# Objective Vision

*Confusing the Subject of Computer Vision*

**James E. Dobson**

There are now numerous accounts of the use of facial recognition by law enforcement agencies and many news stories extolling the capabilities of the technology corporations that market these algorithms.[1] Such algorithms are part of a larger suite of techniques for the automated analysis of digital images known as computer vision. These techniques are recognized for their ability to recover and create knowledge from latent information found in visual data. Uses of this automated image-based knowledge extraction are now ubiquitous, from popular smartphone apps that identify objects like plants and insects to semiautonomous robotic surveillance systems deployed by private corporations and the US military. As computer vision methods become increasingly complex and trained on large archives of everyday images, the lack of distinction between background and foreground objects may increase uncertainty about the content of the image. Information latent in digital images might pool together through algorithmic operations and become more significant, thus escaping the frame provided by image content labels. In these image-based learning systems, the traces of the past that inform decisions made in the present make use of visual associations, sometimes producing what machine learning researchers call *spurious correlations*. Such traces are decomposed and reassembled through the operation of computer vision algorithms. When these are used in automated surveillance systems, the very real possibility of rendering individuals guilty by association and the displacement and transfer of background object features to subjects in the foreground make it crucially important to understand how computer vision conceives of its field of vision.[2]

DOI 10.1215/01642472-10613653

This article examines the persistence of a mid-twentieth-century ontology of the digital image in contemporary computer vision technologies. *Ontology*, in this instance, functions both as a philosophical concept and as a domain-specific technical term; it is used to name the object/subject relation, as well as the logics of disassembly and reassembly that are core computing processes and enacted in the case of computer vision through the manipulation of image data.[3] The specific computer vision technologies of concern to this article are convolutional neural networks (CNNs), multidimensional and layered neural networks based on neurophysiological models of perception systems that make use of a transform known as a *convolution* between some of the layers in the network. Within the computer vision research community, CNNs have received much praise as powerful techniques for image captioning and object recognition. These networks have also been used for many other nonvisual applications, including text analysis. While CNNs are relatively new, especially for complex visual objects and facial recognition, they have been rapidly adopted for many tasks and applications.[4] These applications are being deployed in many everyday locations, even while the technology is not fully understood. It is this automated categorization and sense making of images and their contents that makes the ontology of computer vision incredibly important. These decisions are based on a series of interlinked, image-subregion-based categorizations and an understanding of these as an assemblage that, within computer vision, defines the meaning of the image as a whole. Because of the ontological issues involved in computer vision and especially in the design of CNNs, these methods and their classifications of image data content are shot through with numerous social and cultural implications. These issues give rise to deeply problematic classifications that learn, among other things, stereotypical correlations between racialized subjects and objects in image data. In applying this opaquely determined knowledge, CNNs reproduce biases and amplify ideological constructions found in visual culture. Sometimes these biases take the form of overt confusion of subjects with objects, but because of the operation of CNNs and the ontological issues underlying them, determining the criteria used for classification decisions is quite difficult and potentially impossible. This article argues that these problems, foundational to the representation of visual information within computer vision and its conception of the image, will not be easily addressed through the removal of suspect and biased data or the correction of previously misclassified data.

• • •

The history of computer vision and its use of neural networks is much longer than one might initially think from the speed of recent developments.[5] The most popular and advanced methods at present are updated versions of those existing at the creation of this field. There were experiments with what we would now call *computer vision* as early as the 1950s. Within this long history, the methodological distinctions between objects and scene, subject and object, and figure and ground have never exactly been neatly drawn. This lack of distinction in the image concept within computer vision results from the influence of the major neurophysiological theories underlying many of these methods. Early computer vision researchers, including Frank Rosenblatt, a cognitive scientist and inventor of the Perceptron, one of the earliest and best-known neural networks, were inspired by simplified mathematical neural models that would connect arrays of sensing devices to simulated electronic neurons. These early neural network models remain the basis of today's much more complex and multilayered networks. In the early 1940s, Warren S. McCulloch and Walter Pitts published their mathematical model of the operation of a neuron, what would come to be called a *McCulloch-Pitts neuron*.[6] This model serves as the foundation for the artificial neural networks used in many artificial intelligence and machine learning projects. More influential for computer vision, however, was psychologist Donald O. Hebb's perceptual theory, published in 1949. While Hebb's account of learning in relation to perception was also based on a neurophysiological model, it modified and remained roughly compatible with gestalt theory, at the time the dominant theoretical account of perception. Hebb's model of perception made no distinction between individual objects within the visual field; a representation of all sensed data was stored as sensory input, and through multiple exposures to these patterns, learning was accomplished. Gestalt theory, Hebb wrote, assumes "that when one perceives a simple figure (such as square or circle) one perceives it directly as a distinctive whole, without need of any learning process and not through a prior recognition of several parts of the figure."[7] Hebb argued that learning to perceive objects is the result of repeated excitation of cells by visual stimuli that form patterns, at the neural level, similar to those produced in response to perceiving a whole image: "Quite simple diagrams are not perceived directly as *distinctive* wholes— . . . though the stimulus has a unitary action in the figure-ground relationship, the perception of identity depends on a series of excitations from the parts of the stimulating diagram."[8] Hebb's theory of perceptual learning enabled computer vision researchers to construct what they called *cell assemblies* from simulated neurons (simulated by customized hardware or in software) that simultaneously addressed responses to patterns of excitations in individual receptors and to the entire visual field.

A decade later, in "What the Frog's Eye Tells the Frog's Brain," Jerome Y. Lettvin, along with McCulloch and Pitts and Chilean cyberneticist Humberto R. Maturana, developed an influential contribution to the theory of neural networks and computer vision.[9] Drawing on observations of a frog's optic nerve, these researchers articulated a theory that shared with Hebb's model an understanding of higher-level encoded responses to visual stimuli—a theory that has shaped both the neural networks and the image concept in computer vision to the present day. This image concept depends on the understanding that learning functions by repeated exposure to small segments of an image; the resulting reinforcement produces higher-level representations, an encoded response of the segments, that are used for the analysis and identification of images.

CNNs are one of the more popular contemporary computer vision algorithms, but they are not completely new: they are the direct descendants of the early neural networks described above and used in the late 1950s and early 1960s for pattern-matching techniques. These networks are designed with many layers and inputs, operate on variously sized image fragments, and introduce convolution, meaning that they perform addition and multiplication on sets of neighboring input values, "pooling" the outputs of earlier layers. CNNs have been used in several applied areas of computation, but none more widely than computer vision. The rapid increase in activity and adoption of CNNs in computer vision was spurred by the development of highly parallel and specialized computing hardware, including graphics processing units (GPUs) and the creation of large databases of labeled image data scraped from the Internet. The resulting performance on highly artificial benchmarks created by the computer vision community has increased interest in using these algorithms for many different image analysis tasks.

The best-known and most influential large image data set for computer vision is known as ImageNet, initially created by Fei-Fei Li, at the time an assistant professor in the Department of Computer Science at Princeton University. Li designed ImageNet as the visual companion to WordNet, a taxonomy created by researchers working on textual data and linguistic research. ImageNet was developed as a joint research program with a large, highly structured data set of images collected and maintained by researchers at Stanford and Princeton. Each of the 1.4 million images has been categorized into a sprawling hierarchical schema in which the last node in the "network" is a descriptive noun corresponding to a preexisting WordNet "synset." In the language of machine learning, these noun synonym sets typically serve as the labels or ground-truth class assignments for the represented people, animals, and objects. *Truth* in this context refers to accuracy of the classifications.

ImageNet researchers hired Amazon Mechanical Turk (AMT) workers to label images with the identifying noun and to place bounding boxes around the primary objects. Data sets like ImageNet are required for almost any conceivable application of machine learning, and most of these have been created by AMT workers. The pay for this sort of work is incredibly low—available evidence suggests something like seven cents an hour for completing image labeling tasks. In their critique of AMT, Neda Atanasoski and Kalindi Vora reference Amazon founder Jeff Bezos's account of the human labor performed on his company's platform as a form of "artificial artificial intelligence," as if to suggest the planned obsolescence of human workers once the machine learning replacements have been fully trained. Atanasoski and Vora describe some of the more insidious aspects involved in the labeling of data for artificial intelligence by AMT workers: "The work that artificial artificial intelligence does to propagate technoliberalism is to perform a future predetermined by contemporary imaginaries of the kind of work that can most usefully and productively be automated for the purposes of capitalist acceleration and accumulation."[10]

In adapting David Harvey's account of "accumulation by dispossession" to the digital world in her explication of surveillance capitalism, Shoshana Zuboff analyzes the dynamic through which people, like the AMT workers commonly hired to process vast amounts of data extracted from visual culture, are compelled to offer up their data, experience, and knowledge as digital material to be mined. Zuboff uses the term *rendition* to name both the demand and the process through which data are produced and refined.[11] The rendering of classification decisions into training data for computer vision algorithms dispossesses people of their labor and knowledge. The human labor that powers computer vision is transformed into criteria that through rendition enable the infinite reproduction of this labor for those using these algorithms, now at near zero cost. This process can be understood as a circuit that processes the raw materials of visual data, separates the meaningful from the insignificant, and assigns values. Seb Franklin describes the stakes of such extractive circuits in his account of digital dispossession. "The techniques and technologies of sorting, integration, values, and discipline," Franklin argues, "are value-extractive counterparts to the procedures through which materials are refined before entering the production process proper."[12] The circuit of knowledge production within computer vision depends on the continued expropriation of visual culture and the ongoing extraction of knowledge about these objects. In some instances, such as social media platforms, people render up their own visual data and knowledge in exchange for access to a closed community or for micropayments in crowdsourced labor markets, but much of this value-increasing extractive dispossession takes place without consent and understanding. One well-known example using computer vision would be Clearview AI, a software and services company that scraped images from the internet without consent, trained a facial recognition system on these data, and sells access to its database and tools to law enforcement departments and potentially to private organizations as well.[13]

Yet the dispossessed knowledge should not be thought of only in terms of identifying personal data or individualized knowledge and expertise, for within these many snap judgments and classifications of visual images are the biases, prejudices, and common sense of ideology. The extraction of these categorical judgments produces, through the collection and processing of large numbers of samples, the sense that these are consensus. Even more, one might argue that the numerical basis of the methods that group and assign individual decisions to the groupings that make up the possibility of categorization establish a greater sense of agreement than would otherwise be found within the population. The numerical methods used in machine learning may contribute to that sense of agreement and reduction of uncertainty by taking an unevenly distributed ideology and rendering back the full force of hegemony, condensing and amplifying fragmented elements of that ideology and discounting and dismissing doubt and mixed responses as noise. While it is tempting to dismiss interactions with commercial machine learning applications as bland and or indecisive—the result of training goals that encourage continued use and reduce liability—any imagined neutrality of these applications and tools obscures the uneven extraction of decisions and preferences and the degree to which their outputs have been filtered and processed through obscure criteria.[14]

In extant critiques of computer vision, the major concerns have been connected to failures of recognition and representation. It is thus the failure of computer vision systems to equally recognize all humans and the identification of biases in the labeling of images or absences of certain categories of people in data sets like ImageNet that have been primary concerns. While the critique of bias in artificial intelligence and machine learning contexts is developing, there has been a considerable amount of research on bias in the areas of recognition and representation. Responding to the architecture of most machine learning operations, researchers typically categorize sites of bias into those found in data, in models, and in the evaluation of the models.[15] For machine learning researchers Joy Buolamwini and Timnit Gebru, such forms of bias are connected primarily to what they characterize as a recognition problem, the core issue being the greater failures of recognition and higher misclassification rates for certain populations, especially, as they demonstrate, for darker-skinned women.[16] Because computer vision is deeply dependent on recognition for higher-level tasks, such modes of failure are examples of bias as a form of exclusion from the frame of computer vision by data and methods not calibrated to diverse bodies and presentations. Computer vision thus can be said to exhibit bias through nonrecognition. Simone Browne understands such biased failures of recognition as a violation, an example of ontological insecurity in which "white prototypicality" forms the criteria by which biometric technologies register and track human subjects.[17] Her analysis of common instances of a "failure to enroll," another name for the lack of recognition, exposes the norms embedded within technology. Other forms of similar failures of recognition within image classification systems include the frequently racialized misclassification of people as animals or differential assessments of some groups of facial expressions, such as smiling and blinking.

While locating evidence of embedded or reproduced cultural bias in pretrained models and in algorithms has been technically challenging—and especially with the increasing opacity and complexity of these models—pointing to bias issues within images used to train these models has been much easier. In computer vision, because such data, at least at their origins, are primarily visual images, identifying a lack of diversity, biased selection criteria, or improper labeling of data can be accomplished through even cursory glances at image databases. The bias in such data sets has thus attracted much more critical attention than the methods and the underlying algorithms themselves. These data sets, as previously mentioned, comprise images that have been categorized, frequently into hierarchical relations or taxonomies, and labeled by humans. The labeling of images, a process involving the adhering of a standardized name that functions to provide a map between the digital image content and a category, is crucial to the training of all kinds of computer vision algorithms. The labeling process and the categories themselves are obvious sites of bias in computer vision. The degree to which outsourced, low-paid, and hidden labor is a necessary component of the labeling of images within data sets also has been critically examined.[18] Kate Crawford and Trevor Paglen locate what they term the *politics of images* in the manually organized and labeled taxonomies of these large image data sets.[19] Crawford and Paglen offer as examples a number of deeply troubling categories and subcategories of human subjects found in ImageNet, which, as noted above, is one of the most important and foundational computer vision image data sets. "As we go further into the depths of ImageNet's Person categories," they write, "the classifications of humans within it take a sharp and dark turn. There are categories for Bad Person, Call Girl, Drug Addict, Closet Queen, Convict, Crazy, Failure, Flop, Fucker, Hypocrite, Jezebel, Kleptomaniac, Loser, Melancholic, Nonperson, Pervert, Prima Donna, Schizophrenic, Second-Rater, Spinster, Streetwalker, Stud, Tosser, Unskilled Person, Wanton, Waverer, and Wimp. There are many racist slurs and misogynistic terms."[20]

Academic and research competitions such as the ImageNet Large

Scale Visual Recognition Challenge (ILSVRC) have attempted to limit bias through a reduction of the data set. In focusing on one thousand selected objects, the challenge attempts to set aside some of the troubling classifications concerning human subjects (especially connected to race and gender), yet this reduced data set, and thus the resulting neural network, still includes representations of many people. The creation of object models from selected object categories, reinforced by the rules of the competition, was an attempt to limit bias by removing these subjective categories. With this reduced set of categories, objects rather than people would now be the primary representation found within the selected training data. Yet objects in computer vision are commonly defined by association, meaning that the representations are of natural scenes depicting the use of objects by humans, rather than of objects as such, objects for and by themselves.[21] This is key: the copresence of humans and objects is necessary for the models eventually, after extensive training and evaluation, to be able to detect objects in images containing people. Training only on closely cropped object images would produce models that would not be generalizable to "natural" scenes, images in the wild containing these actual objects. In some ways it is useful to think of the cropping of an image itself as a major component of the task to be automated in object recognition.

One of the more widely used pretrained CNNs for image recognition tasks is known as Inception. This CNN was created by a team of scholars from Google, the University of North Carolina, the University of Michigan, and a corporation called Magic Leap. These researchers introduced this deep network in "Going Deeper with Convolutions," a technical paper that took its title from a then-popular internet meme featuring the text "we need to go deeper."[22] In their paper, the creators of Inception demonstrate and explain the power of their neural network, which was twenty-two layers deep, for detecting and classifying images. This CNN was designed and trained to perform well on a single and narrowly defined task: the ILSVRC 2014 Detection Challenge. The Inception neural network, like many other CNNs, uses training data to learn how to recognize objects represented as matrices of pixels located within images. Some of the neural network layers are specialized for certain features. Inputting an image through the network returns the most likely classifications in terms of the provided classes, which are WordNet synsets. These image class categories caption the image, assigning an explanation or meaning via descriptive labels. Captioning is a seizing, an arresting of the slipperiness of meaning within the images through the halting assertion of language. The Inception network attempts to fix the meaning of an image as it is filtered through the network and generates probable classifications. The algorithm, to put it fancifully, becomes taken by certain features identified within the image. These features or image-part objects, as "seen" or recognized by computer vision, are assembled into something resembling a map—a map that, in the case of a face, does not resemble the subject as subject but instead offers a high-dimension graph of select image features from across the entire image.

CNNs trained on ImageNet object classification tasks return class probabilities: the statistically determined confidence that the returned class label is the most appropriate one for a particular image. Several classes are typically returned by querying the model, sorted in order of these probabilities. For example, running an image of a golden retriever through such a neural network might return the following five object-class and class-probability pairs: ("golden_retriever," 0.82016516), ("Labrador_retriever," 0.024441281), ("Leonberg," 0.015037035), ("Tibetan_mastiff," 0.0124089625), ("bloodhound," 0.010461364). The returned classes, as this example demonstrates, are sometimes from similar categories or even from within what we might consider the same taxonomic tree; we see here that while the first class was the correct classification, the following four were from the same tree: different dog breeds with increasing dissimilarity. Sometimes the returned categories are incoherent, such as an object not included within the reduced set of object categories or not similar enough to the samples included within the training data. Such incoherent results might also suggest that the extracted samples and learned features from these samples are not specific enough but, rather, are tied to generic features of the images, that is, to the presence of similar pixel values in similar locations within the images. While the presentation of probability values suggests a slipperiness to classification, the structure of operations and the affordances of the machine learning operations used in computer vision reinforce the notion of correctness or truth—a troubling notion when applied to the longer history of visual representation. An insidious ideological effect is active within such operations: one tacitly or explicitly endorses a standard of correctness based on matches with prior decisions that are not available for inspection and for critique.

Understanding how these classification decisions are made is a growing area of research in computer vision. One of the problems with computer vision is our not being able to see as the machine sees. Paul Virilio conceives of computer vision as "sightless vision," a form of sensing of visual information from a scene characterized by the absence of a perceptive human subject.[23] Computer perception as such does not exist. Perception, as early computer vision researchers noted, requires the decomposition of the various elements of a scene and an understanding of how they fit together through experience.[24] Putting aside the question of whether or not computers "see" and what that seeing looks like, the crucial problem remains understanding how visual data are organized and used to inform decisions within computer vision. As methods increase in complexity and features become less definable and recognizable by humans and more determined by neural networks from what would appear to be random patches of pixels, understanding how decisions are made and what informs these decisions becomes more and more difficult. This problem of understanding in machine learning is deeply connected to contemporary discussions of explainability, the degree to which a model can be said to be explainable in human terms. Explainable models allow inspection of the most important features used to determine classifications and the link between these and the various probabilities generated by the model.

The task of neatly segmenting digital images into distinct regions corresponding to different objects is not trivial. This segmentation, in part, means the task of object recognition involves defining objects by their location within image regions. Within CNNs, data derived from digital images might be thought of as porous; the pooling of multiple feature maps, which are composed of smaller patches or matrices of pixels, draws data from multiple, overlapping image regions. In the case of object recognition with specialized neural networks, frequently co-occurring visual objects will inform the classifier. In their research to understand the criteria by which humans might explain predictions made by computational models, Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin used Google's Inception to train a classifier to distinguish between wolves and huskies. All of their selected images featuring wolves deliberately had snow in the background. "We trained this *bad* classifier intentionally," Ribeiro and his colleagues wrote, "to evaluate whether subjects are able to detect it."[25] The presence of snow became an important feature for the classifier to distinguish between images of wolves and huskies. While this study didn't demonstrate that the Inception network by itself was making an incorrect classification decision between these two existing categories, it does show how spurious correlations, such as the presence of snow in the background of these images, can inform models and the difficulty in understanding the presence of such spurious data. Yet spurious data are not the only problem. The lack of presence, too, as Jessica Marie Johnson remarks in her account of the "shadowed spaces of empty cells, null values" found in historical data of slavery, disproportionally dismantles and dismembers some subjects.[26] Like the empty values found in slavery's archive that register through absence of uncounted lives, object-based computer vision techniques include but do not mark as present human subjects. Attentive and overfitted to some visual features, CNNs break apart the visual field and alternate between noticing and losing track of image components. The computational acts of making and remaking image representations involved in the CNN can thus amplify existing inequities, absences, and partial representations found within data.

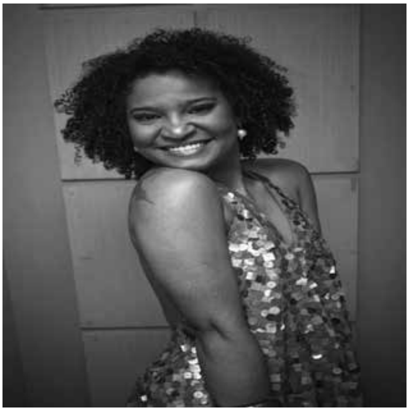

Figure 1. Renato Lorenzetto, "Português: Larissa Luz vocalista da Banda Ara Ketu." Licensed under Creative Commons Attribution-Share Alike 2.0 Generic license.

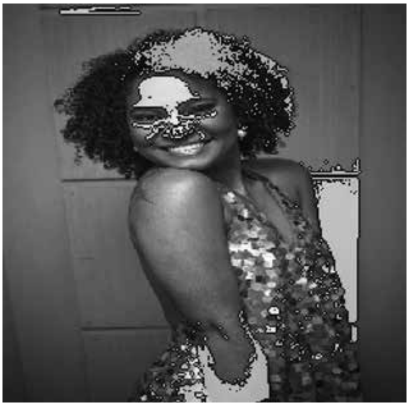

Figure 2. Predicted object classes and class probabilities: ("wig," 0.28289402), ("miniskirt," 0.21179298), ("feather_boa," 0.114811346), ("maillot," 0.059417136), ("maillot," 0.043977804).

The problem motiving this inquiry at present concerns the presence of so-called spurious data already within trained models. How would we know if ImageNet, rather than the researchers mentioned above, had unintentionally made selection choices that resulted in a large number of training images of wolves with snow in the background? Given the narrow scope in terms of trained categories found in many ImageNet-based algorithms, including Inception, determining the influence of representations not identified as central to the images is incredibly difficult. The complexity of such a task is compounded when we consider not only background data influencing foreground objects, as in the case of snow in images of wolves, but the presence of humans within the field of images purported to contain primarily objects. As these individuals are used to give meaning to the objects—holding, wearing, using, examining—they are necessary elements for the image qua image, but in terms of computer vision they represent potential sources of spurious data. Two kinds of data leakage are at work here. The first is a leak of these representations from training data into criteria for classification. The second is a leak of attributes from subjects to objects and from objects to subjects. These leaks trouble any classifications produced by these algorithms. Take the case of an object like a bonnet, an image contained within the reduced-set ImageNet object categories. Many training images labeled "bonnet" will contain a bonnet present on the head of a baby, so images of a baby face might be classified as a bonnet, even if the full head is visible and there is no bonnet in the scene. In the case of the snow in the background of images of a wolf, the snow might become a carrier of information about the wolf for use in classification. For the bonnet, the baby face features become important to the classification of the object. The trained network will recognize a bonnet without the presence of a baby, but these features are latent within the classification criteria. This problem becomes much more pressing, however, when representations attached to human visual features leak across the imaginary boundaries between subjects and objects and into future classifications. In exploring with my students the application of ImageNet-trained CNNs to a variety of images extracted from social media, these theoretical issues quickly became very real problems. Using Inception or VGG16, another popular pretrained CNN, some images of Asian people, no matter their location or action, would return as the highest confidence object category "jinrikisha" or "kimono," and some images of Black people, wearing a variety of hairstyles, would return "wig." These two example object categories demonstrate two different problems: the former pulls subject features from object images that are highly correlated with subjective features from a single population, whereas the latter overprivileges textures and collapses differences across hair styles to overidentify a population with a single object category.

The experiment involving images of the husky and wolf by Ribeiro and colleagues was part of a larger project to explain predictions of machine learning classifiers. They wanted to develop an algorithm and data-agnostic approach to understanding the results of algorithm-based classification decisions and the criteria by which they were made. The criteria are highly dependent on the task at hand and notoriously difficult to pinpoint. In the case of CNNs used for image classification, the multiple layers used in the models make it difficult to point to influential features that can then be mapped back onto an image's individual identifiable pixels. The method selected by Ribeiro and colleagues, called *local interpretable model-agnostic explanations* (LIME), produces a data representation that they argue remains faithful to decisions used by the classifier while also being much more interpretable by humans. LIME allows what we might think of as reverse mapping from decision to evidence or, rather, to features that, in the case of an image, can be displayed or projected on top of the original pixel data to display the regions of the image most likely informing the decision. Provided with input comprising images that feature humans (figs. 1, 3, and 5), the CNN will return predictions, in the form of probabilities, for a set of object classes. Figures 2, 4, and 6 are captioned with the set of returned predicted object classifications, and the images show overlays of image regions returned by LIME. The regions identified in such images are usually located in the background, but in many of these images small foreground groups of pixels are overlayed on individuals, attached to their face, hair, or other features. The presence of

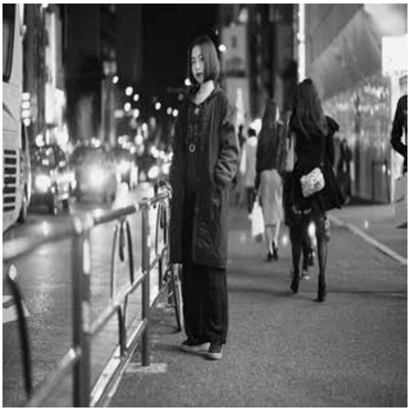

Figure 3. Dick Thomas Johnson, “Harajuku Fashion Street Snap (November 11, 2017) (46370495214).” Licensed under Attribution 2.0 Generic (CC BY 2.0).

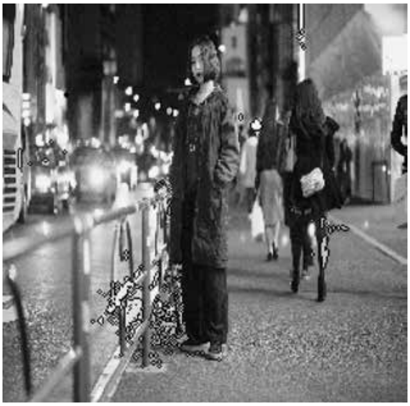

Figure 4. Predicted object classes and class probabilities: (“jinrikisha,” 0.34868827), (“streetcar,” 0.1608768), (“unicycle,” 0.03484652), (“crutch,” 0.03411216), (“parking_meter,” 0.023157952).

these regions in unexpected locations is a symptom of the problems inherent in computer vision’s flattened and confused ontology.[27]

One way to conceptualize the problem presented by this article is through the theorization of transfer learning. The *transfer* in transfer learning typically names the process by which what would otherwise be a mismatch between a training task and the applied target problem becomes a solution through the generalization of the task. To make use of transfer learning, one must cross the boundaries that separate target and training data. Transfer learning is thus nonspecialized learning that seeks to exploit already acquired knowledge. The generalization of the training problem—in the case of computer vision, generalization would include learning how to discriminate among a subset of object classes and then the application of the generated criteria on another, different subset of objects—involves a whole series of assumptions about the ability of task- or domain-specific knowledge to have predictive power for multiple tasks. Applying the concept of transfer learning to the unstable ontology of computer vision enables two distinct ways for thinking through the movement of features from objects to subjects and from subjects to objects and back to subjects. The first includes thinking about data “leakage” within computer vision as enabling transfer learning by allowing knowledge about different types of visual objects to enter into the training data and thus into the decision-making mechanisms embedded within neural networks. Training on images of cars on the road or streets, for example, might produce networks that can also recognize other types of vehicles. The second perspective is to understand transfer learning as always already present within these object models through the transference of learning across the entire image. The flattened ontology and the multiple feature maps make everything within the image a possible feature. When images representing human subjects are brought through neural networks trained primarily but not exclusively on objects, high-level image features are transferred, as it were, from subject to object and back to subject. This transference of features from a person present within a scene by a CNN that has been trained primarily to recognize an object is another form of data leakage. We can consider it a leak because data not expected to be included as part of the training data set have leaked into the data. This data leakage might not be noticeable in many applications of the object classifier, because the trained classifier will most likely produce high probability values for the "correct" class. This transference of features between objects and humans through what I am characterizing as data leakage might become visible only when applied to images containing representations of humans.

Leakage, as it turns out, might flow more easily from certain categories of visual features. In attempting to understand how the architecture of CNN models influences decisions, computer vision researchers have noticed the surprising preference for textures in models trained on ImageNet data.[28] Many computer vision algorithms since the origin of the field have been in large part modeled on recognizing shapes rather than textures. Within these theories of human vision, people learn to differentiate between objects primarily by recognizing their shapes. Because these computer vision models are difficult to understand in simple terms, identifying the important features used for decision making is not a simple task. Different layers within these deep models, researchers propose, have different biases for certain types of features, where features may include blobs, textures, colors, and edges. Texture, of course, can easily map onto both the animate and the inanimate: human skin and hair are highly defined by texture, as are textiles, those woven threads that give etymological origins to *text* and *texture* alike. While textures in themselves contain boundaries—the lines invoked by the filaments common to human hair and artificial woven fibers—they also flow across images. The blurring of such boundaries renders indistinct flesh and surface. These features are not explicitly described but "learned" through the repeated sampling of image data from labeled objects. This is an important aspect of contemporary computer vision. In the past, specialized algorithms composed of programmed descriptions were developed to recognize certain image features, and these were used to train image classifiers. Edge detection algorithms, for example, were used to extract the presence of lines and circles that formed the basic outline or shape of objects. Given a large sample set, these various shapes would define the range of pos-

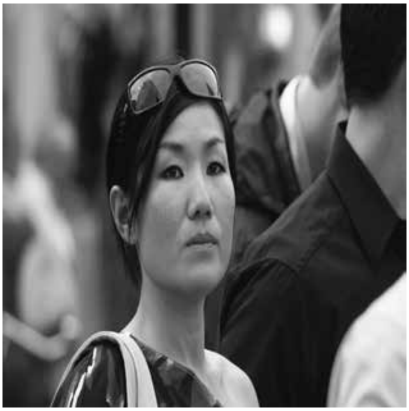

Figure 5. Dick Vos, "Portrait of an Asian Woman." Licensed under Attribution-NonCommercial-ShareAlike 2.0 Generic (CC BY-NC-SA 2.0).

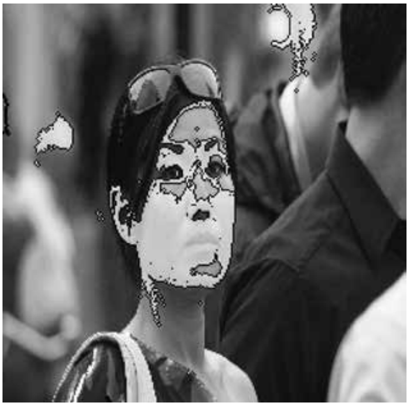

Figure 6. Predicted object classes and class probabilities: ("kimono," 0.5265296), ("sunglass," 0.047480945), ("groom," 0.041185725), ("sunglasses," 0.03571364), ("academic_gown," 0.030526483).

sible shapes found within the diversity of object samples. These feature detection algorithms have been supplanted by neural networks that learn a series of patterns for discriminating among objects without being told what constitutes a pattern.

The appearance of DeepDream, a tool to visualize what patterns CNN has learned from images, led to the widespread sharing of disconcerting visualizations of fractal-like repeating images of animal faces and eyes overlayed on everyday objects. These were produced by applying Google's Inception network, trained to recognize animals and human faces, to images not containing these objects, even on nonrepresentational images of random, noisy data. Fabian Offert and Peter Bell call such visualizations "technical metapictures" that recursively show perceptions of perception.[29] In her account of DeepDream, Hito Steyerl argues that these images can become tools by which we understand the logic of computer vision. Such experimentation, Steyerl argues, "reveals the presets of computer vision, its hard-wired ideologies and preferences. The result: a rainbow-colored mess of disembodied fractal eyes, mostly without lids, incessantly surveilling their audience in a strident display of pattern over-identification."[30] What Steyerl calls *over-identification* we can understand through the results of research in computer vision as a preference for certain kinds of features. These preferences are the result of model architecture and training data. The overlayed eyes and animal faces show a model working overtime to recognize something of importance, even when these features are not at all important to an image. CNNs locate minor features and turn these into the criteria by which a classification can be made. In computer vision, as the image space itself is constructed as an unstructured and flattened ontology, these minor features leak from background to foreground, from subject to object, and from object to subject. Algorithmic preferences, known or unknown, run freely through this space as the algorithms apply these preferences as filters through which they see the world.

What if the problems in the image ontology in computer vision identified above—what some might call merely spurious correlations within image data—function more frequently like what Ruha Benjamin, in her analysis of technical systems, calls "racist glitches"? Benjamin argues that, while we generally think of the glitch as the "fleeting interruption of an otherwise benign system," we might better understand it as a "slippery place" located "between fleeting and durable, micro-interactions and macro-structures, individual hate and institutional indifference," as the signals of the operation of a system rather than a "fleeting interruption of an otherwise benign system."[31] Like the hidden attributes informing the predictive policing models examined by Virginia Eubanks, or any number of other big-data predictive applications, these glitches will occasionally spill out into results—and into the public.[32] Stereotypes are being learned and transferred from scenes and objects to people, from mountains of scraped images used as training images. The glitches that appear, even within carefully constrained data sets and categories, register systemic issues. These are patched together, informing and forming the archive of visual culture. Such systemic issues, of course, reside in the larger frame, and thus saturate the ontological field deployed by many common computer vision applications. This means that these methods cannot escape recording, reproducing, and occasionally exposing the operation of this system. After all, higher-level feature correlations resulting from the comparison of image data make the identification of images possible. When one is attempting to identify objects, correlations in terms of features learned from training on similar images is assumed to mean that these objects are similar, that they all belong to the same class and have some family resemblance. The shape of this class of objects, which is to say, the breadth and range of possible variation, can be established only from the historical comparison of samples and is then used to make future decisions. These historical data and any knowledge produced from them for making decisions in the present will continue to influence and inform future decisions. Wendy Hui Kyong Chun succinctly summarizes the problems involved in using such mechanisms: "Correlations, again, do not simply predict certain actions; they also form them. Correlations that lump people into categories based on their being 'like' one another amplify the effects of historical inequalities."[33] Historical image data—the archive of training images—contain correlations that become decisive criteria for CNNs to recognize images as belonging to defined categories. The amplifications invoked by Chun are challenging to understand partly because they come from the past and partly because the exact source of these decisions is incredibly difficult to locate. Louis Amoore gives an accounting of the multiple sites—crucial for this analysis not just in terms of the training data but in the multiple variables and sources found in computer vision algorithms—in which we can read the politics of neural networks:

> The recognition of edges, motifs, and familiar arrangements is not designed into rules by a human engineer but is definitively generated from the exposure to data. To be clear, this spatial arrangement of probabilistic propositions is one of the places where I locate the ethicopolitics that is always already present within the algorithm. The selection of training data; the detection of edges; the decisions on hidden layers; the assigning of probability weightings; and the setting of threshold values: these are the multiple moments when humans and algorithms generate a regime of recognition.[34]

Influences from other visual objects frequently co-occurring with those labeled and identified as belonging to a distinct class of objects can leak into the set of features, the knowledge-making apparatus itself, that establishes criteria for object identification. As I have demonstrated, learned correlations are frequently what the computer vision literature calls *spurious*, but to frame the problem in this way is a simplification that reinforces the correctness of human categorizations and understandings about the world.

Despite its sense of mechanical automation and cutting-edge complexity, computer vision is not a rupture with the history of visual culture and the legacies of seeing and interpreting images. As I have demonstrated, CNNs encode and obfuscate quite human ways of seeing the world, and the image repertoire used to train these algorithms is struck through with the residues of prior representations. Within visual culture, people are subject to different histories and received representational practices and repertories. Kimberly Juanita Brown argues that "Black women enter the frame photographically—as a collective body always already in pieces; their inability to unify a collective discourse only lengthens the measure of their malleability."[35] For Brown, Black women photographers have the ability to reframe such prior representations and can thus exploit what she sees as the potential of photography to look back on itself. The ontology of computer vision limits such potentials as it duplicates not the frames but the fragmentation. For those already imaged into pieces, CNNs and similar algorithms conduct further dismemberments and reassemblies not as one would want to be seen but in hegemonic terms.

Using CNNs in computer vision as a case study, Leif Weatherby and Brian Justie have argued that artificial intelligence should be understood as indexical: a mapping of neurons that point to learned generalizations. "The trained net," they argue in an analysis of a CNN architecture used for classifying animals, including breeds of dogs, that is quite similar to the one analyzed above, "possesses no holistic sense of Samoyedness but rather a complex architecture of indexical pathways that point to Samoyedness by capturing salient relations between features."[36] As we have seen, the pathways traced through CNNs point to more than just the indexed object. The patterns discovered by CNNs are complex. The ontological organization of these networks ensures that features will be drawn from fragmented images that have been disassembled and reassembled many times. Like many other decisions made by large-scale artificial intelligence systems, these classifications and decisions are hard if not impossible to question and critique. The opacity of this indexing and the immense data sets used as training make rooting out biased, stereotypical, or imbalanced data almost impossible. The glitches appearing in CNN models trained on image data are ultimately the residue of larger problems found in the social systems that the image archive represents and that these algorithms have modeled and amplified. Biases in image data sets, in short, function like repeated exposures to the part-objects of the whole: the microaggressions that form the background of everyday life. Such biases saturate the visual field, both figure and ground. We cannot expect nonbiased output from algorithms that read the whole frame when the frame itself is already acknowledged as saturated with all the biases and prejudices of culture, both historical and present.

**James E. Dobson** is assistant professor of English and creative writing at Dartmouth College. He is the author of *The Birth of Computer Vision* (2023) and *Critical Digital Humanities: The Search for a Methodology* (2019) and coauthor of *Moonbit* (2019).

## Notes

[1] For a notable example, see Wingfield, "Amazon Pushes Facial Recognition to Police."

[2] Computer vision concerns representational rather than synthetic images, which are the primary objects of computer graphics. Jacob Gaboury's recent archeological analysis of computer graphics examines the construction and ongoing development of this field through its treatment of its two primary digital objects: the screen and the simulated image (Gaboury, *Image Objects*, 23).

[3] Alexander R. Galloway discusses the multiple ways in which the term *ontology* has been deployed in computer and information science as well as in philosophy (*Interface Effect*, 64–67).

[4] The most important early contribution to the development of CNNs was a network that recognizes handwritten numerical digits in the form of zip code data

supplied by the US Postal Service. See LeCun et al., “Backpropagation Applied to Handwritten Zip Code Recognition.”

5. For a history of early computer vision, see Dobson, *Birth of Computer Vision*.

6. McCulloch and Pitts, “Logical Calculus of the Ideas Immanent in Nervous Activity.”

7. Hebb, *Organization of Behavior*, 17–18.

8. Hebb, *Organization of Behavior*, 18.

9. Lettvin et al., “What the Frog’s Eye Tells the Frog’s Brain.”

10. Atanasoski and Vora, *Surrogate Humanity*, 100.

11. Zuboff, *Surveillance Capitalism*, 232–33.

12. Franklin, *Digitally Disposed*, 67.

13. Smith and Miller, “Ethical Application,” 169–70.

14. For an example of these training goals for interactions, see OpenAI’s Reinforcement Learning from Human Feedback procedure as part of the “model-assisted safety pipeline” used in GPT-4 (OpenAI, “GPT-4 Technical Report,” 12–14).

15. For an example of such a schema, see Davinder et al., “Trustworthy Artificial Intelligence.”

16. Buolamwini and Gebru, “Gender Shades.”

17. Browne, *Dark Matters*.

18. Parisi, “Negative Optics in Vision Machines.”

19. Crawford and Paglen, “Excavating AI.”

20. Crawford and Paglen, “Excavating AI.”

21. As these object-oriented CNN models are trained on objects within natural rather than studio scenes, these neural networks might be said to function something like the fin-de-siècle natural history museum’s life group display of historical objects. Like Bill Brown’s sense of the “unstable ontological status between the animate and the inanimate” produced by the life group display, these models that are by necessity trained on images of objects in the wild cannot help but reproduce an unstable ontology of the entire image (*Sense of Things*, 89, 97).

22. Szegedy et al., “Going Deeper with Convolutions.”

23. Virilio, *Vision Machine*, 59.

24. Martin A. Fischler, a key figure in the early development of computer vision and machine learning, makes a sharp distinction between seeing and perceiving in an early conference paper on computer vision: “‘Seeing’ is the passive reception of visual data; we might say that a T.V. camera ‘sees’ the object it is focused on. ‘Perceiving’ is an active creative process in which a visual scene is decomposed into meaningful units. This decomposition is a function of the stimulus pattern, the vocabulary and experience of the observer, and the ‘psychological set’ of the observer” (“Machine Perception,” 630).

25. Ribeiro, Singh, and Guestrin, “‘Why Should I Trust You?’”

26. Johnson, “Markup Bodies,” 70.

27. Ian Bogost has a more positive account of flat ontologies, especially in relation to digital systems, as the condition in which “all things equally exist, yet they do not exist equally” (*Alien Phenomenology*, 12).

28. See Geirhos et al., “ImageNet-Trained CNNs Are Biased,” and Hermann, Chen, and Kornblith, “Origins and Prevalence of Texture Bias.”

29. Offert and Bell, “Perceptual Bias and Technical Metapictures,” 1139.

30. Steyerl, “Sea of Data,” 9.

31. Benjamin, *Race after Technology*, 80.

32. A detailed social analysis of the inequalities appearing in several decision-making systems, models, and algorithms can be found in Eubanks, *Automating Inequality*.

33. Chun, *Discriminating Data*, 58.
34. Amoore, *Cloud Ethics*, 71.
35. Brown, *Repeating Body*, 191.
36. Weatherby and Justie, "Indexical AI," 384.

## References

Amoore, Louis. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham, NC: Duke University Press, 2020.

Atanasoski, Neda, and Kalindi Vora. *Surrogate Humanity: Race, Robots, and the Politics of Technology Futures*. Durham, NC: Duke University Press, 2019.

Benjamin, Ruha. *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity, 2019.

Bogost, Ian. *Alien Phenomenology, Or What It's Like to Be a Thing*. Minneapolis: University of Minnesota Press, 2012.

Brown, Bill. *A Sense of Things: The Object Matter of American Literature*. Chicago: University of Chicago Press, 2003.

Brown, Kimberly Juanita. *The Repeating Body: Slavery's Visual Resonance in the Contemporary*. Durham, NC: Duke University Press, 2015.

Browne, Simone. *Dark Matters: On the Surveillance of Blackness*. Durham, NC: Duke University Press, 2015.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. New York: Proceedings of Machine Learning Research, 2018.

Chun, Wendy Hui Kyong. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA: MIT Press, 2021.

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

Davinder, Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. "Trustworthy Artificial Intelligence: A Review." *ACM Computing Surveys* 55, no. 2 (2022): 1–38.

Dobson, James E. *The Birth of Computer Vision*. Minneapolis: University of Minnesota Press, 2023.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martins, 2017.

Fischler, Martin A. "Machine Perception and Description of Pictorial Data." In *Proceedings of the First International Joint Conference on Artificial Intelligence*, 629–39. Washington, DC: IJCAI, 1969.

Franklin, Seb. *The Digitally Disposed: Racial Capitalism and the Informatics of Value*. Minneapolis: University of Minnesota Press, 2021.

Gaboury, Jacob. *Image Objects: An Archeology of Computer Graphics*. Cambridge, MA: MIT Press, 2021.

Galloway, Alexander R. *The Interface Effect*. Malden, MA: Polity, 2012.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. "ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness." arXiv preprint, November 29, 2018. https://arxiv.org/abs/1811.12231.

Hebb, Donald O. *The Organization of Behavior: A Neuropsychological Theory*. 1949; repr., New York: Taylor and Francis, 2002.

Hermann, Katherine, Ting Chen, and Simon Kornblith. “The Origins and Prevalence of Texture Bias in Convolutional Neural Networks.” *Advances in Neural Information Processing Systems* 33 (2020): 19000–19015.

Johnson, Jessica Marie. “Markup Bodies: Black [Life] Studies and Slavery [Death] Studies at the Digital Crossroads.” *Social Text*, no. 137 (2018): 57–79.

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. “Backpropagation Applied to Handwritten Zip Code Recognition.” *Neural Computation* 1, no. 4 (1989): 541–51.

Lettvin, Jerome Y., Humberto R. Maturana, Warren S. McCulloch, and Walter H. Pitts. “What the Frog’s Eye Tells the Frog’s Brain.” *Proceedings of the IRE* 47, no. 11 (1959): 1940–51.

McCulloch, Warren S., and Walter Pitts. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” *Bulletin of Mathematical Biophysics* 5 (1943): 115–33.

Offert, Fabian, and Peter Bell. “Perceptual Bias and Technical Metapictures: Critical Machine Vision as a Humanities Challenge.” *AI and Society: Knowledge, Culture, and Communication* 36, no. 4 (2021): 1133–44.

OpenAI. “GPT-4 Technical Report.” arXiv preprint, March 27, 2023. https://arxiv.org/abs/2303.08774.

Parisi, Luciana. “Negative Optics in Vision Machines.” *AI and Society: Knowledge, Culture, and Communication* 36, no. 4 (2021): 1281–93.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. “‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1344. San Francisco: Association for Computing Machinery, 2016.

Smith, Marcus, and Seumas Miller, “The Ethical Application of Biometric Facial Recognition Technology.” *AI and Society* 37, no. 1 (2022): 167–75.

Steyerl, Hito. “A Sea of Data: Pattern Recognition and Corporate Animism (Forked Version).” In *Pattern Discrimination*, edited by Clemens Apprich, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl, 1–22. Minneapolis: University of Minnesota Press, 2019.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. “Going Deeper with Convolutions.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9. Boston: IEEE, 2015.

Virilio, Paul. *The Vision Machine*. Translated by Julie Rose. Bloomington: Indiana University Press, 1994.

Weatherby, Leif, and Brian Justie. “Indexical AI.” *Critical Inquiry* 48, no. 2 (2022): 381–415.

Wingfield, Nick. “Amazon Pushes Facial Recognition to Police. Critics See Surveillance Risk.” *New York Times*, May 22, 2018.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs, 2019.