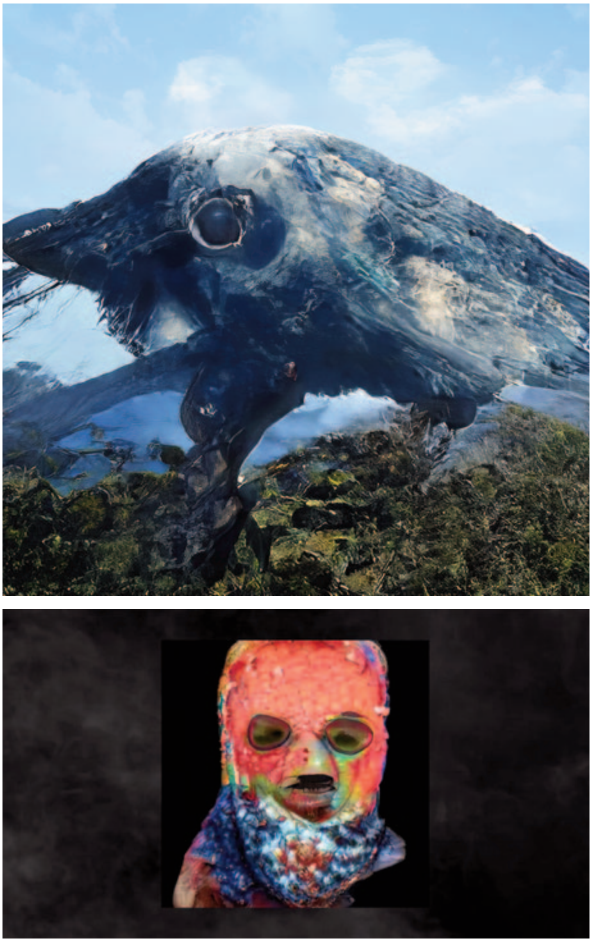
Trevor Paglen. *The Great Hall (Corpus: The Interpretation of Dreams)*, *Adversarially Evolved Hallucination*, 2017. Dye sublimation print, 32 × 40 in. (81.3 × 101.6 cm). © Trevor Paglen. Courtesy the artist; Altman Siegel, San Francisco; and Pace Gallery.

they were fed into the Discriminator, which learned how to detect and classify images belonging to such categories. The Generator, then, began from random noise to produce images that were progressively closer to the ones in the training set, with the goal of eventually fooling the Discriminator. Intervening within this process, stopping the training at specific "checkpoints," Paglen chose which images to extract from the ones that the GAN could generate. The result of this selection are images such as *The Great Hall (Corpus: The Interpretation of Dreams*) that visualize different areas of the latent space.[62]

GAN-generated hybrid images are central also in the work of Chatonsky, who describes them as the product of an "artificial imagination" (rather than "intelligence") that has processed the vast, hypertrophic quantities of images that humans have left on the internet to then release an endless flow of other images out of the latent space. This, for Chatonsky, is a space with a "flat ontology," in which the traditional ontological partitions that once structured the visible world as seen by human beings have dissolved. The images that visualize it are therefore fragments of a fluid, metamorphic, undifferentiated world in which the spectator encounters hybrid entities that preserve only vague traces of recognizable objects.

Second Earth (2019) presents these images as the product of machines that survive in a desolate, fully mineralized landscape after human extinction. After having acquired and processed all the images that human beings left in servers and data centers, deep-learning algorithms start producing other images that are "the hallucinations of a senseless machine, a monument dedicated to the memory of the vanished human species."[63] Among these images stemming from the latent space, one finds fragments of fictions, of possible futures and counterfactual pasts.

Complétion 1.0, instead, focuses on the peculiar, postphotographic "realism" of the images generated by GANs and on its difference from the idea of "realism" that has its roots in the trace-like, imprint-like, indexical nature of analog photographs. Chatonsky calls this new kind of realism an "inductive disrealism": the property of images whose indexical nature is not rooted in the material contact between



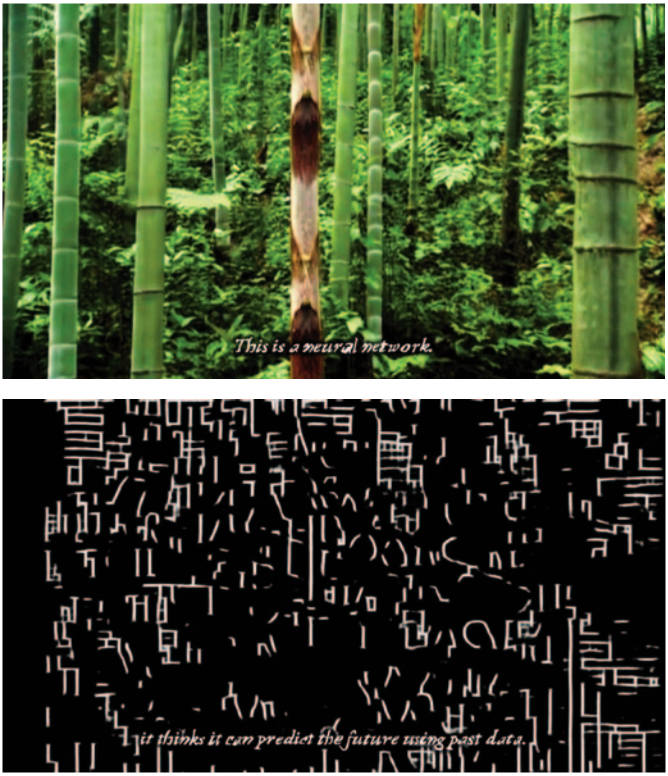
Grégory Chatonsky. *Second Earth*, 2019. Stills from video installation. © and courtesy the artist.

light and a photosensitive surface but rather in their reference to a latent space that GANs explore through a form of inductive learning.[64]

The question of the realistic and documentary nature of GAN-generated images is also raised by Steyerl's *This Is the Future* (2019), a video installation conceived as an expansion of the exhibition *Power Plants* (2018) at the Serpentine Gallery in London. What interested Steyerl in the use of GAN imagery was, in this case, the *predictive* nature of deep-learning algorithms: the fact that they are part of a vast spectrum of predictive systems that operate across contemporary societies, introducing various forms of algorithmic governmentality.

The main video presented in Steyerl's installation, titled *This Is the Future: A 100% Accurate Prediction*, consists for the most part of images produced through a next-frame prediction algorithm—a neural network that, when given a frame from a video, is trained to predict the next frame.[65] In *This Is the Future* the network is presented as a living entity with its own (synthetic) voice and vision—"I am a neural network. This is what I see"—and the images that document what it sees are presented as located "0.04 seconds in the future." They are capable of both *predicting* and *documenting* the future, as paradoxical as this may seem, since, as the synthetic voice says, they "enter into the future by slipping into the cracks between seconds."[66]

In all of these works the images generated by various kinds of GANs are not the output of completely autonomous algorithmic processes. On the contrary, they are always the result of a complex series of interactions between the artists, the programmers that in some cases collaborate with them, the algorithms (with their different versions, possibilities, and limitations), the images that are part of the training set, and the images that were generated out of the latent space. Exploring the latent space and the points and trajectories within it is a crucial moment in the artistic process, so much so that, in the end titles of her video *Animal Spirits* (2022), Steyerl lists, among the many operations of which she has been in charge, "latent space architecture and pathmaking." During the production process, what we see is therefore an inductive series of trials and errors that unfolds through a continuous intertwining


This is a neural network.
it thinks it can predict the future using past data.

Hito Steyerl. *This Is the Future: A 100% Accurate Prediction*, 2019. Stills from single-channel color video, sound, 16 min. From the video installation/environment *This Is the Future* (2019). © Hito Steyerl. Courtesy the artist and Andrew Kreps Gallery, New York.

of human intentions and technical possibilities. The "authorship" of these GAN-generated images is therefore fundamentally *distributed*, spread across layers of actions and operations. This is true also for the most recent generation of deep-learning algorithms dealing with images: the text-to-image and image-to-text models.

The Visible and the Sayable: Prompt-Based Imaging

The third phenomenon highlighted at the beginning of this article—the widespread diffusion of deep-learning algorithms that are capable of *generating images from texts* and of *generating texts from images*—is very recent. *Image-to-text models* have their roots in machine-vision technologies capable of analyzing, labeling, and classifying images, but they have been made more complex in recent years by being combined with LLMs trained with vast textual datasets and used for tasks of natural-language processing, such as text generation and translation. *Text-to-image models* were first tested in the mid-2010s (with programs such as alignDRAW) and then quickly developed during the next few years, mostly through different types of GANs, but became popular only in 2022 with the introduction of so-called diffusion models such as DALL-E 2, Stable Diffusion, and Midjourney.[67]

Image-to-text models usually operate in two steps: *image analysis* and *text generation*. First, a CNN analyzes the image to detect, label, and classify the objects, the actions, and the spatial relations it represents. Then, another deep neural network connects this initial output with a linguistic structure. Models such as DenseCap, for example, analyze an image and generate captions automatically, while CLIP generates descriptions, and other models such as Neural Storyteller can generate short stories. GPT-4 goes one step further: once the description of an image has been generated with an image-to-text model such as CLIP, the description is fed into GPT-4, which then answers questions about the image (e.g., What would be the consequence of making a given change in the image? What is the meaning of a meme? or Why is this image funny?) or performs tasks based on it (e.g., generating a website from a simple drawing).

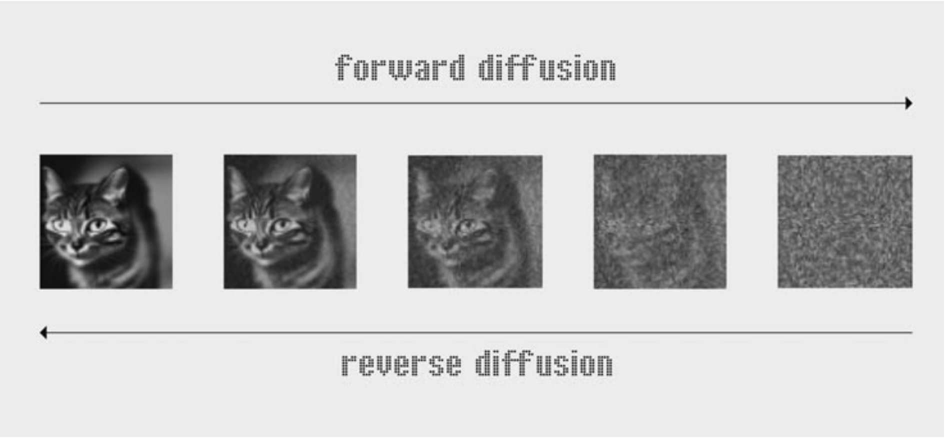
Text-to-image models also operate in two steps: *text encoding* and *image generation*. A transformer model similar to the ones that are used for natural-language processing and translation is used to transform the input text into a latent representation: a series of numerical vectors associated with the individual words and their relations within a sentence. A diffusion model then generates an image from that latent representation.

To better understand how they work, consider the example of Stable Diffusion, which, unlike DALL-E 2 and Midjourney, has openly released its code and model weights.[68] The diffusion model operates in two phases: forward diffusion and reverse diffusion. The first phase, *forward diffusion*, turns the images of the dataset used for training into indistinguishable "noise images" (i.e., random pixels distributed within the grid-like image space). It does so by adding increasing levels of "noise" to each image (e.g., the image of a cat). The term *diffusion*, in this case, refers to the concept of diffusion in physics: the movement of particles from an area of high concentration to an area of low concentration, until equilibrium is reached.

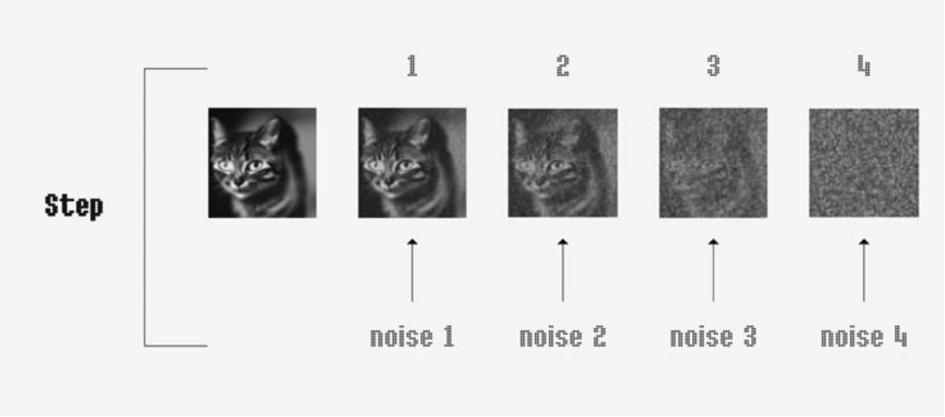
The second phase of a diffusion model is called *reverse diffusion*. During this phase, starting from the "noise images," the model learns how to recover the initial images of the training set. It does this through a "noise predictor" that learns to *predict* how much noise was added to the initial images. This allows the model to subtract for each given "noise image" the layers of noise that were added, until it finds the initial image (e.g., again, the image of a cat).

This double process of *forward diffusion* and *reverse diffusion* would be extremely slow and require very high computing power—inaccessible to most private users—if the images of the training set were not somehow *compressed*. The first version of Stable Diffusion was trained with images having a resolution of 512 by 512 pixels, which means that, keeping in mind that each pixel has three color channels (red, green, and blue), the "image space" of a 512 by 512 image is a 786,432-dimensional space.[69] Instead of operating in such high-dimensional image space, Stable Diffusion operates as a "latent diffusion model." That is, both the forward

Top: Forward diffusion and reverse diffusion in diffusion models. Diagram by the author.



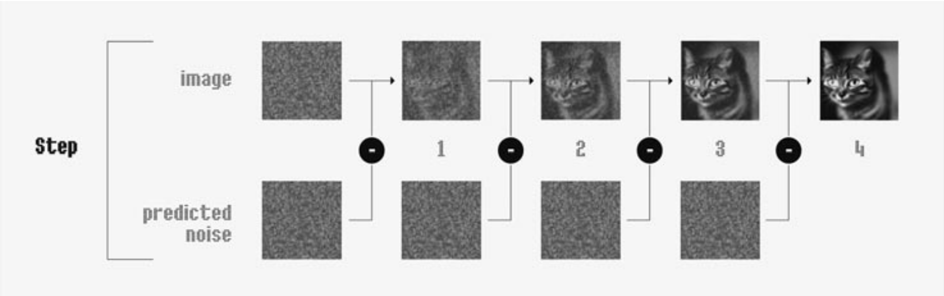
Bottom: From an initial image to a "noise image" through forward diffusion in text-to-image models. Diagram by the author.



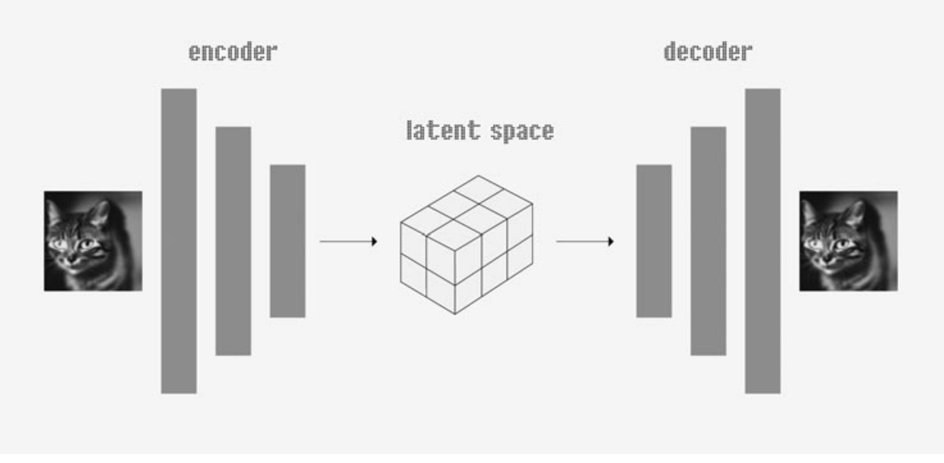
and the reverse diffusion processes happen in a compressed, lower-dimensional latent space, instead of in the initial image space.[70]

Stable Diffusion also makes use of another phase, called "conditioning": its aim is to steer the noise predictor in such a way that, once the layers of noise have been subtracted, the algorithm can produce an image that is perceived as being "aligned"—that is, coherent with—the textual prompts that were used to generate it. Text prompts are first "tokenized." This entails each word of the prompt (e.g., "photo of a cat") being associated, through a deep-learning model called CLIP tokenizer, with a number called a "token" (e.g., *photo* might be associated with the number 50, *of* with 24, *a* with 59, and *cat* with 239). Then, each token is "embedded"; that is, converted into a 768-value vector (a string of numbers indicating coordinates in a 768-dimensional space), which contains the parameters that allow it to locate a given word in relation to other words of a given language (e.g., English), as also happens with language processing algorithms such as GPT-4, ChatGPT, Google Translate, and DeepL. The embeddings are then processed by a "text transformer" and finally fed into the noise predictor, which begins to subtract layers of noise until it reaches an image that is somehow "aligned" with the prompts.

The result of this complex process of *forward diffusion* and *reverse diffusion*, of *text encoding* and *image generation*, is that models such as DALL-E 2, Stable Diffusion, and Midjourney take as input natural-language descriptions or "prompts" and then generate as output a series of still images that are programmed to be different each time, even if one repeats the same prompts, thus promoting the idea of an *infinite generativity* of such algorithms. Further versions of these models, called *text-to-video*, make possible the generation not only of still but also moving images. In the



Top: "Noise images" and the subtraction of predicted noise in text-to-image models. Diagram by the author.



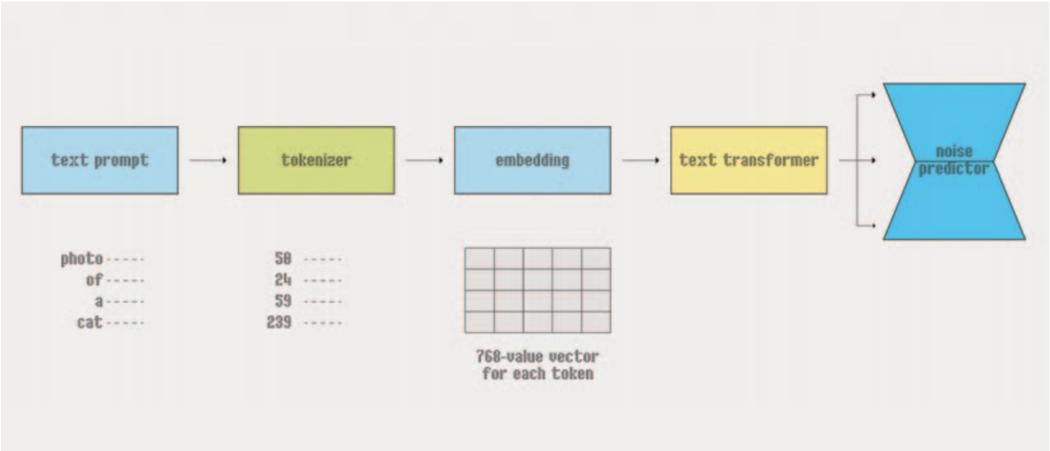
Bottom: Elements of the compression process in a latent-diffusion model of text-to-image models: encoder, latent space, decoder. Diagram by the author.

near future, it will be possible to generate entire videos or films exclusively from textual prompts, replacing or integrating a whole series of analog or digital techniques of image capture, editing, and postproduction.

Prompts are destined to become a key factor in a visual culture more and more based on the operation of submitting words and texts to deep-learning algorithms in order to explore their latent space and have images emerge out of it. Acting as a new kind of interface that, almost as a series of "speech acts," operates a transition from natural language to images, prompts may consist of single words or entire phrases that give indications concerning content (i.e., the elements to be included or omitted from the desired image), composition, color, lighting, style (whether of a single artist or of an artistic movement), and medium (e.g., drawing, painting, animation, photography, video, including detailed specifications about camera models, lenses, etc.). Prompts, in other words, *re-mediate* previous visual media into a set of terms indicating material supports, devices, and operations (e.g., "DVD screengrab") that can be used to generate images out of a latent space in which the formal properties of all these media exist as visual patterns that can flow from image to image.

Prompts may also give *negative* indications: they can indicate qualities perceived as negative ("ugly," "bad," "deformed," "disfigured"), qualities that are not usually searched for ("bad art," "poorly drawn," "out of focus," "out of frame"), or features that are typically unwanted ("extra limbs," "extra legs," "extra arms"). A flurry of prompt tutorials available online helps users refine their queries and bypass, in certain cases, the absence of prompts that have been removed from the platforms for so-called safety reasons, through different forms of content moderation and prompt censorship.

Soon after their launch, text-to-image models such as DALL-E 2, Stable Diffusion, and Midjourney began to proliferate across the internet and to be used for a wide variety of applications, some of which include operations previously performed by GANs, such as inpainting, outpainting, style transfer, and upscaling. Images generated through diffusion models quickly began to appear on book and magazine covers and in advertising campaigns, raising questions about whether entire skill sets and even entire professions in the fields of illustration, graphic design, photography, and video—including the need to use human models in fashion advertising—might soon be radically transformed or even entirely replaced by prompt-based techniques.[71]



The capacity of text-to-image models to generate photorealistic imagery has also been used to generate all sorts of fake, hypothetical, and counterfactual imagery, such as Donald Trump fighting with NYPD police officers trying to arrest him, French President Emmanuel Macron running through a burning Paris during the demonstrations against pension reform, and a video released in April 2023 by the Republican National Committee with the title *Beat Biden* showing the imaginary apocalyptic consequences of a second Joe Biden presidency.[72] After their first widespread diffusion during the mid-2010s, so-called deepfakes are now widely produced through text-to-image models, further undermining the trust in images and highlighting the need for new forms of image forensics and fact-checking.

In other cases, photorealistic images generated by text-to-image models have been used for completely different, "documentary" purposes. In May 2023, Amnesty International received widespread criticism for its decision to use AI-generated images to denounce police violence in Colombia, justifying the choice as a way of protecting the identity of protesters.[73] In a project titled *Exhibit AI—The Refugee Account*, a group of Australian lawyers took as a starting point the thirty-two written statements of survivors of Australia's offshore detention centers in Manus Island, Nauru, and Christmas Island, and, given the lack of photographic or video documentation of these centers, used the statements as prompts to generate images that, after being discussed with the survivors themselves, were meant to somehow "document" events that had not been visually recorded.[74]

As was the case for CNNs and GANs, understanding the sources, the content, the structure, and the guiding principles of the datasets used for the training of text-to-image models such as DALL-E 2, Stable Diffusion, and Midjourney is crucial if one wants to understand the kinds of images they might generate, the values that might be associated with these images, and the operations in which they might be involved. Study of the training sets is also essential for understanding how they connect texts and images by introducing a new, algorithmic correlation between the textual and the visual whose effects, across contemporary visual culture, have yet to be fully understood.

Text-to-image models are all trained through vast quantities of paired images and captions, and while OpenAI has not released information about which datasets were used to train DALL-E 2, we know that Stable Diffusion was trained with the Large-scale Artificial Intelligence Open Network 5B (LAION-5B), released in March 2022: a publicly available dataset derived from Common Crawl data freely scraped from the internet.[75]

To gather the data, LAION-5B searched the crawled websites for the metadata that is connected to all digital images uploaded to the internet,

Elements in the "conditioning" process of text-to-image models: text prompt, tokenizer, embedding, text transformer, noise predictor. Diagram by the author.

focusing in particular on the <img> tags used to embed images in html web pages and treating the "alt attributes"—the "alternative texts" that provide a concise description of the images and replace them when they cannot be visualized—as captions.[76]

As of this writing, LAION-5B contains five billion text-image pairs, sourced from across the internet from search engines, stock-image databases, social-media platforms, and various websites such as Google Images, Shutterstock, Getty Images, Pinterest, WordPress, Flickr, Twitter, ArtStation, DeviantArt, and many others. The images were taken from the internet without permission, raising the question of whether copyright protections should extend to the inclusion of images in large datasets, given their crucial role in the generation of other images through prompts. A series of artists whose works ended up in LAION-5B and whose names in many cases were used in prompts ("in the style of . . .") decided to sue Stability AI and to raise awareness through initiatives such as *Have I Been Trained?*[77]

The criteria according to which the five billion text-image pairs contained in LAION-5B are organized and filtered seem to be continually updated based on feedback from users and developers. After the launch of LAION-5B in March 2022, new criteria were added. In August 2022, "LAION-Aesthetics" was introduced: a subset of text-image pairs containing images with a "high predicted aesthetic score," determined by training a model capable of "predict[ing] the rating people gave when they were asked '*How much do you like this image from a scale of 1 to 10?*'"[78] Behind these generic "people" we actually find individuals generating and rating images through deep-learning text-to-image models on various platforms, or posting images on websites dedicated to photography contests; these images are then gathered in datasets such as SAC (Simulacra Aesthetic Captions) or AVA (Aesthetic Visual Analysis), which introduce further layers of selection and algorithmic processing in these "predicted" aesthetic criteria.

Other upgrades had to do with generating better captions for images, selecting images in higher resolution, eliminating images with watermarks, detecting NSFW ("not safe for work") content in images, or allowing individuals whose names and whose images were found to be in LAION-5B to submit a "takedown form" asking to be removed from the dataset.[79]

LAION-5B shows that the content and formal properties of the images generated by text-to-image models are the outcome of a complex process in which various kinds of technical and normative criteria, statistically established aesthetic preferences, and biases flow from the specific platforms from which the images have been taken (and, as vast and transnational as these platforms may be, they are *never* neutral containers), to the training sets (with their various filters and criteria), all the way down to the images generated through the prompts. These, as Steyerl argues, can be considered "statistical renderings" that "shift the focus from photographic indexicality to stochastic discrimination," given the importance of statistics and probability across all the algorithmic processes involved in their production.[80]

We can easily predict that, in the near future, large quantities of images generated through diffusion models such as Stable Diffusion, DALL-E 2, and Midjourney, after having been uploaded to the internet and rated by users, will end up in new datasets that will, through a continuous feedback loop, contribute to the training of future models, thus increasingly promoting the circulation of specific motifs and specific image styles connected to the different algorithms.

That said, the training set for a model such as Stable Diffusion—or any other deep-learning model—is never final. Users may implement additional training to embed in the dataset new images that were not initially present. This opens up the possibility of using new prompts and therefore of extending the spectrum of possible images that the model may generate out of the latent space.

An interesting example of this possibility appears in Chatonsky's image series *His Story* (2022). Using a model called DreamBooth, the artist added, in the category "person," a series of images of himself, accompanied by the prompt "gchatonsky." He then used this prompt to generate images of himself that became part of a counterfactual autobiography in which a photorealistic but algorithmically generated Chatonsky appears in various spatial and historical contexts.

In yet other works, Chatonsky chose other strategies to test the potential of diffusion models, exploring various areas of their latent space. In *La machine 100 têtes* (*The Hundred Headless Machine*; 2022), for example, he revisited Max Ernst's first collage novel, *La femme 100 têtes* (*The Hundred Headless Woman*; 1929) through the help of DALL-E 2. Using English translations of the legends of Ernst's uncanny, oneiric collages as prompts, Chatonsky generated new images that stem not so much from a surrealist, psychic *automatism* as from a kind of algorithmic *automation*.[81] Another work titled *The Kiss* (2022) takes the script of Alfred Hitchcock's *Vertigo* (1958) and uses parts of it as prompts to generate still and moving images that visualize possible, AI-generated versions of the film.[82]

Grégory Chatonsky.
Image from the series *His Story*, 2022. © and courtesy the artist.

Finally, for *Counterfeits* (2021), Chatonsky used both image-to-text and text-to-image models. Starting from well-known artworks (e.g., Pablo Picasso's 1907 *Demoiselles d'Avignon*), a short description was generated through Neural Storyteller. The text was then given to another program, Zoetrope, which generated an image starting from the text used as a prompt, initiating a potentially endless series of algorithmic translations from images to texts to images and so on.[83]

The questions raised by text-to-image models are many. Space considerations allow me to highlight only some of them, and we also need to keep in mind that these recent algorithms are developing at a rapid pace. As in the case of GAN-generated images, the images generated by text-to-image models are not the output of completely autonomous algorithmic processes. An analysis of their structure and their normative criteria, the datasets used to train them, and the operations performed in order to activate them, shows a series of layered intertwinings between human and technical agencies, even though the distribution of these agencies may change in time (for example, human-written "alt attributes" may soon be replaced by forms of automated captioning).

Each image generated by a text-to-image diffusion model is, as in the case of GANs, a visualization of one of the "points" (a vector) of the latent space generated by the model through its training. If we search for the "referent" of an image generated by one of these models, we are faced with a complex, layered referentiality that cuts across various forms of mediation, involving both images and words and reaching all the way up to the images captured or produced by human beings for other human beings and then uploaded to the internet, along with the words that accompanied them, as captions, in the "alt attributes." Within the flat ontology of the latent space, every point has the same status as all other points: each is defined by a vector (a set of coordinates), which, in the case of diffusion models, results from the embedding of both images and texts. The images that visualize each point, though, may differ significantly in content, formal features, style, and so on, depending on the position of the vector in the multidimensional latent space.

Another important aspect of text-to-image diffusion models is that natural (i.e., human) language becomes the main medium for image generation: what is *visible* is strictly correlated with what is *sayable*, since the images

are generated by textual prompts. What can or cannot be *said*, what can or cannot be *written* in a prompt, determines—together with the other factors analyzed above—what can or cannot be *visualized* and *seen*. The content of the "alt attributes"—together with the various limitations that models such as DALL-E 2, Stable Diffusion, and Midjourney have introduced in the use of prompts (e.g., "not safe for work")—create boundaries that restrict the spectrum of images that may be generated. This introduces new areas of invisibility that prolong in new, still uncharted ways, the long history of "forbidden images."[84]

As was already the case with CNNs and GANs, the new visual culture that these text-to-image models are promoting is one within which images and words are inextricably connected. Against a whole tradition, in both art history and image theory, that tried to underline the autonomy of images, image analysis, and visual experience from texts and text-based disciplines, deep-learning algorithms are now leading us into a new visual landscape in which images and words are increasingly inseparable.[85]

Opposite: Grégory Chatonsky. *The Kiss*, 2022. Still from video.

Left: Grégory Chatonsky. Image from the series *Counterfeits*, 2021.

Notes

I thank Red Giuliano, cofounder and CEO of the machine-learning development platform Zero-True, for the many discussions and suggestions about the technical aspects of this text. I also thank Noam Elcott and Zeynep Çelik Alexander, as well as Erika Balsom, Weihong Bao, Estelle Blaschke, Craig Buckley, Francesco Casetti, Jan Distelmeyer, Laura Frahm, Alexander Galloway, Tom Gunning, Mark Hansen, Christian Joschke, Gertrud Koch, Elisa Linseisen, Olivier Lugon, Pietro Montani, Ariel Rogers, and Holger Schulze for their very useful comments on this text. Finally, special thanks to Grégory Chatonsky, Trevor Paglen, and Hito Steyerl for sharing with me insights about the works discussed in this essay.

[1] The literature on the history of the idea of "artificial intelligence" is immense. For a synthetic account of how the very concept of "artificial intelligence" has been understood since its first introduction in the mid-1950s, see Stephanie Dick, "Artificial Intelligence," *Harvard Data Science Review* 1, no. 1 (Summer 2019), https://hdsr.mitpress.mit.edu/pub/0aytgrau/release/3. In the article, Dick underlines the fact that "there isn't a straightforward narrative of artificial intelligence from the 1950s to today" and that "what counts as *intelligence* is a moving target in the history of artificial intelligence."

[2] On media archaeology, see Thomas Elsaesser, "Film History as Media Archaeology," in *Film History as Media Archaeology: Tracking Digital Cinema* (Amsterdam: Amsterdam University Press, 2016), 71–100; Erkki Huhtamo and Jussi Parikka, eds., *Media Archaeology: Approaches, Applications, and Implications* (Berkeley and Los Angeles: University of California Press, 2011); and Jussi Parikka, *What Is Media Archaeology?* (Cambridge, UK: Polity, 2012).

[3] On algorithmic governmentality, see Antoinette Rouvroy and Thomas Berns, "Algorithmic Governmentality and Prospects of Emancipation," *Réseaux* 177, no. 1 (January 2013): 163–196.

[4] The term *algorithmic images*, which is here used specifically in relation to deep-learning algorithms, could also be used in a broader sense to refer to the longer history of images generated or processed by various kinds of algorithms (i.e., finite sequences of instructions aimed at solving a class of specific problems or at performing some kind of computation). From this perspective, all digital images (including many nondigital ones) are in some way "algorithmic." In "The Algorithmic Turn: Photosynth, Augmented Reality and the Changing Implications of the Image," *Visual Studies* 26, no. 1 (March 2011): 25–35, William Uricchio analyzes what he considers to be a new "algorithmic construction of the image" promoted by the diffusion of "location-aware technologies" and "image recognition–based augmented reality applications" that introduce new "algorithmically defined relations between the viewing subject and the world viewed" (25). See also Ruggero Eugeni, *Capitale algoritmico: Cinque dispositivi postmediali (più uno)* (Brescia: Scholé, 2021), in which "computational images" are treated as "algorithms."

[5] On the concept of "digital objects," see Yuk Hui, *On the Existence of Digital Objects* (Minneapolis: University of Minnesota Press, 2016).

[6] John McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 31 August 1955, in *AI Magazine* 27, no. 4 (2006): 12–14.

[7] For an analysis of the institutional and political reasons behind the split between symbolic and subsymbolic approaches, see Chris Wiggins and Matthew L. Jones, *How Data Happened: A History from the Age of Reason to the Age of Algorithms* (New York: Columbia University, 2023).

[8] See Ronald Kline, "Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence," *IEEE Annals of the History of Computing* 33, no. 4 (2010): 5–16.

[9] For an analysis of the social, political, and environmental implications of the recent

development of AI technologies, see Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven: Yale University Press, 2021). On the relations between AI, statistics, and capitalism, see Justin Joque, *Revolutionary Mathematics: Artificial Intelligence, Statistics, and the Logic of Capitalism* (London: Verso, 2022). On machine learning as a form of knowledge production and a strategy of power, see Adrian Mackenzie, *Machine Learners: Archaeology of a Data Practice* (Cambridge: MIT Press, 2017).

10. On natural-language processing and LLMs with regard to ChatGPT, see Alexandre Gefen, *Vivre avec ChatGPT* (Paris: L'Observatoire, 2023).

11. For a study of the history and the implications of the concept of "visual culture," whose first occurrences can be found during the 1920s and 1930s in the writings of figures such as Béla Balázs, László Moholy-Nagy, and Jean Epstein, see Andrea Pinotti and Antonio Somaini, *Culture visuelle: Images, regards, médias, dispositifs* (Dijon: Les Presses du Réel, 2022).

12. In a rasterized black-and-white (greyscale) image, each pixel has a value from 0 (black) to 255 (white), the numbers between 1 and 244 being different scales of grey. In a rasterized color image, each pixel, alongside its coordinates, has three values, each one corresponding to the intensity of each of the three fundamental colors (red, green, blue). The values range from 0 (the darkest shade of red, green, or blue) to 255 (the lightest shade) for 8-bit screens, from 0 to 1,028 for 10-bit screens, and even higher for the 12- or 16-bit screens used for professional color grading.

13. On the grid as a "cultural technique" (*Kulturtechnik*), 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*, trans. Geoffrey Winthrop-Young (New York: Fordham University Press, 2015), 97–120. On grids, see also Rosalind A. Krauss, "Grids," in *The Originality of the Avant-Garde and Other Modernist Myths* (Cambridge: MIT Press, 1985), 9–22. On the concept of "addressability," see Friedrich A. Kittler, "Computer Graphics: A Semi-technical Introduction," *Grey Room*, no. 2 (Winter 2001): 30–45.

14. Frank Rosenblatt, *The Perceptron, a Perceiving and Recognizing Automaton*, Report 85-460-1 (Buffalo: Cornell Aeronautical Laboratory, 1957). A year after this report, Rosenblatt published an article in which he explained that the Perceptron was part of his attempt "to understand the capability of higher organisms for perceptual recognition, generalization, recall, and thinking." Frank Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review* 65, no. 6 (1958): 386–408.

15. On McCulloch and Pitt's studies on artificial neurons, see Warren McCulloch and Walter Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* 5 (1943): 115–133.

16. The two main "AI winters" happened from 1974 to 1980 and from 1987 to 1993.

17. See Frank Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Washington, DC: Spartan Books, 1962). The Cognitron (1975) and Neocognitron (1980) were developed by Kunihiko Fukushima, who was inspired by Rosenblatt's *Principles of Neurodynamics* and David H. Hubel's and Torsten N. Wiesel's studies on the receptive fields in biological neurons. See Kunihiko Fukushima, "Cognitron: A Self-Organizing Multilayered Neural Network," *Biological Cybernetics* 20, no. 3 (1975): 121–136; and Kunihiko Fukushima, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," *Biological Cybernetics* 36, no. 4 (1980): 193–202. The phrase "back-propagating error correction" was first introduced in 1962 by Frank Rosenblatt. In 1986, David Rumelhart, Geoffrey Hinton, and Ronald Williams published an article that presented an experimental analysis of the technique, whose origins may be found in the Leibniz chain rule. In 1989, Yann LeCun and

colleagues developed a neural network that used the back-propagation algorithm to train a neural network with a dataset of 9,298 handwritten digits compressed in greyscale images of sixteen by sixteen pixels. Geoffrey Hinton and Ronald Williams, "Learning Representations by Back-Propagating Errors," *Nature* 323, no. 6088 (1986): 533–536; and Yann LeCun et al., "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation* 1, no. 4 (1989): 541–551.

18. On CAPTCHAs, see Jimena Canales, "Art in the Age of Captcha," in *Philippe Decrauzat: DELAY*, ed. Mathieu Copeland (Cologne: Walther und König, 2022), 139–143.

19. See Antonio Casilli, *En attendant les robots* (Paris: Seuil, 2019).

20. On machine vision and its aesthetic, epistemological, and political implications, see Mitra Azar, Geoff Cox, and Leonardo Impett, eds., "Ways of Machine Seeing," special issue, *AI and Society* 36, no. 4 (2021): 1093–1312. On the history of machine vision, see James E. Dobson, *The Birth of Computer Vision* (Minneapolis: University of Minnesota Press, 2023).

21. In *supervised learning*, human labor intervenes in both the choosing and the labeling of the images of the training set. Starting from such labeled images, the algorithm then learns how to detect, recognize, and classify entities appearing in new, unlabeled images that were not part of the training set. *Unsupervised learning*, in contrast, operates directly on images that are not labeled. It may help in identifying pattern similarities among images so they can be clustered in separate groups.

22. On the opacity of CNNs as a form of "inscrutability" and "nonintuitiveness," see Andrew D. Selbst and Solon Barocas, "The Intuitive Appeal of Explainable Machines," *Fordham Law Review* 87, no. 3 (2018): 1085.

23. That number was twenty in the case of the datasets used from 2005 to 2010 for the PASCAL Visual Object Classes competition. In 2012, a CNN called "AlexNet" won the ImageNet Large Scale Contest, a moment that is widely considered to be a turning point in the development of machine-vision systems. See Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, ed. F. Pereira et al. (Red Hook, NY: Curran Associates, 2012), 1097–1105.

24. Jia Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: IEEE), 248–255.

25. Fei-Fei Li, as quoted in Dave Gershgorn, "The Data That Transformed AI Research—and Possibly the World," *Quartz*, 26 July 2017, https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world.

26. Kate Crawford and Trevor Paglen, "Excavating AI: The Politics of Training Sets for Machine Learning," 19 September 2019, https://excavating.ai/. On the ethics, origins, and individual privacy implications of face-recognition datasets, see also the project Exposing.ai developed by Adam Harvey and Jules LaPlace: https://exposing.ai/.

27. At the top of the ImageNet taxonomy are nine major categories: "plant," "geologic formation," "natural object," "sport," "artifact," "fungus," "person," "animal," "miscellaneous." Underneath, a whole series of subcategories, organized according to a taxonomy that, especially in the case of the "person" category (temporarily deactivated in 2019), is full of bizarre labels, highly questionable assumptions, and various algorithmic biases. For an image of the Amazon Mechanical Turk basic user interface, see Crawford and Paglen, "Excavating AI."

28. The idea of a total and objective mapping of "the entire world of objects" that lies at the basis of ImageNet could be analyzed in relation to the various understandings of "objectivity" (in particular, "trained objectivity") that are discussed in Lorraine Daston and Peter Galison, *Objectivity* (New York: Zone Books, 2010). I thank Noam Elcott for highlighting this.

29. Crawford and Paglen, "Excavating AI." See also Adam Greenfield, *Radical Technologies: The Design of Everyday Life* (London: Verso, 2017).

30. The controversial case of Clearview AI raised a series of questions that are highlighted in Kashmir Hill, "The Secretive Company That Might End Privacy as We Know It," *New York Times*, 18 January 2020, https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html. Also see Monica Steinberg, "Extralegal Portraiture: Surveillance, between Privacy and Expression," *Grey Room*, no. 87 (2022): 66–99.

31. See, for example, Amanda Wasielewski, *Computational Formalism: Art History and Machine Learning* (Cambridge: MIT Press, 2023); Nuria Rodríguez-Ortega, "Image Processing and Computer Vision in the Field of Art History," in *The Routledge Companion to Digital Humanities and Art History*, ed. Kathryn Brown (New York: Routledge, 2020), 338–357; and Leonardo Impett, "Analyzing Gesture in Digital Art History," in *The Routledge Companion*, 386–407.

32. See, for example, Lior Shamir et al., "Impressionism, Expressionism, Surrealism: Automated Recognition of Painters and Schools of Art," *ACM Transactions on Applied Perception* 7, no. 2 (2010): 1–17; Sergey Karayev et al., "Recognizing Image Style" (2013), arXiv, https://arxiv.org/abs/1311.3715; and Ahmed Elgammal et al., "The Shape of Art History in the Eyes of the Machine" (2018), arXiv, https://arxiv.org/abs/1801.07729.

33. Among the companies currently offering "AI art authentication" services, see, for example, Art Recognition, https://art-recognition.com/.

34. Rodríguez-Ortega, "Image Processing and Computer Vision," 339.

35. Rodríguez-Ortega, "Image Processing and Computer Vision," 341–343. On the role of the double slide projection within the wider field of "comparative looking" in art historiography, see Zeynep Çelik Alexander, *Kinaesthetic Knowing: Aesthetics, Epistemology, Modern Design* (Chicago: University of Chicago Press, 2017), ch. 2. On the way in which slow motion and freeze-frame transformed film viewing and film analysis, see Laura Mulvey, *Death 24x a Second* (London: Reaktion Books, 2005).

36. On the concept of "instrumental image," see Allan Sekula, "The Instrumental Image: Steichen at War," *Artforum* 13, no. 5 (1975): 36–45. On Sekula, see also Marie Muracciole and Benjamin J. Young, eds., "Allan Sekula and the Traffic in Photographs," special issue, *Grey Room*, no. 55 (Spring 2014).

37. Among the texts in which Farocki discusses the notion of "operational images," see, in particular, Harun Farocki "Phantom Images," *Public*, no. 29 (2004): 12–22; and Harun Farocki "Quereinfluss/Weiche Montage," *New Filmkritik*, 12 June 2002, https://newfilmkritik.de/archiv/2002-06/quereinflussweiche-montage/. On Farocki and operational images, see Christa Blümlinger, *Harun Farocki, du cinéma au musée* (Paris: P.O.L., 2022); and Volker Pantenburg, "Working Images: Harun Farocki and the Operational Image," in *Image Operations: Visual Media and Political Conflict*, ed. Jens Eder and Charlotte Klonk (Manchester, UK: Manchester University Press, 2017), 49–62. On operational (or "operative") images, see also Aud Sissel Hoel, "Operative Images: Inroads to a New Paradigm of Media Theory," in *Image—Action—Space*, ed. Luisa Feiersinger, Kathrin Friedrich, and Moritz Queisner (Berlin: De Gruyter, 2018), 11–27; and Jussi Parikka, *Operational Images: From the Visual to the Invisual* (Minneapolis: University of Minnesota Press, 2023).

38. Trevor Paglen, "Operational Images," *e-flux Journal*, no. 59 (November 2014), https://www.e-flux.com/journal/59/61130/operational-images/.

39. Andreas Broeckmann, "Optical Calculus" (paper presented at the Images beyond Control conference, FAMU, Prague, 6 November 2020), video available at https://www.youtube.com/watch?v=FnAgBbInMfA; Adrian MacKenzie and Anna Munster,

"Platform Seeing: Image Ensembles and Their Invisualities," *Theory, Culture and Society* 36, no. 5 (2019): 3–22; and Fabian Offert and Peter Bell, "Perceptual Bias and Technical Metapictures: Critical Machine Vision as a Humanities Challenge," *AI and Society* 36, no. 4 (2021): 1133–1144. In their essay, Offert and Bell also argue for "critical machine vision as an important transdisciplinary challenge, situated at the interface of computer science and visual studies/*Bildwissenschaft*." On the concept of "invisual," see also Parikka, *Operational Images*.

40. For a first attempt in this direction, see Antonio Somaini, "L'impact de l'intelligence artificielle sur la culture visuelle contemporaine," in *Culture visuelle*, 367–417. See also Joanna Zylinska, *Nonhuman Photography* (Cambridge: MIT Press, 2017).

41. See, for example, *Training Humans*, curated by Kate Crawford and Paglen at the Osservatorio of the Fondazione Prada in Milan in 2019–2020, https://www.fondazioneprada.org/project/training-humans/?lang=en.

42. Trevor Paglen, "Invisible Images (Your Pictures Are Looking at You)," *New Inquiry*, 8 December 2016, https://thenewinquiry.com/invisible-images-your-pictures-are-looking-at-you/. "Unlearning to see like humans" is the goal Paglen pursued in a series of works realized from 2017 to 2019. Large installations such as *From Apple to Kleptomaniac (Pictures and Words)* (2019) and *From "Apple" to "Anomaly" (Pictures and Labels)* (2019) visualize in space the connections between words and images within the ImageNet dataset. *Machine-Readable Hito* (2017) and *"Fanon" (Even the Dead Are Not Safe)* (2017), part of an exhibition titled *A Study of Invisible Images* (2017), visualize the way in which face and emotion recognition systems work, extending their reach also to the field of historical images. The fact that "contemporary perception is machinic to large degrees" because "information is passed on as a set of signals that cannot be picked up by human senses" is underlined also by Hito Steyerl in "A Sea of Data: Apophenia and Pattern (Mis-)Recognition," *e-flux*, #72, April 2016, https://www.e-flux.com/journal/72/60480/a-sea-of-data-apophenia-and-pattern-mis-recognition/.

43. The soundtrack of the video, realized by Holly Herndon, uses samples of noises and voices created to teach machine listening systems to recognize speech and other acoustic phenomena.

44. Faced with the implications, limits, and possibilities of machine-vision technologies, other artists and groups have opted for different strategies; for example, that of seizing machine-vision technologies to retrain them through new datasets and redirect them toward new targets. See, for example, Forensic Architecture's *Triple Chaser* (2019), https://forensic-architecture.org/investigation/triple-chaser; and Paolo Cirio's *Capture* (2019), https://paolocirio.net/work/capture/.

45. For a discussion of the possibility of considering these images as part of the longer history of "composite images," see "Disréalismes: Une conversation entre Grégory Chatonsky, Christian Joschke, et Antonio Somaini," in "Images composites," special issue, *Transbordeur*, no. 7 (2023): 98–109.

46. Alexander Mordvintsev, Christopher Olah, and Mike Tyka, "Inceptionism: Going Deeper into Neural Networks," *Google Research Blog*, last updated 13 July 2015, https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html; "DeepDream—A Code Example for Visualizing Neural Networks," *Google Research Blog*, 1 July 2015, https://ai.googleblog.com/2015/07/deepdream-code-example-for-visualizing.html; and Christian Szegedy et al., "Going Deeper with Convolutions" (2015), arXiv, https://arxiv.org/abs/1409.484.

47. Min Lin, Qiang Chen, and Shuicheng Yan, "Network in Network" (2014), arXiv, https://arxiv.org/abs/1312.4400.

[48] After presenting the DeepDream algorithm, Alexander Mordvintsev and Mike Tyka began to use it to generate a series of works that can be accessed through their websites. For Alexander Mordvintsev, see https://znah.net/. For Mike Tyka, see https://www.miketyka.com/.

[49] The term *technical meta-pictures*, which refers to W.J.T. Mitchell's notion of "meta-picture," is used in Offert and Bell, "Perceptual Bias and Technical Metapictures." See also W.J.T. Mitchell, *Picture Theory* (Chicago: University of Chicago Press, 1994), 35–82.

[50] See Steyerl, "A Sea of Data."

[51] See, for example, Hubert Damisch, *A Theory of /Cloud/: Toward a History of Painting*, trans. Janet Lloyd (Stanford, CA: Stanford University Press, 2002); and Dario Gamboni, *Potential Images: Ambiguity and Indeterminacy in Modern Art* (London: Reaktion Books, 2004). The question of "images hidden within images" was tackled by the exhibition *Une image peut en cacher une autre: Arcimboldo, Dalí, Raetz* (An image can hide another: Arcimboldo, Dalí, Raetz) curated by Jean-Hubert Martin and Dario Gamboni at the Grand Palais in Paris in 2009. (An exhibition catalogue with the same name was published by Réunion des musées nationaux in 2009.)

[52] For the work of Refik Anadol, see https://refikanadol.com/.

[53] 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.

[54] Michael Castelle, "The Social Lives of Generative Adversarial Networks," in *FAT* '20: Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2020), https://doi.org/10.1145/3351095.3373156.

[55] I thank Emilie K. Sunde for our conversations about this point. She discusses the idea of all the images generated by a latent space as a possible "atlas" of such space in an unpublished conference paper titled "Latent Reality: The Shifting Ground-Truth of Photography" (presented at Expanded Visualities: Photography and Emerging Technologies, 6th International Conference of Photography and Theory, 17–19 November 2022, Nicosia, Cyprus).

[56] On deepfakes, see Graham Meikle, *Deepfakes* (London: Polity, 2022).

[57] The project *This Person Does Not Exist* consisted of a website created in 2019 by Philip Wang using StyleGAN, a GAN introduced by Nvidia researchers in 2018–2019. Each time the page is refreshed, the algorithm visualizes a hyperrealistic portrait of a nonexistent person. See https://thispersondoesnotexist.com/. The project *DoppelGANger.agency* can be accessed through the website http://doppelganger.agency:3000/.

[58] On the impact of AI on photography, see Milo Keller, Claus Gunti, and Florian Amoser, eds., *Automated Photography* (Lausanne: ECAL/University of Art and Design Lausanne; Mörel Books, 2021).

[59] See Mario Klingemann, "Photography through the Eyes of a Machine," interview within the framework of the 2017 EyeEm Photography Festival, https://www.eyeem.com/blog/mario-klingemann-ai-art; and "Mario Klingemann: Neurophotography," The Photographers' Gallery, video, 21 March 2018, 5:37, www.youtube.com/watch?v=iJl-pM3FzSw.

[60] See the transcription of a conversation between Anadol, Casey Reas, Michelle Kuo, and Paola Antonelli: "Modern Dream: How Refik Anadol Is Using Machine Learning and NFTs to Interpret MoMA's Collection," MoMA, 15 November 2021, https://www.moma.org/magazine/articles/658.

[61] For a broader view of artists using GAN and other deep-learning algorithms, see Arthur I. Miller, *The Artist in the Machine: The World of AI-Powered Creativity* (Cambridge: MIT Press, 2020); Joanna Zylinska, *AI Art: Machine Visions and Warped Dreams* (London: Open Humanities Press, 2020); Ruggero Eugeni, *Capitale algoritmico: Cinque dispositivi*

postmediali (più uno), ch. 5; and Lev Manovich and Emanuele Arielli, *Artificial Aesthetics: A Critical Guide to AI, Media and Design* (2021–2023), http://manovich.net/index.php/projects/artificial-aesthetics. Among the artists working with GANs are Anadol, Nora Al-Badri, Klingemann, Egor Kraft, Christopher Kalendran Thomas.

62. Other examples include *Comet (Corpus: Omens and Portents*), *Venus Flytrap (Corpus: American Predators*), and *Vampire (Corpus: Monsters of Capitalism)*. On the series *Adversarially Evolved Hallucinations*, see Luke Skrebowski, "Resistance at a Moment of Danger: On Trevor Paglen's Recent Work," in John P. Jacob and Luke Skrebowski, *Trevor Paglen: Sites Unseen* (Washington, DC: Smithsonian American Art Museum, 2018), 128–186; and Lila Lee-Morrison, *Portraits of Automated Facial Recognition: On Machinic Ways of Seeing the Face* (Bielefeld: Transcript, 2019), ch. 8.

63. On Chatonsky's installation *Second Earth* (2019), see https://chatonsky.net/earth/. On his formulation of the concept of "artificial imagination" in his works and writings, see http://chatonsky.net/category/journal/ima/.

64. On Chatonsky's notion of "disréalisme," see Grégory Chatonsky, "*Complétion 1.0*: Tout a lieu deux fois (ou presque)," in *L'image à l'épreuve des machines*, ed. Ada Ackerman, Alice Leroy, and Antonio Somaini (Dijon: Les Presses du Réel, forthcoming). See also http://chatonsky.net/disrealisme/.

65. For this work, Steyerl collaborated with two programmers, Damien Henry and Jules LaPlace. The first is the author of a video titled *A Train Window* (2017) in which a next-frame prediction algorithm called pix2pix—part of a specific class of GANs called conditional generative adversarial networks and used for image-to-image translation—was trained to predict the next frame of a video showing a moving landscape seen from a train window. The video is available at https://magenta.tensorflow.org/nfp_p2p. I thank Steyerl, Henry, and LaPlace (through an interview with Quentin Emery) for the information they gave me about this work.

66. 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).

67. On images produced by text-to-image models, see "Generative Imagery: Towards a 'New Paradigm' of Machine Learning-Based Image Production," special issue, ed. Lukas R.A. Wilde, Marcel Lemmes, and Klaus Sachs-Hombach, *IMAGE: The Journal of Interdisciplinary Image Science*, no. 37, pt. 1 (2023).

68. Among the various websites and tutorials explaining the functioning of diffusion models such as Stable Diffusion, see Andrew [Wong], "How Does Stable Diffusion Work?," Stable Diffusion Art, last updated 13 June 2023, https://stable-diffusion-art.com/how-stable-diffusion-work/. The diagrams presented in this section of my article are modeled on the ones from this tutorial. On DALL-E 2, see Aditya Ramesh, "How DALL-E 2 Works," http://adityaramesh.com/posts/dalle2/dalle2.html.

69. Stable Diffusion version 2 is trained instead on images with a resolution of 768 by 768 pixels.

70. To achieve this, another neural network, called a "variational autoencoder," compresses the initial 512 by 512 images of Stable Diffusion version 1 into a latent space that is forty-eight times smaller than the initial image space and then brings it back to its initial resolution. This also happens in two phases: first, an encoder compresses the image from its initial 512 by 512 resolution into the lower-dimensional representation in the latent space; then, a decoder restores the image, bringing it back to the 512 by 512 resolution. If the final desired output is an image in a resolution higher than 512 by 512, one needs to use another

algorithm explicitly conceived for upscaling; for example, Enhanced Super-Resolution Generative Adversarial Networks.

71. See Anthony Masure, *Design sous artifice: La création au risque du machine learning* (Geneva: HEAD Publishing, 2023).

72. *Beat Biden*, GOP [Republican National Committee], 25 April 2023, video, 0:32, https://www.youtube.com/watch?v=kLMMxgtxQ1Y.

73. The images have since been removed from Amnesty International's website. See Luke Taylor, "Amnesty International Criticised for Using AI-Generated Images," *Guardian*, 2 May 2023, https://www.theguardian.com/world/2023/may/02/amnesty-international-ai-generated-images-criticism.

74. The project, developed by Maurice Blackburn Lawyers, can be accessed at "Exhibit A-i: The Refugee Account," https://www.exhibitai.com.au/home.

75. On LAION-5B, see Romain Beaumont, "LAION-5B: A New Era of Open Large-Scale Multi-modal Datasets," LAION, 31 March 2022, https://laion.ai/blog/laion-5b/.

76. Yuk Hui discusses the possibility of considering image html metadata as an example of a digital object's "milieu." See Hui, *On the Existence of Digital Objects*, 47–73.

77. The initiative, launched by Holly Herndon and Mat Dryhust, can be accessed at https://haveibeentrained.com/.

78. Christoph Schuhmann, "LAION-Aesthetics," LAION, 16 August 2022, https://laion.ai/blog/laion-aesthetics/.

79. The information is provided in various sections of the LAION-5B website, https://laion.ai/.

80. See Hito Steyerl, "Mean Images," *New Left Review*, n.s., no. 140/141 (March/June 2023), https://newleftreview.org/issues/ii140/articles/hito-steyerl-mean-images. Jens Schröter describes generative imagery as "stochastic," "statistic," or "probabilistic"; see Jens Schröter, "The AI Image, the Dream, and the Statistical Unconscious," in "Generative Imagery," *IMAGE*, 112–120.

81. For Chatonsky's project *His Story* (2022), see http://chatonsky.net/his-story/. *La machine 100 têtes* (*The Hundred Headless Machine*) (2022) exists in multiple versions, one of which was just published as a book. Grégory Chatonsky, *La machine 100 têtes* (Paris: Rrose, 2023).

82. "The Kiss," Gregory Chatonsky, November 2022, http://chatonsky.net/kiss-4/.

83. "Contrefaits/Counterfeits," Gregory Chatonsky, May 2021, https://chatonsky.net/counterfeits/

84. See Alain Besançon, *The Forbidden Image: An Intellectual History of Iconoclasm* (Chicago: University of Chicago Press, 2001). On more recent forms of image censorship, see Katja Müller-Helle, *Bildzensur: Infrastrukturen der Löschung* (Berlin: Wagenbach, 2022); Katja Müller-Helle, ed., *Bildzensur: Löschung technischer Bilder*, volume 16 of *Bildwelten des Wissens* (Berlin: De Gruyter, 2020).

85. I am referring, for example, to the tradition of German *Bildwissenschaft*, as theorized by art historians and philosophers such as Gottfried Boehm, and to the debate, during the 1990s and at the beginning of the 2000s, about concepts such as the "iconic turn" and the "pictorial turn." On the "iconic turn" (*ikonische Wende*) and the idea of an autonomous, non-text-based "logic of images" (*Logik der Bilder*), see Gottfried Boehm, "Die Wiederkehr der Bilder," in *Was ist ein Bild?*, ed. Gottfried Boehm (Munich: Fink, 1994), 11–38; Gottfried Boehm, "Jenseits der Sprache? Anmerkungen zur Logik der Bilder," in *Iconic Turn: Die neue Macht der Bilder*, ed. Christa Maar and Hubert Burda (Cologne: DuMont, 2004); and Gottfried Boehm, *Wie Bilder Sinn erzeugen—Die Macht des Zeigens* (Berlin: Berlin University Press, 2007). On the "pictorial turn," see Mitchell, *Picture Theory*, 11–34.