Digital War (2025) 6:3
https://doi.org/10.1057/s42984-024-00098-9

BRIEF COMMUNICATION

Influencing machines: Trevor Paglen and Anthony Downey

Trevor Paglen[1] · **Anthony Downey**[1]

Abstract
How do you train an artificial intelligence (AI), or automated image processing model, to classify and recognize images? This question is central to Trevor Paglen's Adversarially Evolved Hallucination series (2017–ongoing), a project that employs a generative adversarial network (GAN) to classify, identify and crucially, produce unique images. Paglen's series demonstrates how images produced by AI image processing platforms—in this instance, a GAN—are, despite claims, never predictable or, indeed, accurate in their classifications. A significant indicator of this unreliability is evident in the potential for GANs, alongside other generative AI (GenAI) models, to hallucinate and erroneously classify images. Notwithstanding this systemic failing, automated image processing platforms remain central to classification tasks, including those associated with facial recognition and surveillance. They remain, for that reason, central to defining, if not pre-defining, how we perceive and look at the world *through* automated models of machine vision. Encouraged to see like machines, or at least take their classifications seriously and act upon them accordingly, we now inhabit a realm of perception defined by "machine realism", if not algorithmic delusion. Enquiring into how we can better understand the degree to which AI encodes our perception of the world, it is this regimen of "machine realism" that Paglen and Downey explore throughout the following conversation: If AI models of image perception replace ocular-centric ways of seeing, they ask, do these apparatuses have the capacity to not only (pre)define but, in time, further estrange and alienate us from the world?

Keywords Artificial intelligence (AI) · Visual culture · Practice-based research · Digital methodologies · Generative adversarial networks (GAN) · Trevor Paglen · Psyops · Facial recognition · Surveillance

Anthony Downey I want to start with your work on artificial intelligence (AI) and how it relates to your broader research on neural networks and computer vision.[1]

Trevor Paglen I started looking seriously at computer vision and AI in 2010 and 2011. It was difficult to get your hands on a lot of the tools at the time, so I worked with my studio assistants to build a homemade software platform for working with computer vision, which we called Chair.[2] We would assemble different pieces of code or algorithms—a lot of which had been published in the academic literature at the time—and incorporate them into Chair. Working from there, we could instruct the platform that we wanted to look at any picture or video through the "eyes" of a particular algorithm. For example, we could say to Chair, "I want to look at this picture of a cloud through the eyes of a computer vision system similar to those found in a guided missile". This led to an exploration of how AI was producing models of seeing the world.

AD What are the difficulties involved in instructing an AI platform to see, which is to say identify and classify, images?

TP The initial aim was to experiment with object classification—to teach the system to recognize particular classes of images, say, oranges or bananas. It was around this time that I started to understand certain distinctions between classical algorithms and the newer machine learning algorithms. The older algorithms are more Euclidean, or geometric, for want of better words. They tend to discern edges, shapes, gradients, and then use those geometries as abstractions for objects that exist in the world. For example, you can think about a street as being an abstraction of three lines—a line on the left, a line on the right, and a line down the middle. That is how, in broad terms, classical algorithms operate: they look for geometric correspondences among abstractions in order to classify images of objects (such as a street or a car or a building). However, with machine learning,

✉ Anthony Downey
anthony.downey@bcu.ac.uk

[1] Birmingham City University, Birmingham, UK

Published online: 20 November 2024

and deep learning more generally, you take a really different approach: you use huge datasets and instruct the machine learning platform to perform statistical analyses on those datasets (Fig. 1). The model can then, in theory, learn to distinguish among the various images in the dataset.[3]

I was particularly interested in the composition of these datasets and how you can modify and take advantage of the processes involved in their creation. You create a dataset by collecting a huge number of images of one thing, assigning labels (categories) to groupings of these images (for example, "orange" for pictures of oranges, "banana" for bananas, and so on), and then, you use that as a training set to train a neural network to see images that it had not yet encountered. When you are building the image classification model for a machine learning algorithm, it will analyse everything in those datasets, identify self-similar features among all the images in a particular category, and make other connections across all the images. These connections are not simply geometric, however, as is the case with more classical algorithms; they rely on many more parameters, including colours, shades, and shapes.

AD Can you give an example here?

TP If you decide to build a machine learning model to recognize the difference between a turtle, a banana, and the sun, you start by building a dataset—putting lots of pictures of each of those objects into a collection that you will use to train the model. When you do the training, the machine learning algorithms will analyse those images and look for what is self-similar in each object (what constitutes an image of a turtle, a banana, and the sun, for example). A turtle, as an image class, will most likely have an ellipse on the top of the shell; it might have some feet, possibly a tail; the colour green might be dominant. For a banana the self-similar elements could relate to more yellow gradients, with occasional black spots; the overall shape could be defined by arcs. The sun, on the other hand, might be a luminous blob. Once you have trained your model to classify images, it will look at images it has not seen before and try to decide whether there are any turtles, bananas, or suns in the image. Is this new image more akin to a luminous blob, a series of arcs and yellow gradients, or an elliptical green shape? The algorithm

Fig. 1 Image of datasets relating to *Rainbow* (All images courtesy of Trevor Paglen). *Note*: The quality of certain images in these datasets varies, with some having a more pixelated appearance than others. Insofar as image processing algorithms essentially assign discrete numerical values to pixels based on intensities of colours, the level of pixelation in an image is critical to the procedures involved in training a GAN

tries to recognize things by looking for the features most associated with a given object that it has been trained on.

AD In your series Adversarially Evolved Hallucinations (2017–ongoing), there is a move towards encouraging viewers to not only reflect upon AI and its apparatuses but to think from within them—to think from *within* machine models of seeing.

TP When you work with neural networks, you realize that the training data (input) is the key element in the process of image classification, and this is a good starting point for figuring out what is happening inside the so-called black box technologies that constitute the machine learning approach. The Adversarially Evolved Hallucinations series is very much about what happens inside these processes, and how you excavate that and make it more explicit. When someone develops datasets and training sets, for example, they are making assumptions about the relationship of images to objects, and that needs to be made more overt and unambiguous.

AD Can you expand on that: What do you mean by assumptions?

TP Creating a training set involves the categorization and classification, by human operators, of thousands of images. There is an assumption that those categories, alongside the images contained in them, correspond to things out there in the world. There are a few metaphysical assumptions here: first, that there is an uncomplicated correspondence between images and the things they represent out there in the world (i.e., that there is nothing complicated about taking a picture of a turtle and labelling the picture "turtle"). The second assumption is that the "world out there" can be neatly organized into a bunch of self-similar categories (turtles, bananas, clouds, etc.). What follows from the first two assumptions is that you can use quantitative approaches to interpret images (i.e., you can build an algorithm to "recognize" turtles). I refer to these assumptions as "machine realism".

The problem with machine realism is that very few people who think deeply about art or representational theory would take these assumptions seriously. They are incredibly reductive in terms of what they assume about images and what they assume about the "world out there". Images are far more complicated than this. It would be ridiculous, for example, to explain Velázquez's *Las Meninas* (1656) as "a group of people in a drawing room with a dog".

Generative adversarial networks

AD To create the Adversarially Evolved Hallucinations series, you began with various taxonomies—or, to use your terminology, corpuses—including OMENS AND PORTENTS, THE INTERPRETATION OF DREAMS, AMERICAN PREDATORS, and EYE MACHINE.[4] If I understand correctly, each of these corpuses is a dataset and in each of these datasets there are labelled categories of images. So, to take one example, the OMENS AND PORTENTS corpus/dataset contains the categories "rainbows", "comets", "eclipses", and "black cats".

TP Yes, that is the general gist of it—and each corpus can include more than one category of images.

AD So each individual dataset/taxonomy (or "corpus") produces the images we see in the Adversarially Evolved Hallucinations series. So, the OMENS AND PORTENTS dataset/corpus produces the final work *Rainbow* (Fig. 2)?

TP Yes, that is exactly the case: each dataset (corpus) produces a work in the overall series.

AD Given that the corpuses you produce seem intentionally abstract, rather than reductive or deterministic, they tend to be very allusive, if not elusive, concepts; they need a considerable amount of political, cultural, and historical context to fully understand their meaning. I suppose the question I am working towards is why they tend to be more speculative taxonomies/datasets, or corpuses, than the more reductive ones we associate with training AI systems?

TP The series is, in part, a critique of "machine realism", as I mentioned. In contrast to how computer vision systems work, humans have all sorts of weird taxonomies that we use to try to makes sense of the world: taxonomies for dreams, tarot cards, historical events, ideas about some things being "lucky" or "unlucky", and even taxonomies of allegories. Humans' relationships to images are very fuzzy and malleable. This is why I wanted to build datasets for the Adversarially Evolved Hallucinations series using taxonomies from literature, philosophy, art, and folklore. These corpuses/ datasets were designed to perform an immanent

Fig. 2 Rainbow (Corpus: Omens and Portents), 2017

critique of "machine realism", alongside the classical, more Linnaean taxonomies used in machine learning models.

AD Can you talk me through the process? How did you deploy these datasets/taxonomies to program an AI image processing model to *see*?

TP In practical terms, I began by collecting a lot of images from various sources, including the Internet, magazines, and anywhere else that I could find an image that added to a specific corpus. I used search phrases for the type of image I wanted—for example "rainbow". You then have to "clean" the data and try to only select images that are good—that is, images that have rainbows in them. These images aligned with certain concepts or taxonomies that I had already pre-defined such as OMENS AND PORTENTS.

AD And a corpus, such as OMENS AND PORTENTS, is then used to train and generate an image classifier model that can both recognize and generate as-yet-unseen images that *fit* into each corpus?

TP Yes, exactly. If you collate enough images in one corpus in order to train a neural network, you can produce an image classifier that can see things it associates with, for example, omens and portents. Given that the categories included in the OMENS AND PORTENTS corpus included images grouped under the headings "comets" and "eclipses" and "black cats", the image classifier can produce images that resemble comets, eclipses, and so forth. In this instance the machine learning and computer vision model is being trained on datasets that I labelled according to the taxonomies I chose. So, I made a computer vision model that looks around and sees images associated with OMENS AND PORTENTS everywhere, and because the training set is limited to seeing only omens and portents, the model can see those and nothing else. It can only produce, or generate, images associated with the corpus OMENS AND PORTENTS.

AD Let us look more closely at the categories of images you were using. To take one example: the corpus AMERICAN PREDATORS had individual categories of images comprised of "Venus flytraps", "drones", "wolves", and "Mark Zuckerberg". One of the works that came out of the AMERICAN PREDATORS corpus is titled *Venus Flytrap*, but it is a somewhat askew image inasmuch as it has only a passing, somewhat uncanny resemblance to the plant in question.

TP Very much so, and that is precisely why I am using the concept of hallucinations, which in AI parlance often describes things that a model might produce in error—images that are untethered from reality. Something similar happens in ChatGPT as well: it might give a footnote to a book that does not exist or "hallucinate" a fictitious answer to a given question.[5] I would suggest that every product of an AI system always exists on a scale of hallucination in the sense that the apparent correspondence between what a model is classifying and how it relates its systems of classification to referents "out there in the world" is not only misleading but hallucinatory.

AD There seems to be an important distinction here between the deterministic function of machine vision in image recognition and the generative, perhaps more allusive function of the practices you have developed in Adversarially Evolved Hallucinations.

TP The training of neural networks is a stochastic, or random, process inasmuch as you could train a neural network on the same data multiple times and end up with different results.[6] This has to do with the weights and augmentations and so forth that were used in the process. However, facial recognition technologies tend to present themselves as reliable. That is a big assumption, not only about the efficiencies of these technologies, which remains doubtful and prone to error (or hallucination), but also the supposition that the creation of a given training set will adequately and without fail identify future images. In facial recognition technologies, this has serious implications when we consider false positives and how you redress—or even address—the errors of such apparently closed systems. One of the elements in Adversarially Evolved Hallucinations is to draw attention to this and highlight the less reliable elements involved in machine learning systems.

AD The datasets that comprise each corpus are trained through generative adversarial networks, or GANs, to produce artificial neural networks that are capable of recognizing a certain image. Can you talk through the technological aspect of this, the actual workings of a GAN system, and the broader implications of such technologies?

TP Generative adversarial networks use two neural networks that operate in a competition of sorts: one neural network is the "generator" and the other is called a "discriminator" (Goodfellow et al. 2014). In simple terms, the generator can draw things and the discriminator can say what a drawing looks like. Let us take an example: for the individual corpus OMENS AND PORTENTS, I first trained the discriminator on image groupings (categories) that included "black cats", "comets", and other elements associated with prophecies. Once trained through a given dataset, the discriminator can distinguish between categories of images—it can distinguish an image of a black cat, for example, from a comet. Then I start the generator. At first, the generator creates "noise", or random image data that does not resemble anything at all. The generator starts sending completely random drawings to the discriminator, and the latter begins to evaluate how similar that "noise" is to the type of image it is looking for. The discriminator responds to the generator with a "confidence score". The goal of the generator is to create an image that generates a high confidence score from the discriminator. Based on this feedback loop, the generator

will revise the initial image and send a new one back to the discriminator, receive more feedback, then reweigh the generative process until the confidence score improves. This back-and-forth goes on until the generator is able to make images that the discriminator believes are examples of the type of image it is looking for. The discriminator, in short, coaches the generator to create an image that looks more and more like the assignment. I am simplifying here for the sake of illustration, but that is the gist of it—the discriminator and generator go back and forth until they reach some degree of agreement on an image.

AD Hence the term "adversarial"?

TP Yes, this is the adversarial process: the generator produces images until the discriminator accepts them as looking adequately similar to the type of image it is looking for. This is a very streamlined account of a highly complex mathematical process, but suffice to say the generator will eventually produce a synthetic instance of one of the objects (images) described in the model.

AD So, in effect, the original dataset, which consists of image categories such as "black cats", "comets", "rainbows", and so on, are the starting points for training the discriminator neural network in the GAN?

TP Yes, precisely, but the interesting element here is that through training the discriminator and the generator neural networks to be an effective image classification system, via the architecture of a GAN, they can nevertheless only recognize instances of the objects (images) they have been trained on. Say you train a neural network to recognize the difference between a comet and a cat, then that is the only thing the model can classify. If you hook that model up to a webcam and show it a picture of a plate and it might say, "Well, this looks more like a comet than a cat because it has some kind of ellipses and it has more gradients, so it is a comet". You could then show it a picture of a keyboard and it might say, "This looks more like a black cat than a comet, since it is black and white and has stripes", and so on. The model will give each classification a specific assignment that is roughly to do with whether the designation has been made with a high or low confidence score, and the system will operate according to such scores.

AD And then the generator will, based on the feedback of a low-confidence designation, for example, synthesize more images to "match" what it deduces—through the process of looped feedback—the discriminator is looking for?

TP The generator will basically begin to "evolve" in the direction that the discriminator guides it in over time, hence the term "adversarial *evolved* hallucination".

AD You mentioned that there are stages in this process and used the term "primitives", which are images, or parts of images, that reside in the latent space of machine learning systems (Figs. 3 and 4). Could you explain what they are exactly and where they fit into this adversarial process?

Fig. 3 Image from "primitives"

Fig. 4 Image from "primitives". Note: The "primitives" included here are for the purpose of illustrating the multidimensional and latent space of a GAN, and are therefore not intended to be viewed as finished works

TP When you are training a discriminator, you are basically feeding it a thousand pictures of a black cat or a comet or a banana, and the training process will break each of those images down to component parts. For example, a banana is likely to have two arcs ranging from the

top to the bottom of the fruit; it could have yellow colour gradients, some brown spots, a stem at the bottom, and so on. These subcomponents effectively constitute the "latent" space of the model. To classify an image, the model will look for specific patterns of these primitives. This explains, in part, how we are able to take complex images, break them down, and quantify them for image classification systems and machine learning more broadly. The image classification system (discriminator) would be looking out for these subcomponents while the generator, through applied weights and various iterations of data, would be training its system to better produce them over time.

AD These primitive stages of the image would be like sketches or sections of the final image, perhaps?

TP I think about each of the primitives as being akin to a pencil or brush stroke. A specific pattern of pencil strokes makes a face, and a different pattern makes a banana. It took me a while to wrap my head around this: the process is very similar to how you learn to draw. You learn your basic "primitive" or generic shapes. You draw a circle, an ellipse, a triangle, a square, and so on, and then, you conjoin them into more complex objects and images. These are the basic building blocks of drawing. As you evolve as an artist, you tend to forget that this was how you thought about images, but it is also how neural networks operate while they are being trained: the generation of primitives, where an image is broken down into its component parts, will be used as the basis for assembling more and more complicated images in a machine learning system. And these primitives reside in what we might call the latent space of the classification model. Again, I am translating a highly complex mathematical process into everyday language, but that is the general substance of the process.

AD How complex does it get?

TP The latent space of the model contains all the arcs, gradients, lines, colours, shapes, and other components, but the primitives can be much more complex than that—certainly more complex than the shapes that you might learn in a drawing seminar. As you move through the latent space of that model, they become even more complex. In our Chair AI platform, I can task the GAN with making a synthetic image of a place in the latent space of the model. For example, I can propose something like, "Instead of generating an image of a cat, generate an image of neuron 7382", or any other place (neuron) in the latent space. The generator then evolves an image in the direction of criteria dictated by the specificities of what is in the latent space.

AD This returns us to my opening question, albeit with a twist: How can we move away from reflecting upon the operative logic of AI as a "black box" system—an apparatus that is often presented as opaque—and towards a method of thinking from *within it*, if not disrupting its apparatuses?

TP In part that is what I am doing, but I should also note that this is only possible with simple models (and remember that the work we are talking about here was done in the 2010s and came out in 2016 and 2017, which is Palaeolithic in machine learning time). When you start to look at later generative models, like ChatGPT or Midjourney, it is far more difficult (and expensive) to look inside them. I am not sure the kind of work I was able to do on our studio computers in the 2010s would be possible on newer models without a well-funded laboratory with a lot of computing resources.[7]

Uncanny technologies

AD I was thinking earlier of Freud's *The Interpretation of Dreams*, which was first published in 1900, and how that volume relates to his 1919 essay on the uncanny (Freud 1900).[8] Throughout *The Interpretation of Dreams*, there are multiple examples of how the "work" of dreams resembles uncanny or unsettling features related to vision, or visions. I have always understood the uncanny, in this sense, to be a primarily visual phenomenon—an ocular discombobulation of sorts, or an apparition, that disrupts reality and our perception of it. It is about disturbing a frame of reference, or appropriating and inhabiting a given reality. Although AI sees or constructs a reality that is invariably presented as definitive or authoritative, the presence of the uncanny image, in these contexts, would appear to have the potential to subvert such definitiveness or computational certainty.

TP AI systems, be they recommendation algorithms or generative AI models, actively perform processes of manipulation: they *want* you to see something. So, yes, it is about machinic perceptions of reality and how we can disrupt them through alternative models of visualization. We could mention here recommendation algorithms or AI more generally: they are basically influencing machines, and they achieve influence through the extraction of value and the use of classification systems to define certain features of the world we live in. This is also where hallucinations come in: these influencing machines generate hallucinations, uncanny ways of seeing, that make you see something in a particular way or believe something in a particular way. These processes have a military history in the field of psychological operations, or psyops, where military operatives construct realities for an adversary in order to convince them that the world exists and behaves in a particular way. I have been thinking about this process—the notion of how machines operate upon or *influence* the world—with a number of works that I am currently working on. These works use the framework of psyops—partly express in the US Army Challenge Coins I have been collecting (Fig. 5)—to consider how reality is perceived, or how realities are *influenced* by various structures of ways of looking at the world, including those defined by

Fig. 5 US army challenge coin (collection of Trevor Paglen), 2023

political or governmental pressures and demands.[9] Psyops is about influencing the realm of perception; it is the strategic or tactical release of information designed to make people see things that you want them to see, whether they exist or not. If you can make people perceive something, you can make them believe it, but to do that effectively you need to exploit certain features or aspects of human perception or common belief systems. The emotional life and group dynamics of people and communities can become the entry point for effectively altering their psychological attitudes towards the world. And AI is increasingly playing a significant role in this process.

AD I recently recalled, when thinking about Adversarially Evolved Hallucinations, a quote from Sol LeWitt's short essay "Paragraphs on Conceptual Art", which picks up on some of the crossovers between creative and mechanistic forms of production. LeWitt wrote: "The idea becomes the machine that makes the art" (LeWitt 1967). In part, this picks up on a dichotomy of sorts between thought and computation, or the relationship between the probabilistic idea and the deterministic machine. In many ways, it goes to the ontological question of what it is to create: What is the metaphysical or elusive context of creation that you cannot create through a mathematical calculus or with an algorithm?

TP I think that would be a good place to start to try to untangle the dichotomies between human thinking and machine computation, or creativity and the models of statistical analysis that algorithms perform. Thinking about Adversarially Evolved Hallucinations and my more recent work on psyops, it seems the connection lies in how we are striving to get a clearer understanding of what types of "worlds" machine learning is creating. Revealing the systems that produce images or classify them has two functions: One is about demystification. It is difficult to understand intuitively how a statistical engine, or an algorithm, could generate an image that you as a human experience as something that has an inherent meaning and an affinity to your world. This has to do with an implied intentionality, inasmuch as historically it was humans who created images, and there is therefore a tendency to assume that when we see an image it is the product of a thinking, speaking being who exists in the world, even if it is actually a machine learning system that produced the image. Our intuition about creativity cannot fully account for that, so we attribute some kind of supernatural element to the machine. We assume that an machinic image makes sense or that it has meaning. We also assume that the machine must have some level of intentionality, in the way a human might, when it presents us with an image of the world; because we cannot fully explain what is going on, it must be "intelligent". But when you examine the anatomy of how training sets are constructed, you can see the ways in which political and sociocultural assumptions are built into them, and how models of seeing and perceiving the world have become influencing machines, designed to affect and inform our perception of the world and normalize the computational activities involved in that process.

AD There has been a lot of debate recently about how large language models (LLMs) in generative AI are now influencing how we see the world, but this discussion is also about how models, such as ChatGPT and other forms of machine "creativity", are threatening human creativity. You seem to be taking a more nuanced approach to this inasmuch as you are saying AI is not about threatening creativity per se, but realigning, or recalibrating, what we understand the ideal of creativity to be.

TP I bristle a little with the concept of creativity; to me it has a lot of supernatural baggage. I think we have an intuitive understanding of the concept of creativity as being something that comes from somewhere else, something undifferentiated and otherworldly that exists on an imaginative plane. When you look at the fields of cultural production that are now most threatened by AI, they are the ones that have become the most routinized or templated. When you look at things like blockbuster film screenwriting, comic graphics, or stock photography, these are highly refined if not rigid genres, which is why AI models seem to have such an easy time reproducing them.

Machinic ontologies

AD There is a corpus titled EYE MACHINE, which produced the works *A Prison Without Guards* (Fig. 6) and *A War Without Soldiers* (Fig. 7), and I recognize a direct

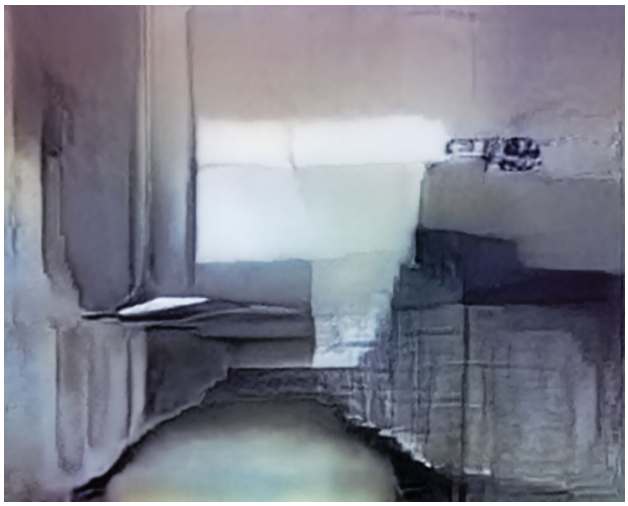

Fig. 6 A prison without guards (Corpus: Eye Machine), 2017

Fig. 7 A war without soldiers (Corpus: Eye Machine), 2017

reference here to Harun Farocki's trilogy *Eye/Machine* (2001–3). When we think about Farocki, and you have written about his practice elsewhere (specifically his work on machine vision and "operational images"), I always think about the role of war in these processes—how image regimes, or machinic ontologies of vision, increasingly define our perception of the world.[10]

TP In media theory, the first Gulf War was a watershed moment in the sense that it is understood as the first war that takes advantage of "smart" weapon systems—weapons that used precision guidance, computer vision, autonomous models of navigation, GPS, and so-called stealth technologies. For Farocki and others, such as Paul Virilio, that war represented a new relationship between images and warfare. It also inaugurated new models of image functionality; images became parts of technical systems—embedded in guidance processes and missile-targeting procedures—and those developments are encased in a black box logic of image production.

AD I want to focus on this for a moment, as it brings us back to the beginning of our conversation, but also raises a further question: How are we being ontologically positioned, if not produced, through and in relation to machinic models of perception? We increasingly appear to inhabit new paradigms of being-in-the-world or subject positions that are being produced *through* the operations of AI systems but, as you observe, these vistas or visions—or hallucinations—are automatically created from models of data extraction and the algorithmic rationalization of data. As machines evolve, our understanding of subjecthood—what it is to be—is evolving too.

TP Do you mean the shift from a what we might call a kind of surveillance or control paradigm of technology towards a generative paradigm of technology—how that affects or produces subjectivity?

AD Yes, I suppose I am drawing here on the Foucauldian notion of disciplinary power: power that evolves according to a given epoch, or era, so that the conventional model of discipline—i.e., control visited upon the body through torture or physical coercion—transforms into a different model of disciplinary control whereby the body, or subject, is rendered "productive" through various technologies, including surveillance. The productive subject, the subject of modernity, can thereafter perform, or conform to, the political and cultural imperatives of a particular era, say neoliberalism, and become established—or positioned—through the apparatuses of statehood or governance. It seems that AI, in our time, has become precisely such an apparatus: it induces the modern subject to perform in a certain way. The technopolitics of machinic vision encourages or induces the subject to perceive reality in a certain way, and to therefore behave accordingly, for want of a better phrase.

TP I think that is exactly it, especially if we consider how contemporary models of disciplinary power are embedded in, say, facial recognition or credit scores, and how that is directed towards the modulation of everyday life based on certain classifications related to consumerism and capital. When I think about the shift towards contemporary notions of discipline, implied in data extraction and applied computation knowledge, and towards the more "productive" turn in AI (how it is increasingly geared towards guiding or influencing decisions or realities), it seems it is designed to efficiently manipulate the subject towards specific ends. This could be getting you to buy something or, more and more, ensuring that you perceive reality in one specific way rather than another. In this sense, it could be about influencing your politics, your biases, your world outlook and allegiances. We do seem to be morphing towards a new turn in AI technologies and, going back to the

idea of psyops, that seems to reside in a shift from the idea of the Internet as a model of surveillance towards an Internet as means to manipulate its users in the direction of certain ends.

AD Machinic visions, it seems, are increasingly becoming *our* vision, or perception of the world. I was thinking here of your 2019 project From "Apple" to "Anomaly" *(Pictures and Labels)*, specifically the focus in that work on how AI networks are taught how to see the world through datasets that perpetuate racist, sexist, and homophobic categories, as well as other models of discriminatory bias (Crawford and Paglen 2019; Paglen and Downey 2020).[11]

TP There are direct connections here between that body of work and Adversarially Evolved Hallucinations, for sure, especially around the normative realms of seeing that machine vision incorporates and replicates. Arguably, Adversarially Evolved Hallucinations and From "Apple" to "Anomaly" explore how the creation of AI models is very much a reality-building exercise.

AD I am wondering here about how calculations of reality impact our political affiliations, and how we understand the politics of activism more broadly. When we think about how algorithms amplify or generate algorithmically defined models of perceiving and engaging with realities, it is as if there is an increasingly ubiquitous algorithmic "command"—both overt and, indeed, covert—that produces reactions to certain political issues, but we still lack the literacy to push back against such techniques or navigate our way through their habitually opaque machinations.

TP I have been thinking a lot about what media literacy would look like in an age of generative AI, and how we need to understand the mechanics of something—be it the "black box" of AI, or the assumptions we make about images in the world. In terms of Adversarially Evolved Hallucinations, I wanted to be able to explain exactly how the images were made and to call attention to the relationship between the datasets and the final images—the relationship between categories such as "comets" and the work *Comet*. I wanted to do this in order to demystify AI as a model of image production. But I also wanted to show how the range of possible things that can be conceptualized, for lack of a better word, by a neural network—or by a generative AI model—is constrained. A model can only produce a version of the world based on its dataset, and it is restricted in that vision of the world by the limits of the dataset it is trained on. That is not to say that generative AI would not have a big impact on culture, because I think it is pretty self-evident that it will.

Notes

1. This conversation, revised for publication here, is an edited version of a conversation first published in Downey (2024a).
2. The software was referred to as Chair because users, in order to operate it, had to sit in what resembled a captain's chair—the designation stuck and became the overall name for the AI platform.
3. Machine learning (ML) and deep learning (DL) are often understood to be subsets of AI. ML is an approach that tends to focus on the use of algorithms to "teach" computers to learn from input data. As a technology, machine learning, as the name suggests, "learns" from input data—datasets of images, for example—in order to make decisions or predictions about future image classifications. The end goal of machine learning is that it will recognize patterns in the input data, eventually make sense of information and, thereafter, "recognize" images that it has yet to encounter. DL is similar, inasmuch as it a type of machine learning. Like machine learning, deep learning uses artificial neural networks (ANNs) to "learn" and to make decisions in manner that is often considered to be similar to the way human brains work. The realm of algorithmic prognostication, once considered in relation to martial models of pre-emptive strikes and Autonomous Weapons Systems (AWS) lends these debates about prediction a far from abstract dimension (Downey 2024b).
4. There are thirteen corpuses in total, including OMENS AND PORTENTS; THE INTERPRETATION OF DREAMS; AMERICAN PREDATORS; EYE MACHINE; THE AFTERMATH OF THE FIRST SMART WAR; MONSTERS OF CAPITALISM; THE HUMANS; THINGS THAT EXIST NEGATIVELY; FROM THE DEPTHS; KNIGHT, DEATH, AND THE DEVIL; SPHERES OF HEAVEN; SPHERES OF PURGATORY; and SPHERES OF HELL. Throughout the Adversarially Evolved Hallucination series, Paglen uses "corpus" as another term for dataset.
5. Such errors, or hallucinations, can be costly. Bard, Google's rival to Microsoft's ChatGPT, incorrectly answered a question about the James Webb telescope in February 2023. This error, or hallucination, saw the share price of its parent company (Alphabet) drop by $100 billion overnight. See Natalie Sherman (2023).
6. Used to define a random sequencing of events, a stochastic process defines the unpredictable evolution of a given event over time. Weather patterns and stock markets are both subject to fluctuations based on numerous factors that render their evolution probabilistic rather than, that is, deterministic. We should note here that AI systems often employ stochastic variables in the algorithms used to train neural networks. A stochastic gradient descent (SGD), for example, is a model used in machine learning to find the best solution by performing steps in random directions—over multi-

ple epochs or iterations—to attain the best possible answer, classification, or prediction (output). These outputs generalize multiple results and then, based on statistical analysis, choose the most likely one.

7. Midjourney is a generative AI program. Launched in 2022, the platform produces images from natural language descriptions ("prompts").
8. Sigmund Freud's *The Interpretation of Dreams* (1900) was originally published as *Die* Traumdeutung and was later published in English in 1913. Throughout his later essay "The Uncanny" (1919), Freud explored the multiple meanings of the term *Unheimlich* (the "unhomely") and its reification as a material or visual object that perturbs the viewer and questions perceptions of everyday objects and events. Freud (1988; orig. 1919).
9. *Psyops* refers to "psychological operations", which involve the dissemination of information to individuals and groups that is designed to influence their subsequent motives in relation to, and understanding of, events. This can begin with the objective reasoning of people and communities, but it can also be designed to impact and manipulate governments and other larger-scale organizations. See Trevor Paglen, "You've Just Been Fucked by PSYOPS", https://paglen.studio/PSYOPS/
10. For a further discussion of automated image processing systems and their deployment in kinetic and non-kinetic warfare, see Trevor Paglen (2014); Paglen and Downey (2020); Downey (2023a; 2024b).
11. For a fuller discussion of *From "Apple" to "Anomaly" (Pictures and Labels)*, see https://www.barbican.org.uk/s/trevorpaglen/. See also Crawford (2021).

References

Crawford, Kate, and Paglen, Trevor. 2019. Excavating AI: The Politics of Training Sets for Machine Learning. September 19, https://excavating.ai.

Crawford, Kate. 2021. *The Atlas Of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Downey, Anthony. 2023a. Neocolonial Visions: Algorithmic Violence and Unmanned Aerial Systems, Ljubliana: PostScript^{UM} 47. https://aksioma.org/neocolonial-visions-algorithmic-violence-and-unmanned-aerial-systems.

Downey, Anthony. 2024a. *Trevor Paglen: Adversarially Evolved Hallucinations*. Berlin: Sternberg Press.

Downey, Anthony. 2024b. Algorithmic Predictions and Pre-Emptive Violence: Artificial Intelligence and the Future of Unmanned Aerial Systems. *Digital War* 5 (1–2): 123–133. https://doi.org/10.1057/s42984-023-00068-7.

Freud, Sigmund. 1988. The Uncanny. In *Art and Literature*, vol. 14, 335–376. London: The Pelican Freud Library, Penguin Books.

Freud, Sigmund. 1900. *The Interpretation of Dreams*. Leipzig: Franz Deuticke.

Goodfellow, Ian J., et al. 2014. Generative Adversarial Networks. *Communications of the ACM* 63 (11): 139–144. https://doi.org/10.1145/3422622.

LeWitt, Sol. 1967. Paragraphs on Conceptual Art, *Artforum*, June, pp. 79–83 (80).

Paglen, Trevor. 2014. Operational Images, *e-flux Journal* 59 (November). https://www.e-flux.com/journal/59/61130/operational-images.

Paglen, Trevor, and Anthony Downey. 2020. Algorithmic Anxieties. *Digital War* 1 (1–3): 8–28. https://doi.org/10.1057/s42984-020-00001-2.

Sherman, Natalie. (2023). Google's Bard AI Bot Mistake Wipes $100bn Off Shares, *BBC*, February 8. https://www.bbc.co.uk/news/business-64576225#.

Trevor Paglen is a multidisciplinary artist known for blending image-making, sculpture, journalism, and engineering into his work. His art, which explores themes like state secrecy and artificial intelligence, has been exhibited globally, including at the Smithsonian Museum of American Art and the Barbican Centre. Notably, Paglen launched an artwork into orbit and contributed to the Oscar-winning film *Citizenfour*. He has also created public art for Fukushima's exclusion zone. An acclaimed author, Paglen's contributions to investigative journalism and art have been recognized with awards like the Electronic Frontier Foundation's Pioneer Award and the MacArthur Fellowship. He holds degrees from UC Berkeley and the Art Institute of Chicago, underscoring his diverse expertise across art, geography, and technology.

Anthony Downey is Professor of Visual Culture in the Middle East and North Africa (Birmingham City University) and the series editor for Research/Practice (2019–ongoing). He sits on the editorial boards of *Third TextDigital War*, and *Memory, Mind & Media*, respectively. Recent and forthcoming publications include *Decolonising Vision: Algorithmic Anxieties and the Future of Warfare* (2025); *Falling Forward: Khalil Rabah—Works, 1995–2025* (2023); and *Topologies of Air: Shona Illingworth* (2022). Downey is the recipient of a series of Arts and Humanities Research Council (AHRC) awards, including four-year multidisciplinary project that focuses on cultural practice and educational provision for children with disabilities in Lebanon, the Occupied Palestinian Territories, and Jordan (2021–2025).