UNIVERSITY of CALIFORNIA PRESS | Media+Environment

Offert, Fabian. 2021. "Latent Deep Space: Generative Adversarial Networks (GANs) in the Sciences." *Media+Environment* 3 (2). https://doi.org/10.1525/001c.29905.

MODELING THE PACIFIC OCEAN

Latent Deep Space: Generative Adversarial Networks (GANs) in the Sciences

Fabian Offert[1] [a]

[1] Department of Germanic and Slavic Studies, University of California, Santa Barbara, United States

Keywords: digital images, artificial intelligence, machine learning, generative adversarial networks

https://doi.org/10.1525/001c.29905

The recent spectacular success of machine learning in the sciences points to the emergence of a new artificial intelligence trading zone. The epistemological implications of this trading zone, however, have so far not been studied in depth. Critical research on machine learning systems, in media studies, visual studies, and "critical AI studies," in the past five years, has focused almost exclusively on the social use of machine learning, producing an almost insurmountable backlog of deeply flawed technical reality. Among this backlog, one machine learning technique warrants particular attention from the perspective of media studies and visual studies: the generative adversarial network (GAN), a type of deep convolutional neural network that operates primarily on image data. In this paper, I argue that GANs are not only technically but also epistemically opaque systems: where GANs seem to enhance our view of an object under investigation, they actually present us with a technically and historically predetermined space of visual possibilities. I discuss this hypothesis in relation to established theories of images in the sciences and recent applications of GANs to problems in astronomy and medicine. I conclude by proposing that contemporary artistic uses of GANs point to their true potential as engines of scientific speculation.

Introduction

> We must have images because only images can teach us. Only pictures can develop within us the intuition needed to proceed further towards abstraction. [...] And yet: we cannot have images because images deceive. Pictures create artifactual expectations, they incline us to reason on false premises. We are human, and as such are easily led astray by the siren call of material specificity. (Galison 2006, 236)

[a] offert@ucsb.edu

Fabian Offert is Assistant Professor for the History and Theory of the Digital Humanities in the Department of Germanic and Slavic Studies at the University of California, Santa Barbara. His research and teaching focuses on the visual digital humanities, with a special interest in the epistemology and aesthetics of computer vision and machine learning.

At UCSB, he is affiliated with the Media Arts and Technology program, the Comparative Literature program, and the Center for Responsible Machine Learning. He is also principal investigator of the UCHRI multi campus research group "Critical Machine Learning Studies" (2021-23), and the international research project "AI Forensics" (2022-25), funded by the VW foundation.

His most recent articles have appeared in journals like AI & Society, Media and Environment, and the Journal of Art Historiography, as well as in conferences like DH, DHd, CHR and NeurIPS. He has given invited talks at Berkeley, UNSW Sydney, transmediale, Gaîté Lyrique Paris, HU Berlin, EHESS Paris, the Hebrew University of Jerusalem, and many other institutions. He is also founding editor of Construction Kit: A Review Journal for Research Software and Data Services in the Humanities.

Before joining the faculty at UCSB, he served as postdoctoral researcher in the DFG SPP "The Digital Image", associated researcher in the Critical Artificial Intelligence Group (KIM) at Karlsruhe University of Arts and Design, and Assistant Curator at ZKM Karlsruhe, Germany.

Website: https://zentralwerkstatt.org

In late October 2020, in the midst of a global pandemic and political turmoil, the Turing Institute, Great Britain's most important computer science research institute, announced a new project with the remarkable title "AI for Scientific Discovery: Developing Artificial Intelligence Systems Capable of Nobel-Quality Discoveries by 2050."[1] As with all things artificial intelligence, it is important to consider the economic and political interests behind such a claim, particularly when it reaches this far into an uncertain future. But while automated "Nobel-quality" research may remain unattainable, machine learning for science has come a long way since the beginning of the current "AI summer" around 2012. Most recently, a machine learning system seems to have "solved" protein folding in biology, at least within the constraints of the CASP assessment (Senior et al. 2020). Moreover, researchers in physics have started to utilize machine learning systems to (re)discover physical laws (Udrescu and Tegmark 2020b; Iten et al. 2020) or to directly derive symbolic representations from observations (Greydanus, Dzamba, and Yosinski 2019; Udrescu and Tegmark 2020a; Cranmer et al. 2020): "from pixels to physics."[2] Others claim nothing less than to have found a viable way of simulating the quantum foundations of matter itself, with the help of machine learning (Pfau et al. 2020).

Surprisingly, the epistemological implications of this new trading zone (Galison 2011) in the making have so far not been studied in depth. In fact, critical research on artificial intelligence systems, in media studies, visual studies, and "critical AI studies," in the past five years, has focused almost exclusively on either the history of AI (for instance, its relation to cybernetics) or on its contemporary sociopolitical use.[3] And while the significant individual and societal harm of applications like facial recognition and predictive policing has indeed warranted this kind of attention, this narrow focus of the "critical disciplines" has produced an almost insurmountable backlog of technical reality.

In this paper, I argue that, among the artificial intelligence systems that make up this backlog, one warrants particular attention from the perspective of media studies and visual studies: the generative adversarial network (GAN), a type of deep convolutional neural network that operates primarily on image data. While GANs have been studied in the context of so-called "deep fakes"

[1] See https://www.turing.ac.uk/research/research-projects/turing-ai-scientist-grand-challenge. Two examples of similar albeit more modestly designed projects are the new NSF AI Institute for Artificial Intelligence and Fundamental Interactions (https://iaifi.org/) and the "cluster of excellence" "Machine Learning: New Perspectives for Science" at Tübingen University (https://uni-tuebingen.de/en/research/core-research/cluster-of-excellence-machine-learning/home).

[2] Both of these papers were recently presented at a workshop at Emory University, aptly named "Can Machine Learning Learn New Physics, or Do We Need to Put It In by Hand?" A recording is available here: https://www.youtube.com/watch?v=DRh1OlGlRxo. See also Raghu and Schmidt (2020) for an overview of machine learning approaches to "scientific discovery."

[3] One exception is the nascent VW-funded research project "How Is AI Changing Science?" See https://howisaichangingscience.eu/.

(so, again, as an explicitly political technique) and as a creative tool, their distinctive role in the sciences—in astronomy, medicine, chemistry, and biology, among others—has gone largely unnoticed.

The paper's main hypothesis is simple: GANs, as specific technical objects, inevitably produce an epistemic opacity that goes beyond the well-known general technical opacity of artificial intelligence systems (Pasquale 2015). In the sciences, GANs "pass" as optical media. But while they seem to enhance our view of an object under investigation—be it galaxies, cancer cells, or brain waves—they actually present us with a technically and historically predetermined space of visual possibilities: what there is to know is what is already known. The epistemic thing falls back to, and is completely determined by, the technical object (Rheinberger 1997), and technical legacy determines epistemic faculty.

In the following pages, I present a close reading of GANs and GAN-based techniques in the sciences to validate this hypothesis. Specifically, I explore the role of GANs in astronomical and medical imaging, including the GAN-based enhancement of images of galaxies, the use of GANs to translate between different types of MRI image formats, and the application of GANs to the visual interpretation of brain waves. I conclude by proposing that contemporary artistic uses of GANs point to their true potential as engines of scientific speculation.

The Epistemic Oscillation of Scientific Images

Because of their complicated epistemic status, scientific images have long been a core concern of the history of science and science and technology studies. Those same "technical" (Bredekamp, Schneider, and Dünkel 2012), "systemic" (Hinterwaldner 2017), or "operational" (Pantenburg 2016) images present an ongoing challenge to visual studies and art history.

As Peter Galison reminds us in the above epigraph, we desperately need images to make sense of the world. At the same time, we fall so easily for their "siren call of material specificity" (Galison 2006, 300). As objects in the world, their simple presence often obfuscates their inadequacy as vehicles of representation. We should not rely on them too much or hope they will speak the truth. At the same time, without images, we "cannot proceed further towards abstraction," simply because we cannot think in purely symbolic operations. "By mimicking nature, an image, even if not in every respect, captures a richness of relations in a way that a logical train of propositions never can. Pictures are not just scaffolding, they are gleaming edifices of truth itself that we hope to reveal" (Galison 2006, 300). This dialectic of the image, then, materializes as a "battle between iconoclasm and iconophilia" in twentieth-century science.

Examining this same dialectic, Bruno Latour argues that images, specifically diagrams, can be understood as elements in a "chain of reference" that ensures the legibility of the natural world by allowing us to move freely back and forth between the material and the symbolic. As such, they are artificially created and then, nevertheless, serve as “raw data” again.[4] They are released from their mimetic duties and become objects of empirical study as if they had naturally emerged from the concrete, material object under investigation. As Latour writes, the diagram

> is not realistic; it does not resemble anything. It does more than resemble. It takes the place of the original situation [...]. Yet we cannot divorce this diagram from this series of transformations. In isolation, it would have no further meaning. It replaces without replacing anything. It summarizes without being able to substitute completely for what it has gathered. (Latour 1999, 67)

In other words, in both Galison’s and Latour’s readings, the diagram—and the scientific image in general—“oscillates”; it is unstable and subject to forces that pull it in either direction. It “wants”—to invoke W. J. T. Mitchell (2005)—to be something it is not: the concrete, material object under investigation (Latour) or its abstract, symbolic description (Galison).

This epistemic oscillation generally concerns both analog and digital images, but for Galison, the computer has rendered the “flickering exchanges” between the abstract and the pictorial more pronounced. The frequency of the oscillation is increased, as computation enables images to easily “scatter into data” and data to “gather into images”: neither “the ‘pictorial-representative’ nor the ‘analytical-logical’ exist as fixed positions. Instead, across a wide span of the sciences, we see that the image itself is constantly in the process of fragmenting and re-configuring. [...] No longer set in motion only in moments of crisis, we find that ordinary, every-day science propels this incessant oscillation” (Galison 2006, 322).

But no matter how “frictionless” computation renders the epistemic oscillation, we have to keep in mind, exactly with Latour, that every phase change requires, in image terms, a lossy conversion, a “violent representation” of reality, as Claus Pias (2003) calls it. The gaps between the abstract and the pictorial—as with those between the material and the pictorial—cannot be bridged; they can just be “jumped.” And every such jump requires human guidance and interpretation. Every time images scatter into data, or data gather into images, a thousand human decisions come into play, no matter the direction of the jump.

Importantly, the image itself conceals these human decisions. They can be recovered only through forensic work, if at all. In other words, a fundamental epistemic opacity, a readiness to deflect how its knowledge is produced, already

[4] This point is later emphasized in Gitelman (2013), who understands it as a foundational fallacy of data visualization.

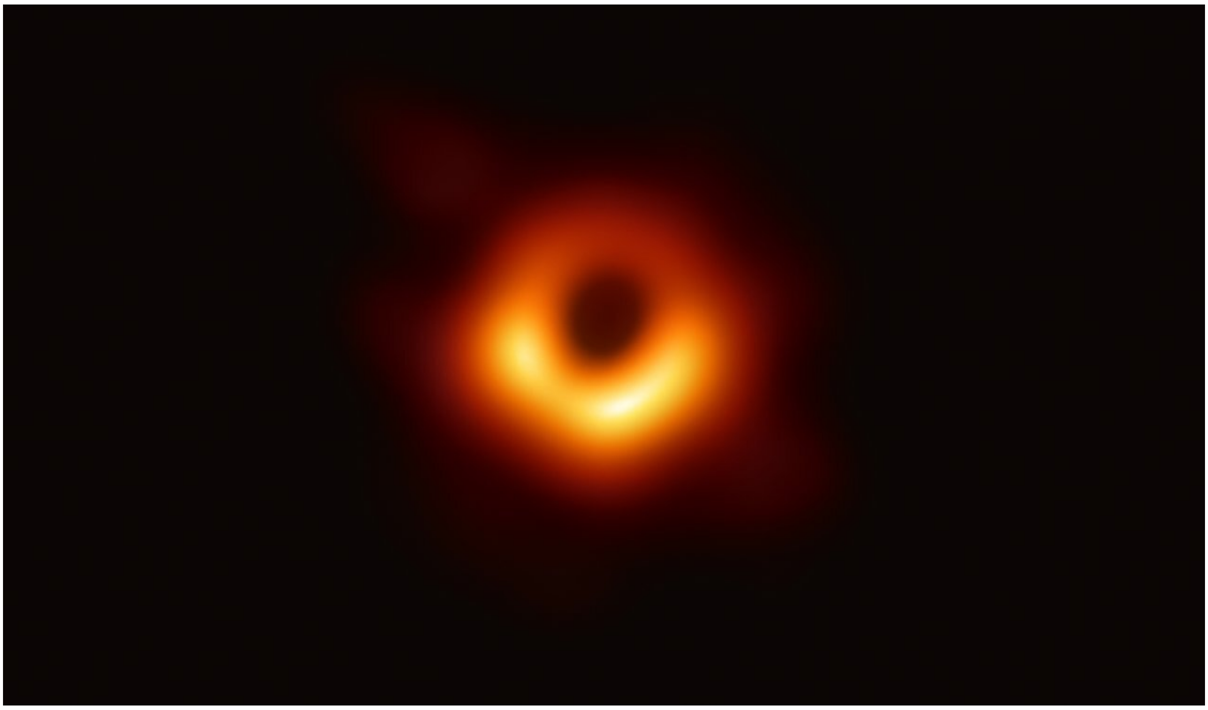


Figure 1: Black hole in the center of the Messier 87 galaxy

Credit: Event Horizon Telescope Collaboration 2019 (CC).

germinating in every image by virtue of its apparent concreteness, is brought to life where the frequency of the image's epistemic oscillation increases through computation.

Images at the Limits of Perception

The famous "black hole image" from 2019 is a good example of this dilemma. The image's only raison d'être is to give evidence for the symbolic equations that describe the phenomenon it represents. It exists on the threshold to the nonimage, as a last glimpse of what is pictorially possible, or, as Orit Halpern has put it, it "represents the figure of the terminal limits of human perception" (Halpern 2021, 229). Interestingly, Galison, at the end of "Images Scatter into Data, Data Gather into Images," chooses the black hole as a metaphor to describe this exact threshold:

> General relativity gives a fascinating description of an object falling into a black hole. As the object approaches the event's horizon—the point of no return—an outside observer sees that object slow down as it approaches the horizon, its image gradually shifting towards the red. Eventually the scene of the falling object freezes in dimming redness at just the instant it passes beyond the visible. That scene resembles ours. Just when the scientific image moves towards abstraction we are left with the last glimpse of a frozen picture and ignore what happens next. (Galison 2006, 323)

The black hole image can exist only because computation enables "data to gather into images," the actual "photographic" process involving terabytes of data being rearranged (Event Horizon Telescope Collaboration 2019). Even without that knowledge, and without any intuition for the physical implications of an outrageous astronomical phenomenon like a black hole, however, it is immediately obvious that the image cannot be a "normal" analog photo (some equivalent of photons hitting a photon-sensitive medium), that it, instead, has to involve some advanced computational processes. And yet the image "wants" to be treated as if it were the result of photons hitting a photon-sensitive medium. It does not give away its constructed nature, at least not within the image space. To not lie, it relies on contextualization.

The epistemic opacity at play here becomes particularly obvious if we consider that we can easily approach the task of creating an image of a black hole from the other side of the material-to-symbolic continuum—for instance, with a physically based rendering system that allows the precise computation of light scattering for arbitrary objects. While, as Jacob Gaboury (2015) has pointed out, even physically based rendering has to make significant compromises to arrive at realistic results, one thing is obvious: between such two images, intuitively, there would be no way to tell which is which. Not because of sophisticated photo manipulation or 3D rendering techniques but because both images would represent a phenomenon at the threshold of what can be depicted in the first place, where, quite literally, on the event horizon, the light that would give birth to an image is stopped in its tracks. To quote Halpern again, the borderlands of perception require a "turn to automation and big data as modes of managing extreme uncertainty" (Halpern 2021, 232). It is here where GANs start to become relevant.

Generative Adversarial Networks

Generative adversarial networks[5] leverage game theory (Goodfellow et al. 2014) to approximate the probability distribution that defines a set of images by means of a minimax game between two deep convolutional neural networks (LeCun et al. 1989; Krizhevsky, Sutskever, and Hinton 2012; LeCun, Bengio, and Hinton 2015). Effectively, GANs define a continuous noise distribution p_z which is mapped to a discrete data space (we could also say "image space") via $G(z)$ where G is a "generator," an "inverted" convolutional neural network that "expands" an input variable into an image, rather than "compressing" an image into a classification probability. G is trained in conjunction with a "discriminator," a second deep convolutional neural network D that outputs a single scalar $D(x)$ which represents the probability that x came from the data rather than from G.

[5] Théo Lepage-Richer has analyzed the general importance of the notion of adversariality for the history of artificial intelligence in Lepage-Richer (2021).

Less technically put: the generator G learns to transform any high-dimensional (e.g., 512-dimensional) latent vector z into an image, while the discriminator learns to distinguish such artificially created images from a set of "real" images. With Galison: data (z) gather into images $(G(z))$, and images scatter into data $(D(G(z)))$, at every iteration of the GAN training process. At every iteration, an image is created and destroyed. The epistemic oscillation of digital images is thus the defining feature of generative adversarial networks. Also note that the system effectively learns a lossy compression:[6] a high-dimensional data space with dimensions $> z$—for instance, a set of images—is compressed to be reproducible from a latent space with dimensions z. Another aspect of the Galison-Latour model is thus operationalized in the technique.

The original paper by Goodfellow (Goodfellow et al. 2014) demonstrates the potential of GANs by using them to synthesize new handwritten digits from the MNIST dataset. The MNIST dataset, however, has a resolution of 28 × 28 pixels—that is, several orders of magnitude below standard photo resolutions—and scaling up the approach proved difficult. While a lot of effort was made, and a lot of "compute" was spent, to go beyond marginal resolutions, progress was slow (for machine learning) until very recently, when StyleGAN (Karras, Laine, and Aila 2018), a generative adversarial network that implemented several significant optimization tricks to mitigate some of the inherent architectural limitations, was introduced. Current-generation models like StyleGAN2 (Karras et al. 2019), which presents another improvement over the original StyleGAN, are now able to produce extremely realistic samples from large image corpora.

But what kind of problem could such a system potentially solve? What kind of scientific application does a system have that has learned to "imitate" a certain kind of image, or, more precisely, that has learned the defining features of a certain set of images and is able to construct new samples from this information? As it turns out, these exact properties come in handy in the approximation of so-called inverse problems. Inverse problems are a broad class of problems in the sciences, where causal factors are to be reconstructed from a limited number of observations. When it comes to images, inverse problems often imply the reconstruction of an original image from a version that has been perturbed by noise (G. Wang et al. 2018; Z. Wang, Chen, and Hoi 2019). The inverse problem is the reconstruction of the noise function (i.e., of the exact signal that has altered the image)—the reconstruction of the original image, then, is trivial. GANs can facilitate such a reconstruction because they are generative classifiers (Ng and Jordan 2002): in theory, they can learn "an explicit low-dimensional manifold" for every "natural signal class," based on multiple samples (Asim et al. 2020).

[6] GANs and related generative techniques like variational autoencoders have also been proposed as a solution for creating efficient compression algorithms—for example, in Cao, Wu, and Krähenbühl (2020) and Mentzer et al. (2020).

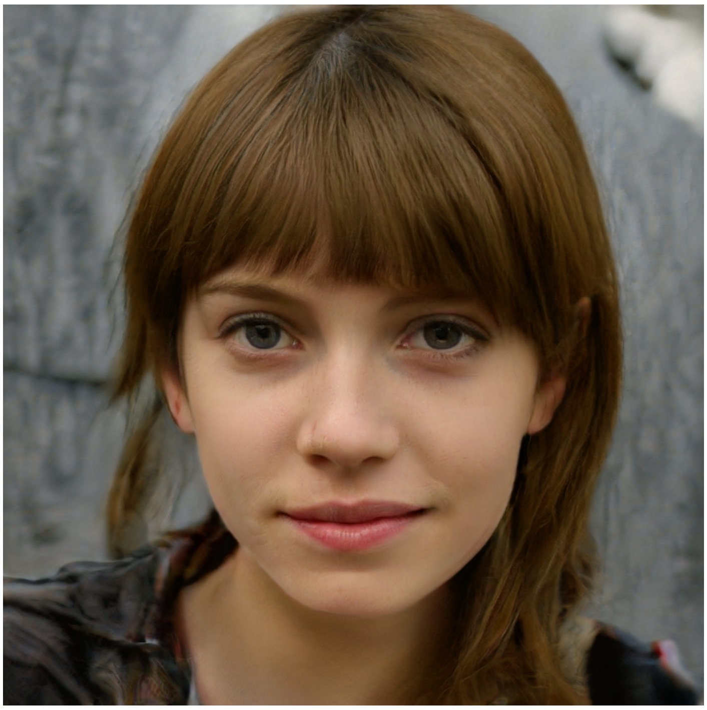


Figure 2: Image of an imaginary person, produced with the StyleGAN architecture

Going beyond the Deconvolution Limit with Pix2Pix

Inverse problems commonly emerge where science deals with objects that need to be observed but that are somehow "out of reach." In astronomy, which is concerned with objects that are often millions of light-years away, trying to obtain a "clear view" commonly becomes an inverse problem. Importantly, it becomes an inverse problem with solutions that are impossible to verify. To further complicate things, with the invention of more and more elaborate optical media, the most important inverse problem in astronomy has become the reconstruction of images that have been perturbed by noise *produced by the very instruments that facilitate the imaging*. Advanced telescopes enable the observation of previously unknown astronomical objects. As optical media, they shift the border of visibility (Kittler 2010). The images they produce, however, are subject to specific noise introduced by advanced telescopes only. We thus find ourselves in an interesting situation: observation and perturbation come from the same source and are often hopelessly entangled. Disentangling signal and noise thus becomes a difficult, nonlinear exercise: an area where artificial neural networks usually excel.

An example is the "denoising" of images of galaxies. In a recent paper called "Generative Adversarial Networks Recover Features in Astrophysical Images of Galaxies beyond the Deconvolution Limit," Schawinski et al. describe a GAN-based technique to reconstruct images of galaxies that have been perturbed by "various sources of random and systematic noise from the sky background, the optical system of the telescope and the detector used to record the data"

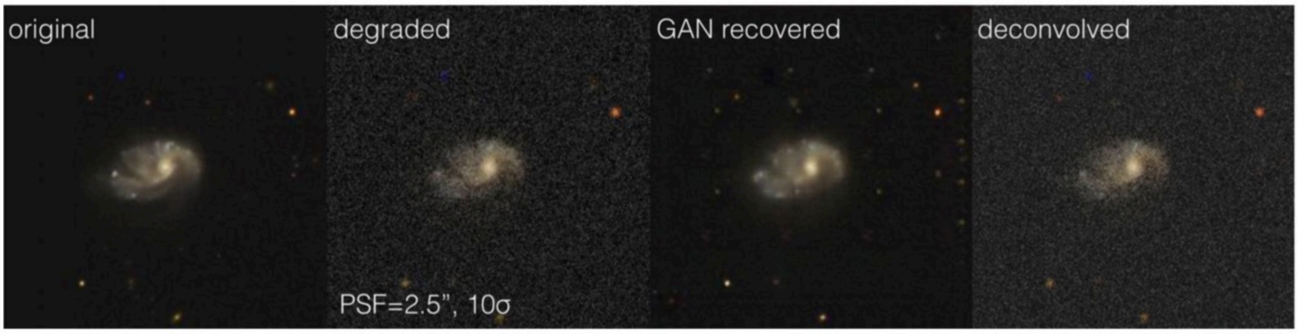


Figure 3: Denoising of galaxy images

Figure from Schawinski et al. (2017).

(Schawinski et al. 2017, 110). The reconstructions facilitated by the technique, also called "GalaxyGAN," transcend the deconvolution limit, according to the authors.

The deconvolution limit sets a lower bound for all kinds of digital reconstructive techniques (not only in the image domain), based on the Shannon-Nyquist sampling theorem (Shannon 1949)—that is, a limit of what is "naturally" reconstructible from a limited amount of information. The deconvolution limit is a "hard" limit. Hence, transcending this limit has, for decades, been a topos in science fiction. The most prominent example is, of course, *Blade Runner*, where Harrison Ford gives voice commands to a computer system to almost infinitely "enhance" an image. From there on, transcending the deconvolution limit has become somewhat of an internet meme, with shows like *NCIS* intentionally exploiting the ridiculousness of infinite "enhancement." Thus, it seems like there must be some sort of catch.

To go beyond the deconvolution limit, Schawinski et al. propose to utilize a GAN that has learned a mapping from artificially degraded, noisy images to undistorted original images. The authors utilize an established GAN-based architecture for paired image-to-image translation for this task, called Pix2Pix. Pix2Pix was first introduced by Isola et al. (2017) and was the first application of GANs to the task of image-to-image translation that went beyond style transfer—that is, introduced semantic aspects to the translation process. If Pix2Pix is trained on the artificially created training set described in the paper, the network learns a mapping from noisy to "clear" images. The authors provide a thorough evaluation of their results as well, which shows that both quantitatively (compared to previous methods with PSNR, or peak signal-to-noise ratio, as a measure) and qualitatively, the method works well.

And indeed, at the very end of the paper, the authors point to some important limitations: "In general, the GAN fails on rare objects that were absent or low in number in the training set, stressing that the performance of our method is strongly tied to the training set; it cannot reconstruct features it has not learned to recognize." Intuitively, this seems like an expected limitation. GANs cannot reproduce what they have not seen—that is, seen at least once in theory, and seen a significant number of times in practice. Moreover, because a GAN latent space is essentially a lossy compression of an image space, some features inevitably get lost in the training process and thus cannot be reconstructed by the GAN.

Here, suddenly, the inverse problem comes back to haunt the very mechanism devised to solve it. The defining feature of an inverse problem is the fact that information (about a signal) is missing, and it is impossible to know exactly *what* is missing. In the case of GANs, we cannot know *which* features are lost in the training process. There is no precipitate of unique artifacts, no list of special cases, no box of rejected samples. This means that essentially, one noise source has been replaced by another noise source. Other than the first noise source (the telescope), however, the perturbation introduced by the second noise source (the GAN) is entirely dependent on the training set fed to the mechanism. In other words, GANs are able to take the reconstruction process beyond the deconvolution limit only because they introduce epistemic priors, additional knowledge about the problem domain. These epistemic priors, then, define what "can be seen" with the GAN.

To be clear: this is a widely acknowledged problem when it comes to inverse problems,[7] and Schawinski et al. (2017) mention it specifically in their introduction:

> Deconvolution has long been known as an "ill-posed" inverse problem because there is often no unique solution if one follows the signal processing approach of backwards modelling. Another standard practice in tackling inverse problems like these is integrating priors using domain knowledge in forward modelling. [...] In this paper, we demonstrate a method using machine learning to automatically introduce such priors.

Similar disclaimers can be found in almost every paper that tackles "denoising" and the related problem of "super resolution." But—and this is the crucial difference—the epistemic priors introduced by this method are concealed in the GAN. This means that GANs fail silently. The solution space facilitated by a GAN is a "dense" solution space: there is always *a* solution, as the generator has learned a mapping from all possible inputs to all possible outputs. GANs always give coherent answers, even if the question is ill-posed. While "regular noise" is an obvious perturbation, when it is removed by a GAN it is "secretly" replaced by the epistemic priors inherent in its latent space. The fact that the image *remains* perturbed, only in a different way, is concealed. While it thus seems like the GAN provides the "clear image" we desire, what we get is an image that "pretends" to be clear while being fully dependent on a set of

[7] In Offert (2020), I describe the case of face super resolution, arguing that there is no (nonmalicious) real-world use case for this specific application of GANs.

epistemic priors. This is how GANs "pass" as media. But in reality, looking at galaxies with a GAN is looking into the GAN latent space exclusively. GAN vision is not augmented reality, it is virtual reality.

If GAN images were to serve as a means of discovery—for example, by facilitating the discovery of unknown properties of galaxies—we suddenly have to face the problem that the space of discovery in these images is exactly the latent space of the neural network that improved them: all that is potentially "unknown" about the galaxies in these images is modeled from existing data by the neural network. GANs, in a peculiar way, thus operationalize the epistemological distinction between invention and discovery by rendering the space of discovery a technically determined space. This determination is a *historical* determination: where GANs serve as a medium, what there is to know is what is already known.

GANs thus not only amplify the frequency of the epistemic oscillation of digital images in the sciences, but they also amplify their potential to conceal the human decisions involved in the high-frequency back-and-forth between image and data. While the black hole images conceal only the specific technical process necessary to produce them, GAN-facilitated images can and do conceal whole corpora of images, a complete history of all the attempts to depict what they represent. The GAN-facilitated image is a "summary image" in Mitchell's sense, a "piece of moveable cultural apparatus [...] that encapsulates an entire episteme, a theory of knowledge" (Mitchell 1995, 49).

Causing Cancer with CycleGAN

So far, the problematic implications of GANs seem to have emerged from the unavailability of the objects under investigation: we could argue that black holes, galaxies, and the like invite, even require, some degree of epistemic imagination to become addressable. But in fact, we do not have to venture into space to see these implications at work. In recent years, GANs have become "human-centered," to intentionally reframe the technical term, and have found their way into the medical field.

In a recent paper, "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation," Cohen, Luck, and Honari (2018) describe a system based on CycleGAN (Zhu et al. 2017), a GAN-based approach to unpaired image-to-image translation. Other than Pix2Pix, CycleGAN does not require pairs of images (e.g., original and degraded) but simply two datasets of images from the two translation domains. Translation is learned by adding an additional "circular" constraint to the training process—that is, a constraint that makes sure that A can be translated into B and vice versa. In medical

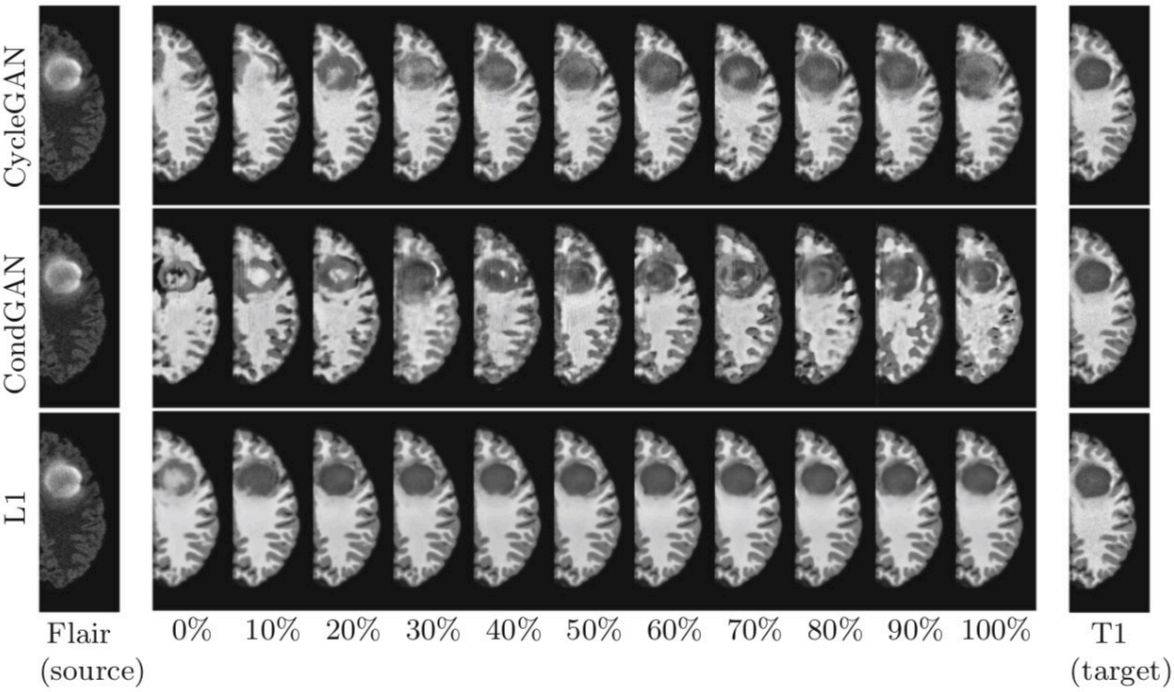


Figure 4: "Hallucinated" tumors from unbalanced training sets in image-to-image translation models

Figure from Cohen, Luck, and Honari (2018).

imaging, where image-making is complex and expensive, attempts have been made to employ such techniques to translate between different variations of imaging techniques—for example, so-called "flair" and "T1" MRI images.[8]

Cohen, Luck, and Honari now show that, if the training sets from both domains are unbalanced—for instance, if there are too many or too few images of cancerous tissue in one of the datasets—CycleGAN "hallucinates" features in the translated image. As the authors state, they "demonstrate the problem with a caricature example [...] where we cure cancer (in images) and cause cancer (in images) using a CycleGAN that translates between Flair and T1 MRI samples." Moreover, in a thorough analysis of the amount of difference needed to induce such "hallucinations," they show that the imbalance in the dataset does not have to be total (only cancerous tissue in one set and only healthy tissue in the other) but that there is a threshold where feature-different samples become irrelevant to the GAN.

Visualizing Imagination with GAN-Made Natural Image Priors

Importantly, the paper discussed above is one of very few examples dedicated to the limitations, rather than the seemingly infinite potential, of GAN-based techniques in medical imaging. In "Deep Image Reconstruction from Human Brain Activity," for instance, Shen et al. (2019) describe a technique to (literally) "make sense of" MRI data of humans looking at, or imagining,

[8] The speculative nature of MRI images in general has been discussed, for instance, in Joyce (2010) and Jonas and Kording (2017).

images of objects. The approach is introduced by the authors as "a method for visual image reconstruction from the brain that can reveal both seen and imagined contents" and is part of a whole line of recent research seeking to "decode" brain activity[9] into images or video (Le et al. 2021) with the help of deep neural networks.

In the paper, a deep neural network is trained on tuples of "image looked at" and corresponding MRI data—that is, the training set consists of images and the brain wave responses of test subjects in an MRI machine looking at these images. Once trained, the network produces images from unseen MRI data, providing a visual approximation of a test subject's imagination. Where do these images come from? They come from a GAN latent space. More technically, they are produced by utilizing a technique called feature visualization with natural image priors,[10] which in turn uses a GAN latent space as search space to find the closest visual match for a targeted image.

While the authors also run the experiment without using GAN images as natural image priors, interpretable images—images that show something in the literal sense of an identifiable object ("some thing")—depend on the use of a GAN. Without it, the images that can be reconstructed barely resemble the forms seen or imagined by the person in the MRI machine: "While the reconstructions obtained without the DGN [deep generator network, the GAN] also successfully reconstructed rough silhouettes of dominant objects, they did not show semantically meaningful appearance" (Shen et al. 2019, 5).

In other words, meaningfulness itself is necessarily restricted, again, to the latent space of the GAN. Moreover, the images seen by the study participants in the MRI machine come from the ImageNet dataset. This means that it is even "easier" for the GAN, which has also been trained on ImageNet, to produce "almost perfect" reconstructions, as it has been trained on the same dataset that the images it is reconstructing are taken from. As a pretrained generator from Dosovitskiy and Brox (2016) is reused instead of training a new generator on a subset of ImageNet omitting the test images, the generator necessarily "knows" the test images already, albeit only through their contribution to the learned probability distribution. Whatever the person in the MRI machine is thus imagining, it is expressed in terms of the legacy GAN.

GANs as Engines of Scientific Speculation

One could rightfully ask, then, if GANs can ever have a place in science at all—if their incredible potential to conceal is not fundamentally opposed to scientific principles. At the end of the day, all GAN images are "deep fakes,"

[9] The historical and epistemological entanglement of AI and neuroscience warrants a separate investigation that lies outside the scope of this paper; see, for instance, Bruder (2017, 2019).

[10] Feature visualization by itself is a highly speculative visual interpretability technique, as shown in Offert and Bell (2020b).

images designed to deceive.[11] One could argue, then, that the only real benefit of GANs is exactly their tendency to capture and distort, in interesting ways, what is already known. The fact that the black hole image has been acquired by the Museum of Modern Art, in New York, already speaks of this aesthetic potential. And more recently, contemporary artists working with artificial intelligence have explicitly turned to GAN images in the sciences[12] to generate speculative environments.

French artist Pierre Huyghe, for instance, utilized the methods described in Shen et al. (2019) to create images for his *UUmwelt* exhibition, originally shown at Serpentine Gallery and reexhibited in 2021 online under the title *Of Ideal*, facilitated by Hauser & Wirth. The images were created in collaboration with the Kamitani Lab at Kyoto University, where the paper by Shen et al. originated.[13] In the original exhibition space, the images are presented as video loops that capture the optimization process (the latent space search) of the GAN system, juxtaposed with one hundred thousand flies occupying the gallery space. Temperature, humidity, smell, and sound are also controlled elements in this hybrid digital-organic environment, mirroring previous works by Huyghe, like *After ALife Ahead* at Skulptur Projekte Münster 2017. The gallery becomes "a porous and contingent environment, housing different forms of cognition, emerging intelligence, biological reproduction and instinctual behaviors."[14] And it is through these organic elements that the exhibition emphasizes the distorted nature of the reconstructions shown and their inadequacy as representations of thought.

Another example is Tega Brain's 2019 installation *Asunder*. *Asunder* is a three-channel video installation that speculates on the potentiality of climate interventions. In the installation, a neural network is used that has learned to generate imaginary satellite images according to certain parameters. Real-life satellite images are changed by this network to suggest radical geoengineering interventions: removing the city of San Francisco, planting trees in the desert, rerouting rivers. The resulting composite image is then fed into a scientific climate model, which runs on a high-performance computer in the exhibition space, and the resulting changes in the global climate are displayed. The work is relevant exactly because it puts the peculiar epistemic implications of GANs on their feet: the speculative nature of inverse problems is embraced, and utilized to project the ridiculous exit strategies that humankind still has left. The literal "alienated" perspective of machine learning systems is transformed into productive "alienation" in the Brechtian sense.

[11] If we accept the inevitability of this deceptive quality, other potential mitigation strategies emerge, like the utilization of techniques from interpretable machine learning/explainable artificial intelligence. See Bau et al. (2018, 2019); Offert and Bell (2020a, 2020b).

[12] For an overview of the recent emergence of a separate field of "AI art," see Offert (2019, 2022).

[13] Pierre Huyghe in conversation with Hans-Ulrich Obrist, https://vip-hauserwirth.com/online-exhibitions/pierre-huyghe-of-ideal/.

[14] See https://vip-hauserwirth.com/online-exhibitions/pierre-huyghe-of-ideal/.

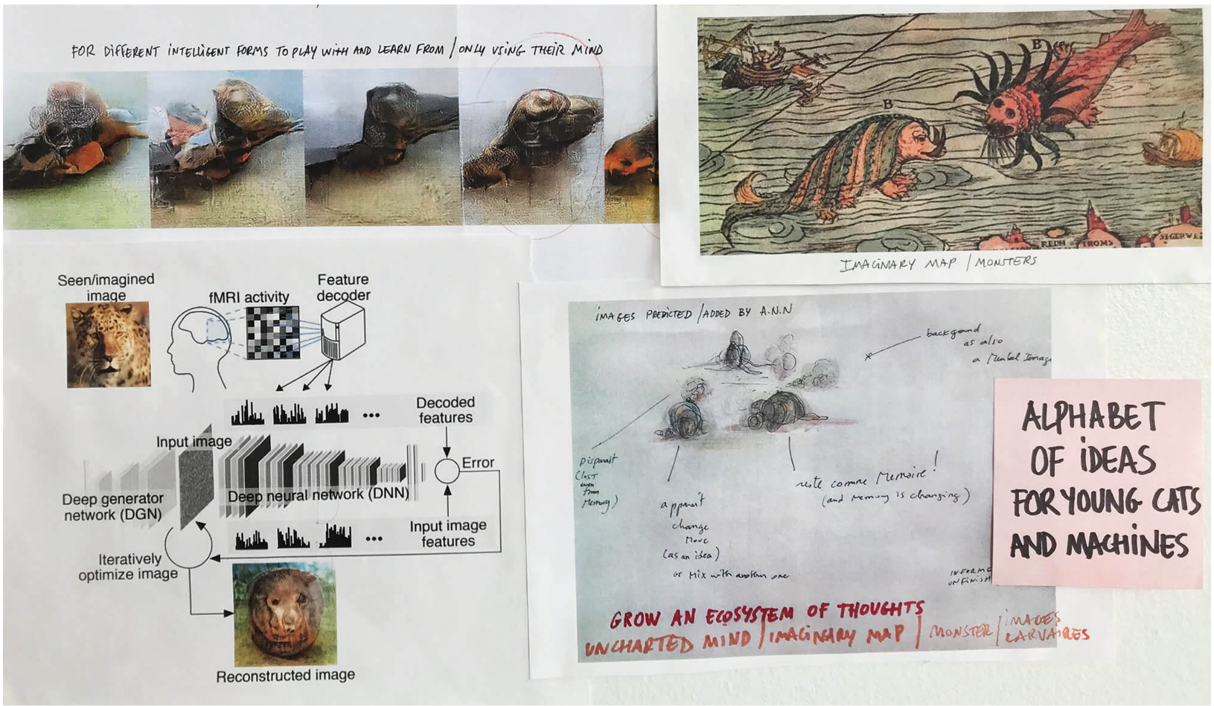
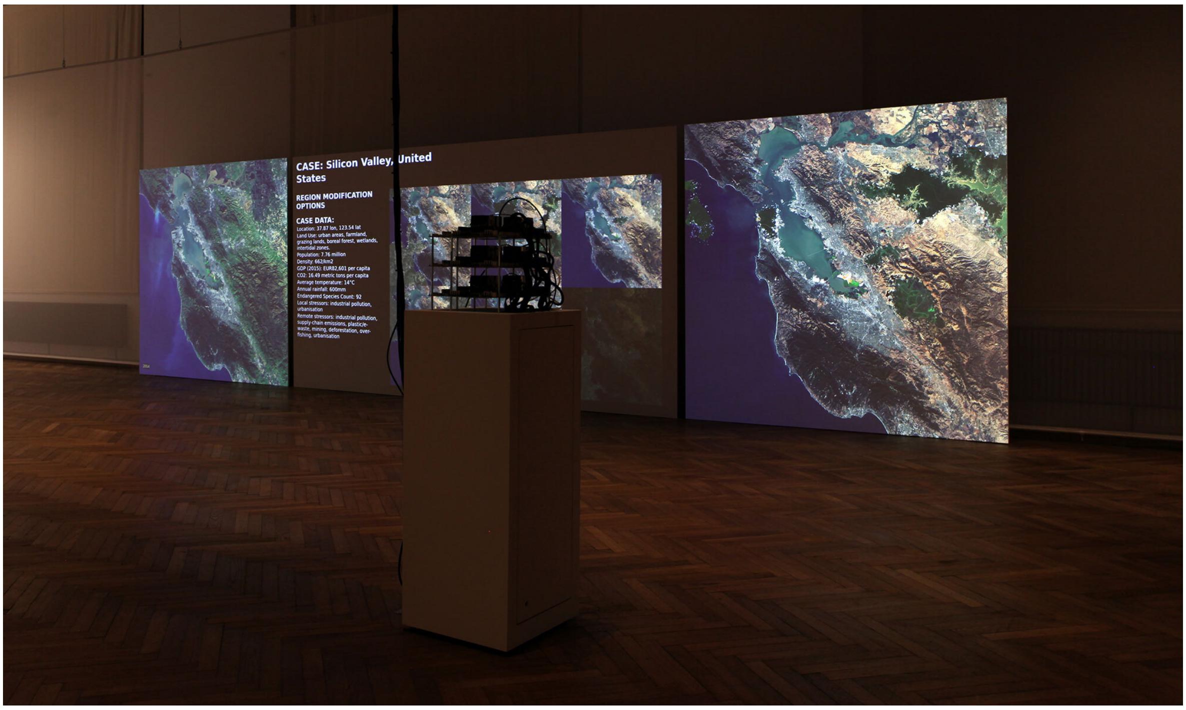

FOR DIFFERENT INTELLIGENT FORMS TO PLAY WITH AND LEARN FROM / ONLY USING THEIR MIND
IMAGINARY MAP / MONTERS
Seen/imagined image
fMRI activity
Feature decoder
Decoded features
Input image
Error
Deep generator network (DGN)
Deep neural network (DNN)
Input image features
Iteratively optimize image
Reconstructed image
IMAGES PREDICTED / ADDED BY A.N.N
background as also a Mental Image
Disparait (lost even from Memory)
apparait change Move (as an idea) or Mix with another one
reste comme Memoire! (and memory is changing)
INFORM UNFINISH
ALPHABET OF IDEAS FOR YOUNG CATS AND MACHINES
GROW AN ECOSYSTEM OF THOUGHTS
UNCHARTED MIND / IMAGINARY MAP / MONSTER / IMAGES LARVAIRES

Figure 5: Pierre Huyghe, study for UUmwelt, 2018

Image courtesy of the artist © Kamitani Lab / Kyoto University and ATR.


CASE: Silicon Valley, United States

REGION MODIFICATION OPTIONS

CASE DATA:
Location: 37.87 lon, 123.54 lat
Land Use: urban areas, farmland, grazing lands, boreal forest, wetlands, intertidal zones.
Population: 7.76 million
Density: 662/km2
GDP (2015): EUR82,601 per capita
CO2: 16.49 metric tons per capita
Average temperature: 14°C
Annual rainfall: 600mm
Endangered Species Count: 92
Local stressors: industrial pollution, urbanisation
Remote stressors: industrial pollution, supply-chain emissions, plastic/e-waste, mining, deforestation, over-fishing, urbanisation

Figure 6: Tega Brain, Asunder, 2019

Credit: Tega Brain, 2019.

What is clear, regardless, is that GANs in particular, and machine learning in general, present a new kind of trading zone within, or even replacing, the trading zone of computer simulation, quite literally living up to Peter Galison's idea that "the computer came to stand not for a tool, but for nature itself" (Galison 2011, 157). The wide adoption of machine learning across the sciences indeed links "by strategies of practice [what] had previously been separated by object of inquiry" (Galison 2011, 157). We can find evidence for this transformation in the increasing appearance of "general" scientific machine learning approaches: in machine learning systems learning to solve partial differential equations (Li et al. 2020), gaining theorem proving capabilities (Polu and Sutskever 2020), adopting symbolic mathematics skills (Lample and Charton 2019), or acquiring specialized domain knowledge just from "reading" the relevant literature (Tshitoyan et al. 2019).

While none of these general approaches are based on GANs, the epistemic implications of GANs described in this paper point to a much deeper problem that is architecture independent: the idea that learning from examples is a sufficient approach to modeling the world might be irreversibly flawed. This is not a new insight. In fact, the question if we need to reintegrate (at least some) innate skills—physical intuition, shape preference, symbolic reasoning—into machine learning systems has been widely discussed since at least 2017 and has attracted some vocal support (Marcus 2020) and some interesting new technical proposals (Sabour, Frosst, and Hinton 2017; Lake et al. 2017; Geirhos et al. 2019). What a close reading of GANs can tell us, then, is how this purely technical discourse is also a discourse about the epistemic faculty of technical images, and how the existing flaws of technical images are amplified in these contemporary mechanisms.

Competing Interests

The author reports no conflicts of interest.

Submitted: December 09, 2020 PST, Accepted: May 13, 2021 PST

REFERENCES

Asim, Muhammad, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. 2020. "Invertible Generative Models for Inverse Problems: Mitigating Representation Error and Dataset Bias." In *International Conference on Machine Learning*, 399–409. PMLR.

Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. 2018. "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks." arXiv Preprint arXiv:1811.10597.

Bau, David, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. 2019. "Seeing What a GAN Cannot Generate." In *Proceedings of the IEEE International Conference on Computer Vision*, 4502–11.

Bredekamp, Horst, Birgit Schneider, and Vera Dünkel. 2012. *Das Technische Bild: Kompendium Zu Einer Stilgeschichte Wissenschaftlicher Bilder*. Walter de Gruyter.

Bruder, Johannes. 2017. "Infrastructural Intelligence: Contemporary Entanglements Between Neuroscience and AI." *Progress in Brain Research* 233: 101–28. https://doi.org/10.1016/bs.pbr.2017.06.004.

———. 2019. *Cognitive Code. Post-Anthropocentric Intelligence and the Infrastructural Brain*. Montreal: McGill-Queen's University Press.

Cao, Sheng, Chao-Yuan Wu, and Philipp Krähenbühl. 2020. "Lossless Image Compression Through Super-Resolution." arXiv Preprint arXiv:2004.02872.

Cohen, Joseph Paul, Margaux Luck, and Sina Honari. 2018. "Distribution Matching Losses Can Hallucinate Features in Medical Image Translation." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 529–36. Springer. https://doi.org/10.1007/978-3-030-00928-1_60.

Cranmer, Miles, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. 2020. "Discovering Symbolic Models from Deep Learning with Inductive Biases." arXiv Preprint arXiv: 2006.11287.

Dosovitskiy, Alexey, and Thomas Brox. 2016. "Inverting Visual Representations with Convolutional Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4829–37.

Event Horizon Telescope Collaboration. 2019. "First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole." *Astrophys. J. Lett* 875 (1): L1.

Gaboury, Jacob. 2015. "Hidden Surface Problems: On the Digital Image as Material Object." *Journal of Visual Culture* 14 (1): 40–60. https://doi.org/10.1177/1470412914562270.

Galison, Peter. 2006. "Images Scatter into Data, Data Gather into Images." *Images: A Reader* 236.

———. 2011. "Computer Simulations and the Trading Zone." In *From Science to Computational Science*, edited by Gabriele Gramelsberger, 118–57. Zurich: Diaphanes.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. "ImageNet-Trained CNNs Are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness." arXiv Preprint arXiv:1811.12231.

Gitelman, Lisa. 2013. *"Raw Data" Is an Oxymoron*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems*, 2672–80.

Greydanus, Samuel, Misko Dzamba, and Jason Yosinski. 2019. “Hamiltonian Neural Networks.” In *Advances in Neural Information Processing Systems*, 15379–89.

Halpern, Orit. 2021. “Planetary Intelligence.” In *The Cultural Life of Machine Learning*, 227–56. Springer. https://doi.org/10.1007/978-3-030-56286-1_8.

Hinterwaldner, Inge. 2017. *The Systemic Image: A New Theory of Interactive Real-Time Simulations*. MIT Press. https://doi.org/10.7551/mitpress/9780262035040.001.0001.

Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. “Image-to-Image Translation with Conditional Adversarial Networks.” arXiv Preprint. https://doi.org/10.1109/cvpr.2017.632.

Iten, Raban, Tony Metger, Henrik Wilming, Lídia del Rio, and Renato Renner. 2020. “Discovering Physical Concepts with Neural Networks.” *Physical Review Letters* 124 (1). https://doi.org/10.1103/physrevlett.124.010508.

Jonas, Eric, and Konrad Paul Kording. 2017. “Could a Neuroscientist Understand a Microprocessor?” *PLoS Computational Biology* 13 (1): e1005268. https://doi.org/10.1371/journal.pcbi.1005268.

Joyce, Kelly. 2010. *Magnetic Appeal: MRI and the Myth of Transparency*. Cornell University Press. https://doi.org/10.7591/9780801460517.

Karras, Tero, Samuli Laine, and Timo Aila. 2018. “A Style-Based Generator Architecture for Generative Adversarial Networks.” arXiv Preprint arXiv:1812.04948. https://doi.org/10.1109/cvpr.2019.00453.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2019. “Analyzing and Improving the Image Quality of StyleGAN.” arXiv Preprint arXiv:1912.04958. https://doi.org/10.1109/cvpr42600.2020.00813.

Kittler, Friedrich A. 2010. *Optical Media*. Cambridge: Polity.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” In *Advances in Neural Information Processing Systems*, 1097–1105.

Lake, Brenden M., Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. “Building Machines That Learn and Think Like People.” *Behavioral and Brain Sciences* 40. https://doi.org/10.1017/s0140525x16001837.

Lample, Guillaume, and François Charton. 2019. “Deep Learning for Symbolic Mathematics.” arXiv Preprint arXiv:1912.01412.

Latour, Bruno. 1999. *Pandora’s Hope: Essays on the Reality of Science Studies*. Harvard University Press.

Le, Lynn, Luca Ambrogioni, Katja Seeliger, Yağmur Güçlütürk, Marcel van Gerven, and Umut Güçlü. 2021. “Brain2Pix: Fully Convolutional Naturalistic Video Reconstruction from Brain Activity.” bioRxiv. https://doi.org/10.1101/2021.02.02.429430.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep Learning.” *Nature* 521 (7553): 436–44. https://doi.org/10.1038/nature14539.

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. “Backpropagation Applied to Handwritten Zip Code Recognition.” *Neural Computation* 1 (4): 541–51. https://doi.org/10.1162/neco.1989.1.4.541.

Lepage-Richer, Théo. 2021. “Adversariality in Machine Learning Systems: On Neural Networks and the Limits of Knowledge.” In *The Cultural Life of Machine Learning*, 197–225. Springer. https://doi.org/10.1007/978-3-030-56286-1_7.

Li, Zongyi, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. 2020. "Fourier Neural Operator for Parametric Partial Differential Equations." arXiv Preprint arXiv:2010.08895.

Marcus, Gary. 2020. "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence." arXiv Preprint arXiv: 2002.06177.

Mentzer, Fabian, George D. Toderici, Michael Tschannen, and Eirikur Agustsson. 2020. "High-Fidelity Generative Image Compression." *Advances in Neural Information Processing Systems* 33.

Mitchell, W. J. Thomas. 1995. *Picture Theory: Essays on Verbal and Visual Representation*. University of Chicago Press.

———. 2005. *What Do Pictures Want?: The Lives and Loves of Images*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226245904.001.0001.

Ng, Andrew Y., and Michael I. Jordan. 2002. "On Discriminative Vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes." In *Advances in Neural Information Processing Systems*, 841–48.

Offert, Fabian. 2019. "The Past, Present, and Future of AI Art." *The Gradient*. https://thegradient.pub/the-past-present-and-future-of-ai-art/.

———. 2020. "There Is No (Real World) Use Case for Face Super Resolution." *Zentralwerkstatt*. https://zentralwerkstatt.org/post_PULSE.html.

———. 2022. "KI und/als bildende Kunst." In *Handbuch Künstliche Intelligenz und die Künste*, edited by Stephanie Catani and Jasmin Pfeiffer. De Gruyter.

Offert, Fabian, and Peter Bell. 2020a. "Generative Digital Humanities." In *CEUR Workshop Proceedings*, 202–12.

———. 2020b. "Perceptual Bias and Technical Metapictures: Critical Machine Vision as a Humanities Challenge." *AI & Society*, October. https://doi.org/10.1007/s00146-020-01058-z.

Pantenburg, Volker. 2016. "Working Images: Harun Farocki and the Operational Image." In *Image Operations*, edited by Charlotte Klonk and Jens Eder. Manchester University Press. https://doi.org/10.7228/manchester/9781526107213.003.0004.

Pasquale, Frank. 2015. *The Black Box Society. The Secret Algorithms That Control Money and Information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061.

Pfau, David, James S. Spencer, Alexander G. D. G. Matthews, and W. Matthew C. Foulkes. 2020. "Ab Initio Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks." *Physical Review Research* 2 (3): 033429. https://doi.org/10.1103/physrevresearch.2.033429.

Pias, Claus. 2003. "Das digitale Bild gibt es nicht. Über das (Nicht-) Wissen der Bilder und die informatische Illusion." *zeitenblicke* 2 (1).

Polu, Stanislas, and Ilya Sutskever. 2020. "Generative Language Modeling for Automated Theorem Proving." arXiv Preprint arXiv:2009.03393.

Raghu, Maithra, and Eric Schmidt. 2020. "A Survey of Deep Learning for Scientific Discovery." arXiv Preprint arXiv:2003.11755.

Rheinberger, Hans-Jörg. 1997. *Toward a History of Epistemic Things. Synthesizing Proteins in the Test Tube*. Stanford University Press.

Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. 2017. "Dynamic Routing Between Capsules." In *Advances in Neural Information Processing Systems*, 3856–66.

Schawinski, Kevin, Ce Zhang, Hantian Zhang, Lucas Fowler, and Gokula Krishnan Santhanam. 2017. "Generative Adversarial Networks Recover Features in Astrophysical Images of Galaxies Beyond the Deconvolution Limit." *Monthly Notices of the Royal Astronomical Society: Letters* 467 (1): L110–14. https://doi.org/10.1093/mnrasl/slx008.

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, et al. 2020. “Improved Protein Structure Prediction Using Potentials from Deep Learning.” *Nature* 577 (7792): 706–10. https://doi.org/10.1038/s41586-019-1923-7.

Shannon, Claude Elwood. 1949. “Communication in the Presence of Noise.” *Proceedings of the IRE* 37 (1): 10–21. https://doi.org/10.1109/jrproc.1949.232969.

Shen, Guohua, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. 2019. “Deep Image Reconstruction from Human Brain Activity.” *PLOS Computational Biology* 15 (1): e1006633. https://doi.org/10.1371/journal.pcbi.1006633.

Tshitoyan, Vahe, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. “Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature.” *Nature* 571 (7763): 95–98. https://doi.org/10.1038/s41586-019-1335-8.

Udrescu, Silviu-Marian, and Max Tegmark. 2020a. “Symbolic Pregression: Discovering Physical Laws from Raw Distorted Video.” arXiv Preprint arXiv:2005.11212.

———. 2020b. “AI Feynman: A Physics-Inspired Method for Symbolic Regression.” *Science Advances* 6 (16): eaay2631. https://doi.org/10.1126/sciadv.aay2631.

Wang, Ge, Jong Chu Ye, Klaus Mueller, and Jeffrey A. Fessler. 2018. “Image Reconstruction Is a New Frontier of Machine Learning.” *IEEE Transactions on Medical Imaging* 37 (6): 1289–96. https://doi.org/10.1109/tmi.2018.2833635.

Wang, Zhihao, Jian Chen, and Steven C. H. Hoi. 2019. “Deep Learning for Image Super-Resolution: A Survey.” arXiv Preprint arXiv:1902.06068.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks.” arXiv Preprint arXiv:1703.10593. https://doi.org/10.1109/iccv.2017.244.